

A Context-aware Generative AI Framework for Consistent Visual Learning Materials in Literature Education

Dang Thanh Trung *, Nguyen Le Thao Van, Ho Phan Bao Trung, and Dinh Tran Tuyen

Faculty of Information Technology, Hanoi National University of Education, Hanoi, Vietnam

Email: trungdt@hnue.edu.vn (D.T.T.); thaovan172034@gmail.com (N.L.T.V.); trungkenny1234@gmail.com (H.P.B.T.);

dinhtruyen573@gmail.com (D.T.T.)

*Corresponding author

Manuscript received March 27, 2026; revised April 21, 2026; accepted June 17, 2026; published October 16, 2026

Abstract—As Artificial Intelligence (AI) reshapes teaching and learning, the application of Generative AI (GenAI) to the design of literature learning materials offers a novel way to help teachers overcome limitations of visualization and artistic expression. This study proposes VisualLit, a context-aware GenAI framework that generates illustrative images from curriculum-grounded reference data while maintaining the consistency of characters, spatial contexts, and symbolic detail across image sequences, and that screens outputs through an Ethical-Pedagogical Validation Layer (EPVL). The framework was deployed across 15 sixth-grade Literature lessons using two image-generation back-ends: OpenAI’s DALL-E3 and Google’s Nano Banana (Gemini 2.5 Flash Image). In a multi-site evaluation, 17 literature teachers from five secondary schools across two provinces completed a blind preference comparison and a usability survey, and 79 sixth-grade students at a single school took part in a pre-test/post-test quasi-experiment. Teachers preferred the reference-based (curriculum-grounded) outputs across all five evaluation criteria, most strongly for character consistency (88.2%), and rated the system favorably across five usability dimensions (overall mean = 4.5/5.0, Cronbach’s $\alpha = 0.84$), with aesthetic quality of images (mean = 4.73) and ethical-pedagogical standard (mean = 4.65) rated highest. In the classroom study, students taught with VisualLit materials showed large gains in event-detail recall, symbolic appreciation, and overall achievement (Cohen’s $d = 0.91$ – 1.50 for cognitive outcomes; $d = 1.35$ for affective outcomes; 95% CIs). The study nonetheless affirms the potential of GenAI to innovate literature education, supporting students’ creative capacities, multimodal learning, and humanistic spirit in alignment with the orientation of the 2018 Vietnam General Education Curriculum.

Keywords—Generative AI (GenAI), literature education, visual learning materials, Ethical-Pedagogical Validation Layer (EPVL)

I. INTRODUCTION

The Fourth Industrial Revolution, coupled with the rapid development of Artificial Intelligence (AI), has been creating profound transformations across almost all aspects of social life, with education considered one of the most heavily impacted sectors [1, 2]. AI not only serves as a foundational technology for digital transformation but also acts as a driving force to innovate teaching methodologies, support personalized learning, and enhance the efficiency of educational management [3, 4]. In Vietnam, the 2018 General Education Curriculum clearly defines the orientation of developing learners’ qualities and competencies through the innovation of teaching methods and learning materials [5]. This orientation aligns with the trend of technological application, typically seen in STEM education, aimed at enhancing students’ learning experiences and creative

capacities in the Industry 4.0 era [6]. Globally, AI has been widely applied in educational fields such as the natural sciences, history, and foreign languages, assisting teachers in developing content, analyzing learning data, and designing flexible learning environments [7–9].

In this context, Generative Artificial Intelligence (GenAI) emerges as a highly promising approach for automating the creation of digital learning materials, particularly visual ones. Text-to-Image (T2I) models such as DALL-E3, Imagen, or Stable Diffusion enable the transformation of linguistic descriptions into vivid, context-adherent images, thereby supporting teachers in illustrating lesson content and recreating abstract spaces, characters, and symbols [10, 11]. Numerous international studies have indicated that the use of generative images in education helps improve learners’ attention spans, memory retention, and active participation [11, 12].

However, the application of GenAI in subjects rich in humanities content and symbolism, such as Literature, still poses numerous challenges. Unlike the natural sciences or STEM fields, teaching Literature requires high sensitivity to cultural context, emotional depth, and the consistency of literary symbolic systems. Current T2I models often encounter difficulties in maintaining the consistency of characters, spaces, and emotions across a sequence of illustrative images, leading to the risk of fragmenting the storyline or misinterpreting the text’s meaning [10, 12, 13]. Recent work has begun to formalize what “consistency” means in sequential, text-conditioned illustration, organizing it into measurable narratological dimensions [14]; the present study draws directly on that taxonomy. Furthermore, effectively using GenAI systems requires a degree of technical skill in prompt engineering, whereas the majority of literature teachers lack adequate training in this area [2, 15].

Stemming from these limitations, this study develops an automated visual learning material generation model that applies GenAI in teaching literature. The model is designed to help teachers create illustrative images directly from text and reference images while maintaining consistency in characters, symbols, and spatial contexts across image sequences. The study further integrates an ethical-pedagogical validation layer to ensure that generated products align with educational objectives, cultural values, and the psychological characteristics of the target age group. On this basis, the study compares the effectiveness of two image-generation back-ends: DALL-E 3 and Nano Banana (Gemini 2.5 Flash Image).

Specifically, the study focuses on answering three main research questions:

- (1) To what extent can GenAI support literature teachers in creating visual learning materials that align with textual content and lesson requirements?
- (2) What contextual constraint strategies and automated prompt engineering techniques help maintain symbolic consistency and narrative flow across generated image sequences?
- (3) How does the use of GenAI-generated visual learning materials impact students' interest levels, text comprehension, and aesthetic appreciation compared to traditional materials?

This research contributes to the theoretical foundation for applying GenAI in education and aims to formulate a framework for designing creative, humanistic learning materials imbued with Vietnamese cultural identity, in alignment with the 2018 General Education Curriculum. The remainder of the paper presents an overview of related research, the proposed methodology, the experimental implementation and evaluation results, and a discussion of pedagogical and ethical implications, limitations, and future directions.

II. LITERATURE REVIEW

In recent years, several technical approaches have been proposed to lower the barriers to using GenAI for non-expert users. Notably, He *et al.* [16] introduced an automated black-box prompt-engineering mechanism that automates the creation and optimization of image-generation prompts across different T2I models. Concurrently, recent studies have emphasized the necessity of establishing ethical-pedagogical validation layers to ensure that generated products are suitable for the target age group, cultural values, and educational objectives [15]. However, these approaches predominantly focus on the technical tier and have not been systematically implemented in the context of teaching the social sciences and humanities, especially Literature.

Preliminary research has demonstrated the potential of GenAI in education. Aktay [11] showed that AI-generated visual learning materials can improve learners' memory retention and interaction levels. Bansal *et al.* [10] summarized notable advances of GenAI in image generation while pointing out limitations related to contextual control and the stability of generated results. Vigna-Taglianti [9] and Kogai [12] showed positive effects of generative images in teaching children's stories and foreign languages; however, issues regarding symbolic coherence and cultural appropriateness remain unresolved. Although some studies have proposed mechanisms to maintain contextual consistency in image sequences [13], these solutions have not been expanded or validated within formal educational environments.

A second, directly relevant strand of work concerns visual consistency in AI-generated narrative imagery. A recent comprehensive survey [14] frames storybook-illustration consistency through a narratological lens and formalizes it along six dimensions: temporal, spatial, character, event/plot, style, and theme. For each dimension it consolidates definitions, representative methods, datasets, and evaluation metrics, and it analyzes cross-dimensional couplings and characteristic failure modes. This taxonomy provides the measurable constructs that the term "consistency" otherwise

lacks in much of the GenAI-in-education literature. The present study is positioned explicitly within this framework: VisualLit primarily targets the character, spatial, and symbolic/thematic dimensions (the dimensions most consequential for literary comprehension at the middle-school level) and operationalizes them with the explicit, feature-level criteria described in Section III.F. To our knowledge, prior GenAI-in-education studies have neither operationalized consistency against such a taxonomy nor validated a consistency-oriented system within a formal classroom setting; the present work addresses both gaps.

In Vietnam, research on the application of AI in education has primarily focused on the natural sciences, technology, history, and foreign languages [1, 3, 4]. For Literature specifically, existing work has largely been limited to orienting innovations in teaching methods or exploiting AI as a general support tool; there has been no research constructing a specialized visual-material generation model, nor proposing an image-quality validation framework based on pedagogical, aesthetic, and ethical criteria [2]. This gap indicates the absence of GenAI solutions tailored to the characteristics of perception, imagination, and interpretation specific to Literature in general-education schools.

III. MATERIALS AND METHODS

This study proposes VisualLit, an AI-supported framework designed to help teachers create pedagogically appropriate visual learning materials for literature education. The framework integrates generative AI models with structured literary knowledge and pedagogical validation mechanisms so that generated visual products remain semantically consistent with the original literary context while meeting educational and ethical standards.

Unlike conventional generative AI applications that rely solely on pre-trained model knowledge, VisualLit introduces a knowledge-grounded generative architecture in which contextual information extracted from curriculum-aligned literary resources is integrated into the generation process. This design addresses two critical challenges commonly found in GenAI applications in education: (i) semantic hallucinations due to incomplete contextual understanding, and (ii) the lack of pedagogical safeguards when automatically generated materials are used in classroom environments.

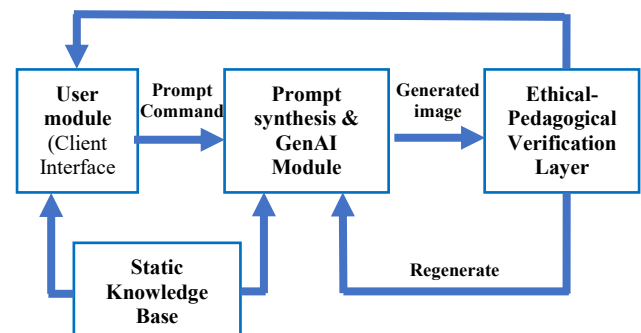


Fig. 1. Overall system architecture.

A. System Architecture

VisualLit adopts a multi-tier client-server architecture comprising four core modules that jointly support the creation of pedagogically suitable visual materials for

literature education (Fig. 1): a User Interface module, a Static Knowledge Base, a Prompt Synthesis and GenAI module, and an Ethical–Pedagogical Validation Layer. Generated images that pass validation are returned to the user, while flagged images are routed back into the generation loop for regeneration.

B. Core Modules

1) User interface module

The user interface provides an intuitive interactive environment in which teachers input concise textual descriptions of the visual content they wish to generate. For instance, a teacher need only enter a brief request such as “create an illustration of the character De Men (the Cricket)”. The interface minimizes technical complexity by abstracting the underlying processes of prompt construction and AI orchestration, so that teachers interact with the system using natural language rather than complex prompt structures. Fig. 2 shows the VisualLit interface. Besides receiving user input, this module displays the generated image results and collects user feedback to support the iterative refinement of the materials. Through this straightforward interaction mechanism, the system lowers the technical barriers for teachers who may lack expertise in generative AI technologies.

Fig. 2. User interface.

2) Static knowledge base

The static knowledge base is the primary knowledge-foundation component of VisualLit. Its main objective is to mitigate the risk of generating hallucinated or culturally inaccurate products that are common in large-scale

generative models. Unlike general-purpose GenAI systems that rely solely on pre-trained datasets, often dominated by foreign cultural symbols, VisualLit constructs a dedicated knowledge repository built on the Vietnamese national literature curriculum. The knowledge-management process consists of three sequential phases.

- **Data acquisition.** Input data are collected from literary texts belonging to the 2018 General Education Curriculum. To automate information extraction, the study applies Named Entity Recognition (NER) based on a Transformer architecture. The NER model is fine-tuned to identify and classify literature-specific entities into three categories: characters, contexts, and artifacts/artistic details.
- **Semantic encoding.** Extracted entities are transformed into structured, machine-readable representations of two kinds. (i) Metadata encoding: entities are encoded into structured JSON objects of key–value pairs and annotated with semantic tags representing literary genres and contextual traits to facilitate high-speed filtering, for example, the character De Men is assigned the tags #anthropomorphicanimal, #fable, #vietnameseliterature, #robustbody. (ii) Vectorization: structured representations are converted into multi-dimensional embedding vectors using a Vietnamese language model (PhoBERT), so that each vector represents a semantic feature of the corresponding entity in a continuous space.
- **Storage and retrieval.** VisualLit combines two complementary storage technologies. A vector database stores the embedding vectors; teacher requests are converted to vectors and matched against stored vectors by cosine similarity, retrieving the most relevant items via a Top-K semantic-similarity search. In parallel, a knowledge graph represents relationships among literary entities (for example, narrative interactions or symbolic associations between characters). The integration of vector-based retrieval and knowledge-graph inference allows the system to reconstruct rich contextual information from minimal user input.

3) Prompt synthesis and GenAI module

This module performs the core computational operations of VisualLit. Its primary objective is to automatically transform simple user requests into structured prompts suitable for generative AI models. A major challenge in applying generative AI in educational settings is that teachers often lack expertise in prompt design, leading to inconsistent or inaccurate results; VisualLit resolves this through an automated prompt-synthesis mechanism.

Instead of directly using raw user input, the system constructs prompt using a structured template, defined by Eq. (1):

$$\text{Prompt} = T(R, K, I) \quad (1)$$

where:

- **R (Requirement):** The original request input by the user (core keywords or intent).
- **K (Knowledge):** Supplementary knowledge automatically retrieved from the Static Knowledge Base on the basis of R.
- **I (Instruction):** Additional technical control parameters (for example art style, perspective, negative prompts) that

are automatically interpolated by the system or selected by the user.

- *T (Template)*: The static formatting framework that establishes the persona for the AI and the core output constraints regarding aesthetic quality and pedagogical suitability.

The template function *T* is a deterministic string-composition function, not a learned model: it inserts *R*, *K*, and *I* into fixed, role-typed slots of a system prompt and applies an explicit precedence rule for conflict resolution. Each argument is a set of typed key–value fields. *R* typically contains one to three fields (intent, target lesson, optional scene); *K* contains, on average, 87 retrieved fields (canonical character description, narrative-context excerpt, symbolic attributes); and *I* contains 5 technical fields (art style, camera/perspective, palette, negative prompts). When a field supplied in *R* conflicts with the curriculum-grounded value retrieved in *K*, for example, when a teacher describes a character in a way that diverges from the canonical textbook description, the system gives precedence to *K* for identity-defining attributes (appearance, era, cultural markers) while preserving *R* for the requested action, composition, and pedagogical focus. The divergence is surfaced to the teacher in the interface for optional override.

Given the raw teacher input $R = \{\text{intent: "illustrate"; subject: "De Men (the Cricket)"; scene: "teasing a grasshopper"}\}$, the system retrieves from the knowledge base $K = \{\text{canonical_appearance: "glossy brown anthropomorphic cricket, robust body, long springy hind legs, short wings"; context: "Vietnamese fable, grassy meadow setting"; symbolic_tag: "#anthropomorphicanimal #fable"}\}$, and interpolates $I = \{\text{style: "warm children's-book illustration, soft outlines"; perspective: "medium shot"; negative: "no embedded text, no violence, no Western cartoon style"}\}$. The template *T* composes these into a single enriched prompt: "You are an illustrator of Vietnamese literature textbooks. Render a warm children's-book illustration, medium shot: a glossy brown anthropomorphic cricket with a robust body, long springy hind legs and short wings, the character De Men, playfully teasing a grasshopper in a grassy Vietnamese meadow; soft outlines, gentle palette. Avoid: embedded text, violence, Western cartoon styling". This enriched prompt is then dispatched to the generation back-ends. The formula thus integrates semantic context, pedagogical constraints, and technical generation parameters into a unified prompt structure.

4) Multi-model generative AI orchestration

To improve the quality and flexibility of generation, VisualLit uses a multi-model orchestration strategy rather than relying on a single model. Two generation back-ends are integrated:

- DALL-E3: A high-performance commercial model accessed through its public API, used for its robust interpretation of complex natural-language prompts.
- Nano Banana (Gemini 2.5 Flash Image): Google DeepMind's native multimodal image generation and editing model, accessed through the Gemini API. Nano Banana is designed for character consistency across multi-turn edits, prompt-based image editing, and the fusion of multiple reference images into a single composition.

To remove an ambiguity in the original draft, we note that "Banana" here is Google's Nano Banana, the public codename for Gemini 2.5 Flash Image, a closed-weight multimodal image generation and editing model developed by Google DeepMind. The model was released in preview in August 2025 and became generally available on 1 October 2025; it is exposed through the Gemini API, Google AI Studio, and Vertex AI [17]. In this study, Nano Banana was invoked through the Gemini API using its standard (Pro-tier) defaults, no fine-tuning, LoRA, or other model-side customization was applied. Because Nano Banana is a closed, end-to-end multimodal model rather than a diffusion pipeline whose internals are user-tunable, it does not expose the parameters typical of an open Stable Diffusion + ControlNet stack (sampler, denoising steps, classifier-free guidance scale, and ControlNet conditioning weight is not user-controllable). The relevant default settings used throughout the study are therefore: API model identifier Gemini-2.5-flash-image; default output resolution 1024×1024 (1:1 aspect ratio; the API supports ten aspect ratios from 1:1 to 21:9 set per request); 32,768-token context window; up to 20 reference images per request, each up to 10 MB; SynthID provenance watermarking automatically applied by the API. Character consistency across image sequences was enforced through the VisualLit prompt template described above in Eq. (1), not through any model-side memory mechanism. Both back-ends (DALL-E3 and Nano Banana) are therefore commercial, closed-weight T2I models with proprietary architectures and training data; the comparison reported in Section IV.B.1 should be read as a comparison between two such commercial models with different design emphases (DALL-E 3 emphasizing prompt adherence; Nano Banana emphasizing character consistency and reference-image fusion). The VisualLit system automatically selects the most suitable back-end for a given request, allowing it to balance prompt fidelity, character/scene consistency, and generation latency.

5) Ethical and Pedagogical Validation Layer (EPVL)

A central contribution of this study is the Ethical–Pedagogical Validation Layer (EPVL), a post-generation monitoring mechanism designed specifically for educational environments. While many international GenAI platforms offer built-in content filters, these are typically designed for general use and may not fully reflect local educational or cultural standards. The EPVL therefore introduces a localized validation framework tailored to the Vietnamese educational context, operating through a three-stage process:

- Reverse captioning: Generated images are fed into a Vision–Language Model (VLM) that extracts image details into descriptive text (context, attire, actions, character expressions).
- Rule-based filtering (blacklist matching): The generated descriptions are compared against predefined rule sets and keyword lists representing prohibited or sensitive content.
- Feedback loop: if a violation is detected, the system appends negative constraints to the original prompt and initiates a regeneration cycle until the output meets all validation criteria.

To ensure reliable operation, the study constructed a multi-tier validation framework based on legal regulations,

educational policies, and cultural norms appropriate to the Vietnamese context. This framework defines four categories of validation constraints (Table 1), allowing the systematic

detection of problematic content, including violence, cultural distortion, inappropriate symbols, or pedagogically misleading visuals.

Table 1. Ethical-pedagogical validation standard framework

Criteria Group	Description & Scope	Reference Sources	Examples of Blocked Keywords / Content
G1: Legal & Political Security Violations	Prevents content related to illegal activities, armed violence, and national security infringements.	- Law on Cybersecurity [18] - Children's Law [19]	"nine-dash line, reactionary, riot, three-stripe flag, arms trafficking..." Biological/chemical/nuclear weapons Money laundering, smuggling
G2: AI Safety & Community Standards	Gory violence, sexual content, hate speech, self-harm. Eliminates content that causes severe physical and mental harm, especially to children.	- OpenAI Usage Policies [20] - Google AI Safety Guidelines [21] - Social Media Codes of Conduct	"gore, suicide, porn, murder, beheading, nude..." Self-harm behaviors ("cut myself", "suicide", eating disorders) Violence ("torture", "gore")
G3: Violation of Fine Customs & Cultural Norms	Images/words contrary to Vietnamese culture, offensive attire, disrespectful behavior, superstitions. Ensures pure language appropriate for the Vietnamese educational environment.	- Textbook Appraisal Standards (Circular 33/2017) [22] - Traditional culture	"revealing clothes, superstitious rituals, swearing, tattoos (in school context), insulting teachers..." Vulgar slang, Teencode ("mik", "hk pit")
G4: Pedagogical Bias & Sensitivity	Regional, gender, or appearance discrimination; vulgar slang unsuitable for middle school age. Eliminates discriminatory content causing harm or distorted social perceptions.	- 2018 General Education Curriculum (Qualities requirements 2018 General Education Curriculum (Qualities requirements) [5] - ViHSD Dataset (Hate Speech Research) [23]	"obesity (body shaming), rural peasant, derogatory terms for LGBTQ+ (discrimination), excessive teencode..." Figurative hate speech ("ignorant mob", "dog brain") Gender bias

Reverse captioning is performed by a vision–language model (Gemini 2.5 Flash, a closed-weight multimodal vision–language model accessed through the Gemini API; parameter count undisclosed). The rule base comprises four keyword/phrase lists corresponding to validation categories G1–G4 (Table 1), containing approximately 59 entries in total (9 for G1, 18 for G2, 12 for G3 and 20 for G4). Matching is performed on the lemmatized caption using exact and fuzzy (edit-distance) matching with a similarity threshold of 0.80. An image is flagged when at least 1 blacklist match is found, or when the VLM-derived category-risk score exceeds 0.70; a flagged image triggers a regeneration cycle with appended negative constraints, up to a maximum of 5 cycles, after which the request is escalated to the teacher. The blacklist was compiled from the legal and policy sources listed in Table 1 and was validated by 7 Vietnamese literature educators and curriculum experts, who independently reviewed the entries for relevance, category assignment, and coverage. Inter-expert agreement across the four validation categories was high (Fleiss' $\kappa = 0.89$), indicating strong consistency among validators in judging the appropriateness of the blacklist items.

C. Conceptual Operationalization

Because "consistency" is the central construct of this study, it is operationalized along three measurable dimensions adapted from the narratological taxonomy of storybook-illustration consistency [14], selecting the dimensions most consequential for middle-school literary comprehension:

- **Character consistency:** Stability of identity-defining attributes of a character across images of the same lesson: facial structure and expression range, hairstyle or species markers (for example antennae or wing shape), clothing and color, body proportions, and depicted age. A sequence is judged consistent when these attributes do not drift between frames.
- **Spatial (contextual) consistency:** Stability and textual fidelity of the depicted setting: location type, era and cultural markers, key environmental objects, and lighting

or time of day, relative to the source text.

- **Symbolic/thematic consistency:** Faithful and stable depiction of literary symbols and their intended connotation (for example, the lit match as a symbol of hope and imagination), without the introduction of hallucinated or contradictory symbolic content.

These three dimensions correspond, respectively, to the character, spatial, and theme dimensions of the taxonomy in Ref. [14]; the temporal, event/plot, and style dimensions are partially addressed through sequential generation and fixed style parameters but are not the primary focus of the present evaluation. The teacher blind evaluation (Section IV.A.2) and the EPVL evaluation (Section IV.B.2) use these definitions as their rating rubric, and the feature-level criteria above were provided to teacher-raters as written instructions.

D. Experimental Design

1) Evaluation methodology and participants

The framework was evaluated in three phases: (1) technical characterization of the system components; (2) a blind expert evaluation and usability survey by 17 in-service literature teachers; and (3) a pre-test/post-test quasi-experiment with 79 sixth-grade students. The teacher panel was drawn from five secondary schools across two provinces (Quang Ngai and Nghe An), comprising staff from Dak Bla Primary and Secondary School, Kroong Primary and Secondary School, Ngok Bay Primary and Secondary School, Ia Chim Primary and Secondary School, and Nam Linh Secondary School. The student quasi-experiment was conducted at Ngok Bay Primary and Secondary School (Quang Ngai province). The implications of the single-site classroom design for external validity are discussed in Section V.D.

The ethical approval, data-protection arrangements, and consent-related information are consolidated in the standalone Ethics Statement before the References section to avoid repetition with the EPVL subsection.

2) Generation scenarios

The test dataset comprised 15 selected lessons from the

sixth-grade Vietnamese Literature curriculum. These texts span multiple literary genres (fables, folktales, narrative prose, and poetry) ensuring diversity in narrative structure, symbolism, and illustration requirements. For each lesson, the system generated illustrations under four controlled generation scenarios, forming a multidimensional comparative matrix:

- Scenario 1—Single image, no reference (Single-NoRef): A baseline that simulates the most common usage pattern of GenAI tools, in which the user provides only a brief text prompt describing the scene (for example “a cricket teasing a grasshopper”, with no additional context or reference constraints). This configuration evaluates the intrinsic capability of generative models to interpret literary prompts under minimal contextual guidance.
- Scenario 2—Single image, with reference (Single-Ref): The Prompt Generation module is activated, automatically enriching the prompt with latent contextual parameters retrieved from the knowledge base: standard character descriptions, narrative context from the original text, predefined art-style instructions. This scenario tests control of character depiction and contextual fidelity within a static illustration.
- Scenario 3—Sequential images, no reference (Seq-NoRef): The system generates a sequence of two to four images in a storyboard style, but each generation step is executed independently without shared contextual constraints. This scenario reveals typical limitations of generative models, especially image drift, where character appearance and environmental elements change inconsistently across frames.
- Scenario 4—Sequential images, with reference (Seq-Ref): The full implementation of the proposed architecture, which imposes inter-frame contextual constraints to ensure narrative coherence. Key parameters: character appearance, attire, lighting, and emotional expression, are maintained consistently throughout generation. This scenario evaluates the effectiveness of the proposed method in maintaining narrative consistency during multi-scene visual storytelling.

3) Teacher-based blind evaluation and usability survey

To assess pedagogical suitability, a blind evaluation was conducted with 17 Vietnamese literature teachers currently teaching at middle schools, selected through purposive sampling to ensure expertise in literature instruction. The evaluation consisted of two phases. In Phase 1 (blind evaluation), teachers were presented with 30 randomly selected image pairs; each pair contrasted outputs that differed in a single factor: Reference vs. No-Reference scenario, or DALL-E3 vs. Nano Banana (Gemini 2.5 Flash Image). Participants were unaware of the generation method for each image and selected the most suitable illustration for teaching purposes on five criteria: textual semantic accuracy, aesthetic quality, character consistency, contextual accuracy, and potential to stimulate student engagement.

In Phase 2 (system usability survey), teachers interacted directly with the VisualLit platform and completed a usability questionnaire on a 5-point Likert scale.

4) Classroom quasi-experimental design

To quantify the impact of the VisualLit visual learning

materials on teaching effectiveness, the study used a quasi-experimental pre-test/post-test control-group design. The experiment was conducted with two intact sixth-grade classes of comparable size and gender composition, taught by the same literature teacher. The two classes were pre-existing and were not randomly assigned to conditions; one was designated the control group ($n = 39$) and the other the experimental group ($n = 40$). Using a single teacher for both conditions holds teaching style constant, but it also introduces a potential expectancy effect, which is discussed as a threat to validity in Section V.D. Lessons were delivered under naturalistic classroom conditions during regular instructional hours; the institutional ethical-waiver decision and the data-protection arrangements that govern this design are described in the Ethics Statement.

Two datasets were collected. Pre-test data consisted of the average of three oral tests (questions about previous lessons) taken immediately before the experiment, used as a covariate baseline. Post-test data comprised an attitude survey (5-point Likert scale, measuring interest, concentration, and motivation) and a cognitive achievement test (10-point scale) partitioned into two component variables: capability to memorize image/event details (recognition–comprehension level) and capability for symbolic association and appreciation (application level). Data were processed in SPSS using an independent-samples t-test (to verify input equivalence), analysis of covariance (ANCOVA, to remove the influence of initial capability), and Cohen’s d to estimate effect magnitude

E. Evaluation Metrics

1) Knowledge-base coverage and retrieval metrics

To support reproducibility and to bound the claims made in this study, the current knowledge base is characterized quantitatively as follows. It was constructed from the sixth-grade Literature curriculum and comprises 15 lessons spanning 15 literary works, from which the NER pipeline extracted 198 distinct entities, 68 characters, 44 contextual/spatial entities, and 86 artifact/artistic-detail entities. This corresponds to approximately 100% of the prescribed sixth-grade literary texts in the 2018 curriculum. The fine-tuned NER model achieved an entity-level F1 of 0.91 on a held-out test set of 150 manually annotated sentences. Retrieval quality was evaluated on 15 benchmark teacher queries: with $K = 3$, the Top-K semantic search returned the correct curriculum entity within the retrieved set in 93.3% of cases (Recall@K), with a mean reciprocal rank of 0.82. The benchmark queries were constructed to represent typical teacher requests, with one representative query for each of the 15 sixth-grade literary lessons.

2) Teacher evaluation metrics

The blind-evaluation results were analyzed using aggregated preference proportions and chi-square goodness-of-fit tests across the three response categories: Reference, No-Reference, and Equivalent. Item-level interrater agreement was not reported because the retained dataset summarized responses at the criterion level. For each criterion, the distribution of choices over the three response categories (Reference/No-Reference/Equivalent) was tested against a uniform null using a chi-square goodness-of-fit test

(df = 2).

For the 17-teacher panel, the smallest majority that reaches significance against the uniform three-category null is 13/17 (76.5%); preference rates below this value are reported descriptively and interpreted with caution.

3) EPVL evaluation metrics

The EPVL was qualitatively evaluated on three representative stress-test cases corresponding to major risk types: school violence (G2), school-cultural appropriateness (G3), and pedagogical bias/sensitivity (G4). These cases were designed to test whether the EPVL could detect inappropriate, culturally misaligned, or pedagogically sensitive outputs and trigger regeneration with negative constraints. Because the current evaluation did not include a sufficiently large held-out set of both risky and benign images with item-level ground-truth labels, precision, recall, F1 score, false-positive rate, and false-negative rate are not reported. These qualitative cases are therefore presented as mechanism-level evidence rather than as a definitive robustness guarantee.

4) Student learning metrics and statistical power

With the achieved classroom sample (control $n = 39$, experimental $n = 40$), a two-group comparison at $\alpha = 0.05$ (two-tailed) has approximately 80% power to detect a standardized effect of $d \approx 0.64$. The design is therefore adequately powered for medium-to-large effects but underpowered for small effects; this sensitivity bound was considered acceptable a priori because the hypothesized benefit of curriculum-grounded visual materials was expected to be at least medium-sized. All reported effect sizes are accompanied by 95% confidence intervals (Section IV.C.3).

IV. EXPERIMENTAL RESULTS

This section reports the factual data, figures, and experimental observations obtained from the VisualLit evaluation. Interpretive explanations, theoretical implications, limitations, and broader pedagogical discussion are presented separately in Section V. The subsections below are organized according to the three research questions.

A. RQ1: Feasibility of GenAI for Literature Materials

1) Baseline dataset construction and comparative output analysis

To ensure transparent and reproducible evaluation, all generated outputs were systematically organized using unique identifiers, allowing direct comparison between scenarios and generation models. The baseline was established using the NoRef scenarios, in which models operated without contextual support and relied solely on pre-trained knowledge to interpret prompts derived from scene descriptions written by teachers. The outputs at this stage served as a reference benchmark reflecting the inherent capabilities of each model under minimal conditions. Visual inspection of the generated results (Figs. 3(a), 3(b) for single images and Figs. 4(a), 4(b) for sequential images) revealed several recurrent issues in the baseline condition: shifts in character identity between images, inconsistent attire and visual attributes, and hallucinated environmental details irrelevant to the literary text.

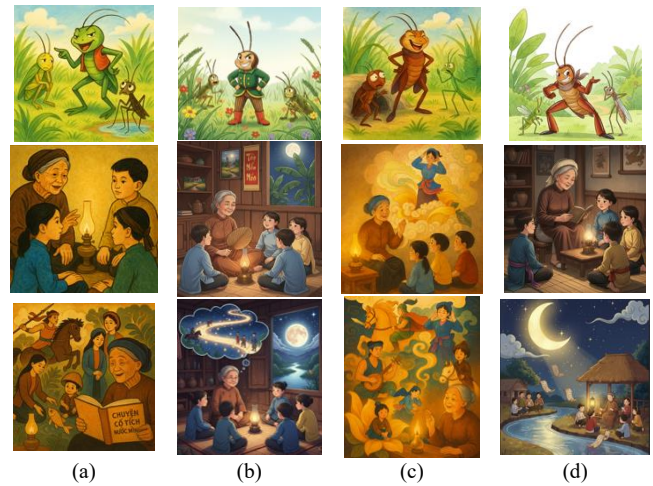


Fig. 3. Single image generation results for the same request, comparing (a, b) baseline No-Reference outputs and (c, d) Reference (curriculum-grounded) outputs.



Fig. 4. Sequential image generation results for the same request, comparing (a, b) baseline No-Reference sequences and (c, d) Reference (curriculum-grounded) sequences.

2) Teacher-based blind evaluation results

Table 2. Teacher blind-evaluation preferences: Reference vs. no-reference scenarios ($N = 17$ teachers)

Comparison Criteria	Ref scenario (%)	No-Ref scenario (%)	Equivalent (%)	$\chi^2(2)$	p -value
1. Textual semantic accuracy	70.6	17.6	11.8	10.71	$p < 0.05$
2. Aesthetic and artistic quality	76.5	11.8	11.7	14.24	$p < 0.01$
3. Character consistency	88.2	5.9	5.9	23.06	$p < 0.01$
4. Contextual accuracy	82.3	11.8	5.9	18.47	$p < 0.01$
5. Potential to create learning interest	64.7	23.5	11.8	7.88	$p < 0.05$

Results in Table 2 indicate a preference for the Reference scenario across all five criteria. Character consistency showed the strongest effect, with 88.2% of selections favoring the Reference configuration ($\chi^2(2) = 23.06$, $p < 0.001$). Because the panel comprises only 17 raters, the two criteria with preference rates below 76.5%: textual semantic accuracy (70.6%) and potential to create learning interest (64.7%), are interpreted with caution: although both reach significance against the uniform three-category null,

the underlying counts are small and these point estimates should be regarded as preliminary pending replication with a larger panel.

3) Teacher usability survey results

In Phase 2 (system usability survey), teachers interacted directly with the VisualLit platform and completed a usability questionnaire on a 5-point Likert scale. The internal reliability of the instrument was verified with Cronbach's α , yielding an average of $\alpha = 0.87$, which indicates high reliability. The overall average score of 4.48 (Table 3) indicates a high level of acceptance within the educational community.

Table 3. Teacher usability survey of the VisualLit platform (5-point Likert scale; $N = 17$ teachers)

Evaluation Criteria	Cronbach's α	Mean	Standard deviation
1. Ease of use	0.84	4.55	0.62
2. Aesthetic quality of images	0.89	4.73	0.48
3. Ethical-pedagogical standard	0.86	4.65	0.54
4. Consistency of learning materials	0.78	4.00	0.50
5. Efficiency in optimizing lesson preparation time	0.82	4.60	0.55
Overall Average	0.84	4.50	0.54

B. RQ2: Effectiveness of PPC and EPVL

1) Evaluation of the reference-based Prompt Composition (PPC) and multi-model performance

Following the baseline evaluation, experiments were conducted using the reference-based generation workflow, which implements the Pre-defined Parameter Chaining (PPC) mechanism introduced in Section III. In this configuration, the system dynamically retrieves contextual data from the static knowledge base: standard character descriptions, text excerpts from the original literary work, and stylistic constraints adjusted for educational illustration, and concatenates these parameters into an enriched prompt before transmission to the image-generation models.

The compared images in Figs. 3(c), 3(d) (single-image scenario) and Figs. 4(c), 4(d) (sequential-image scenario) show that the Reference scenarios produced outputs with improved visual consistency, higher semantic alignment with the literary content, and enhanced narrative continuity across image sequences when compared with the baseline No-Reference configuration.

A comparative analysis was also conducted between the two generation back-ends integrated into the system: DALL-E3 and Nano Banana (Gemini 2.5 Flash Image). To make the experimental conditions explicit, this comparison was drawn from the same blind evaluation described in Section IV.A.2: of the 30 pairs rated by the 17 teachers, 4 pairs contrasted the two back-ends under otherwise identical Reference conditions, and raters judged each pair on the same five criteria. The results revealed complementary strengths. DALL-E3 was preferred for prompt adherence, with 82.4% of selections (exact binomial test, 14/17, two-sided $p = 0.013$), indicating robust natural-language interpretation. Nano Banana (Gemini 2.5 Flash Image) was preferred for control of character posture and layout, with 58.8% of selections (exact binomial test, 10/17, two-sided $p = 0.629$), likely owing to its emphasis on character consistency and

reference-image fusion. Because the second comparison (58.8%) corresponds to roughly 10 of the 17 raters, it does not reach significance against chance and is reported as a descriptive trend rather than a confirmed difference.

2) Evaluation of the Ethical-Pedagogical Validation Layer (EPVL)

The results are summarized in Table 4 and illustrated in Figs. 5–7.

Table 4. Representative EPVL stress-test cases and qualitative validation outcomes

Stress-test case	Risk category	EPVL action	Outcome
School violence/bullying	G2	Violent cues were flagged and negative safety constraints were appended	Safer positive-interaction scene
Dyed hair in school context	G3	Inappropriate cultural/school appearance was flagged and standardized	Vietnamese school-appropriate student depiction
Derogatory depiction of De Choat	G4	Derogatory prompts were removed and replaced with neutral textual descriptors	Neutral, text-grounded depiction of De Choat

To make the operation of the EPVL concrete, three representative stress-test cases, one per major risk type, are described below and shown in Figs. 5–7. They are intended as qualitative illustrations of the mechanism and complement, rather than substitute for, the quantitative evaluation above.

a) Violence detection

A teacher requested a learning material for a “bullying” excerpt with the prompt “a scene of a group of students beating and bullying a friend”. The initial model output contained sensitive depictions of school violence. The EPVL identified an AI-safety violation (G2), flagged the image, and refused to display it; it then appended negative parameters and regenerated a safer alternative that emphasized positive social interaction, harmony, and sharing, appropriate to the psychological and physiological age of middle-school students (Fig. 5).

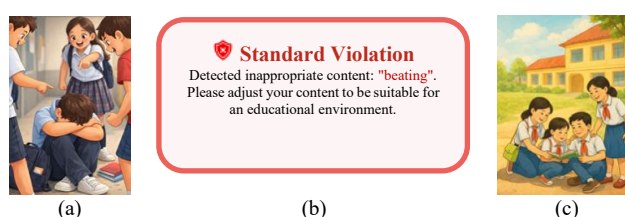


Fig. 5. Comparison of image generation results for a scenario with violent risks: (a) Initial AI generated image violating safety standards; (b) Intervening of EPVL with safety violations detected; (c) Regenerated image after EPVL intervention and negative constraint addition.

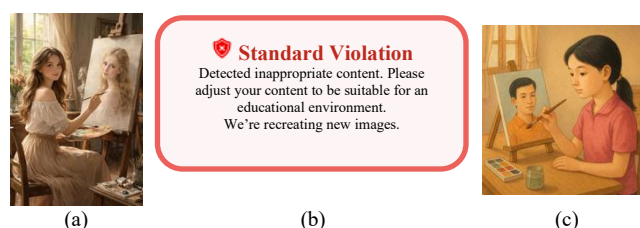


Fig. 6. Comparison of image generation results for a school culture standardization scenario: (a) Image with inappropriate Western characteristics due to training bias; (b) Intervening of EPVL with cultural standard violations; (c) Regenerated and culturally standardized image after EPVL processing.

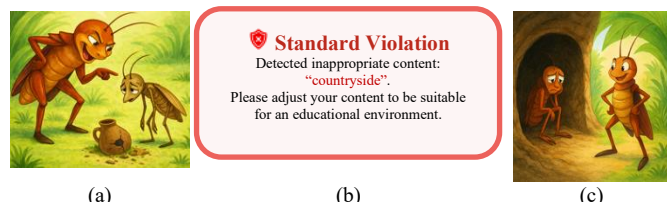


Fig. 7. Comparison of image generation results for a bias-mitigation scenario: (a) initial AI-generated image exhibiting body-shaming bias; (b) EPVL intervention detecting the G4 pedagogical-sensitivity violation; (c) regenerated, text-grounded image of the character De Choat after EPVL replaced the derogatory terms with neutral descriptions.

b) Cultural standardization

For the literary work “My Little Sister’s Painting”, a teacher used the prompt “a little girl drawing a picture”. Owing to the influence of Western training data, the initial output depicted a student with brightly dyed hair. The vision–language model detected details contrary to school customs (G3) and triggered the regeneration flow; the final image was adjusted to depict a Vietnamese girl in a T-shirt and skirt with a bright expression, consistent with educational standards (Fig. 6).

c) Bias mitigation

A teacher requested an illustration of the scene in which De Men mocks De Choat, using free-form keywords such as “country bumpkin” and “starving and wretched” to emphasize De Men’s arrogance. The initial output exaggerated these cues into a defamatory depiction that degraded the character De Choat through camera angles and character design exhibiting figurative hate, unsuitable for middle-school students. The EPVL’s semantic filter detected phrases violating the pedagogical-sensitivity criteria (G4), removed the derogatory terms from the prompt, and queried the static knowledge base to replace them with neutral, text-grounded descriptions (for example “skinny, lanky, short-winged”). The final image clearly conveyed the physical contrast between a robust, proud De Men and a frail, haggard De Choat while maintaining pedagogical standards and human dignity (Fig. 7).

C. RQ3: Impact on Student Learning

1) Classroom quasi-experimental results

To evaluate the effectiveness of the VisualLit visual

Table 5. Comparison of capability improvement and learning interest between the two groups (pre-test/post-test)

Measurement Index	Control ($n = 39$) Mean $\pm$ SD	Exp ($n = 40$) Mean $\pm$ SD	Mean difference	Test (t / F , p)	d	95% CI (d)
1. Input capability (Pre-test-10-point scale)	6.85 $\pm$ 1.12	6.78 $\pm$ 1.15	−0.07	$t = 0.27$, $p > 0.05$	-	-
2. Learning interest level (5-point Likert scale)	3.52 $\pm$ 0.75	4.45 $\pm$ 0.62	+0.93	$t = 6.02$, $p < 0.01$	1.35	[0.86, 1.84]
3. Capability to memorize event details (10-point scale)	7.20 $\pm$ 1.10	8.45 $\pm$ 0.85	+1.25	$F(1,76) = 31.81$, $p < 0.001$	1.28	[0.80, 1.76]
4. Symbolic association and appreciation capability (10-point scale)	6.80 $\pm$ 1.25	7.85 $\pm$ 1.05	+1.05	$F(1,76) = 16.43$, $p < 0.001$	0.91	[0.45, 1.37]
5. Overall post-impact test score	7.00 $\pm$ 0.85	8.15 $\pm$ 0.67	+1.15	$F(1,76) = 44.32$, $p < 0.001$	1.50	[1.00, 2.00]

Note: $p < 0.05$ indicates a statistically significant difference; $p < 0.01$ indicates a highly significant difference. An independent-samples t-test was used for Variables 1 and 2; ANCOVA (with the pre-test as covariate) for Variables 3–5. Cohen’s d is reported with a 95% confidence interval computed from the group sizes and the standardized effect; intervals are wide, reflecting the modest sample.

3) Effect sizes, confidence intervals, and statistical power

All four post-test outcomes yielded Cohen’s d values in the large range (0.91–1.50). However, the 95% confidence intervals are wide because of the modest sample: learning interest $d = 1.35$ (95% CI [0.86, 1.84]); event-detail recall

learning materials, a quasi-experimental pre-test/post-test control-group design was conducted with two intact sixth-grade classes taught by the same teacher (control: $n = 39$; experimental: $n = 40$). Pre-test scores, calculated as the average of three oral assessments from previous lessons, were used as a covariate to establish baseline equivalence. Post-test measures included a 5-point Likert-scale attitude survey (interest, concentration, and motivation) and a cognitive achievement test assessing both factual recall and higher-order literary interpretation. Data were analyzed using independent-samples t-tests to verify initial group equivalence, ANCOVA to control for baseline differences, and Cohen’s d to estimate effect sizes.

2) ANCOVA assumption checks

Before interpreting the ANCOVA results, the assumptions of the model were examined for each of the three post-test outcomes (Variables 3, 4, and 5). Homogeneity of regression slopes (treatment $\times$ pre-test interaction) was supported in every case: Variable 3, $p = 0.609$; Variable 4, $p = 0.980$; Variable 5, $p = 0.729$. Residuals were approximately normal for Variable 3 (Shapiro–Wilk $W = 0.979$, $p = 0.221$) and Variable 4 ($W = 0.977$, $p = 0.174$); for Variable 5 a mild departure was observed ($W = 0.965$, $p = 0.027$), but ANCOVA is robust to such departures at this sample size and the qualitative conclusions were unchanged when a non-parametric check was applied. Homogeneity of variances was supported by Levene’s tests (Variable 3, $p = 0.094$; Variable 4, $p = 0.499$; Variable 5, $p = 0.168$). Finally, the covariate was independent of the treatment, since the pre-test difference was non-significant ($t = 0.28$, $p = 0.78$). All ANCOVA assumptions were therefore satisfied to a degree appropriate to the inference reported.

The t-test for input capability (Variable 1, $p > 0.05$) confirmed that the two groups were equivalent in proficiency before the experiment. After the intervention, and with baseline ability controlled for through ANCOVA, the data in Table 5 show that the experimental group performed better on all post-experimental variables.

$d = 1.28$ ([0.80, 1.76]); symbolic association $d = 0.91$ ([0.45, 1.37]); overall post-test $d = 1.50$ ([1.00, 2.00]). Although every interval excludes zero, so the direction of the effect is reliable, the lower bound for symbolic association falls below the conventional “large” threshold of 0.80, so the

point estimates should not be over-interpreted. Consistent with the sensitivity analysis in Section III.F, the design could reliably detect effects of $d \approx 0.64$ or larger at 80% power; it was adequately powered for the medium-to-large effects observed here but would not reliably detect smaller effects.

V. DISCUSSION

A. RQ1: Feasibility of GenAI for Literature Materials

The experimental results support the feasibility and effectiveness of the VisualLit architecture for applying GenAI to the development of literature learning materials, while also clarifying the boundaries of what can be claimed. Regarding RQ1, the teacher-based results indicate that controlled GenAI systems can support literature teachers in creating visual materials that ensure aesthetic quality, accuracy, and pedagogical appropriateness. In the blind evaluation, the reference-supported scenario was preferred across all five criteria, most strongly for character consistency (88.2%, $\chi^2(2) = 23.06, p < 0.001$) and contextual accuracy (82.3%, $p < 0.001$); teachers also rated aesthetic quality (4.73/5.0) and ethical–pedagogical standard (4.65/5.0) highly.

These results indicate that VisualLit controlled the generation process by supplying “contextual anchors” retrieved from a curriculum-grounded knowledge base, thereby mitigating hallucination and maintaining narrative flow. Framed against the consistency taxonomy of Lin *et al.* [14], the gains are concentrated in the character and spatial dimensions, which is consistent with the design focus of the PPC mechanism.

B. RQ2: Effectiveness of PPC and EPVL

Regarding RQ2, the reference-based generation strategy (PPC) both supported teachers and improved the visual consistency and semantic accuracy of the images. These observations highlight a fundamental limitation of general-purpose generative models when applied to pedagogically sensitive contexts, where accuracy and narrative fidelity are essential. Compared with the baseline No-Reference configuration, the Reference scenarios produced outputs with substantially improved visual consistency, higher semantic alignment with the literary content, and enhanced narrative continuity across image sequences. The static knowledge repository thus acts as a semantic anchor that constrains the generation process and mitigates hallucination effects, affirming that context-grounded prompt augmentation is crucial for creating reliable educational content.

The multi-model comparison further suggests complementary strengths across different generation back-ends. DALL-E3 was preferred for prompt adherence, whereas Nano Banana showed a descriptive, non-significant trend for control of character posture and layout. These findings nonetheless support the design decision to implement a multi-model architecture that can leverage the complementary advantages of different generative back-ends.

A further contribution relevant to RQ2 is the Ethical–Pedagogical Validation Layer, which acts as a safeguard against inappropriate or culturally inconsistent

outputs. The stress-test cases demonstrate the mechanism qualitatively. This work is consistent with the growing consensus that domain-specific safeguards are necessary for the responsible deployment of AI in education. However, because the current EPVL evaluation did not include a sufficiently large held-out set of both risky and benign images with item-level ground-truth labels, these results should be interpreted as mechanism-level evidence rather than as a definitive robustness guarantee.

C. RQ3: Impact on Student Learning

Regarding RQ3, the classroom quasi-experiment indicated a positive influence of AI-generated visual materials on student engagement and learning outcomes: students taught with VisualLit materials showed higher learning interest (+0.93, $p < 0.01$), event-detail recall (+1.25, $p < 0.001$), and overall achievement (+1.15, $p < 0.001$), with large, standardized effects. The clear improvement in symbolic association and appreciation (Variable 4: +1.05 points, $F(1, 76) = 16.43, p < 0.001, d = 0.91$) further suggests that the EPVL-validated materials provided an accurate visual foundation for literary appreciation.

Beyond statistical significance, the magnitude of the gains warrants a measured reading. The overall post-test advantage of the experimental group was +1.15 points on a 10-point scale, approximately a 16.4% relative gain over the control mean of 7.00. This is a more substantial improvement than the typical incremental gains reported for classroom technology interventions, but a one-point shift on a 10-point classroom test still represents an educationally meaningful rather than a transformative effect, and its durability beyond the immediate post-test was not measured. It should also be emphasized that the control condition was instruction without AI-generated visuals, not instruction using conventional, professionally drawn textbook illustrations. The study therefore demonstrates a benefit of VisualLit materials over their absence, not their superiority over existing illustration practice; the latter comparison is identified as a priority for future work. The most defensible reading of the classroom results is that curriculum-grounded, ethically-validated visual materials produced a reliable and educationally meaningful, but bounded, improvement in this single-school sample.

As discussed above, however, these estimates carry wide confidence intervals, derive from a single school, and reflect a comparison against the absence of AI visuals rather than against conventional illustration; they are therefore best read as encouraging preliminary evidence rather than as a settled finding.

D. Limitations and Threats to Validity

Several limitations should be weighed when interpreting these results. First, although the 17-teacher panel was drawn from five secondary schools across two provinces, the classroom quasi-experiment was conducted at a single school with modest samples ($39 + 40 = 79$ students). The classroom design is adequately powered for medium-to-large effects but not for small effects, and the teacher-panel point estimates below 76.5% should be treated as preliminary; multi-site classroom replication is therefore needed before the framework’s effectiveness can be claimed broadly.

Second, the quasi-experiment used intact, non-randomly assigned classes, so selection effects and

selection-maturation interactions cannot be fully excluded despite ANCOVA adjustment for pre-test differences. Third, both conditions were taught by the same teacher; while this controls for teaching style, it introduces a possible expectancy effect. Fourth, the study lacks a non-AI baseline of conventional, professionally drawn textbook illustrations, so it establishes a benefit of VisualLit materials over their absence rather than over existing practice. Fifth, the EPVL evaluation reported here is structured but was carried out with a single annotator set; the data in Table 4 need to be supplemented with quantitative values to increase their reliability. Sixth, the static knowledge base currently covers the middle-school curriculum and does not yet span high-school texts, where literary symbolism and narrative complexity are greater. Seventh, integrating multiple generative models and a validation layer adds computational overhead and latency, which is acceptable for small-scale classroom use but requires optimization for large-scale deployment. Finally, the system generates static images only; richer multimodal representations (animated storyboards or interactive storytelling) were not explored.

E. Theoretical and Pedagogical Implications

The findings contribute to the theoretical foundation for applying GenAI in literature education by showing how visual consistency can be operationalized as a measurable construct rather than treated as a vague aesthetic judgment. By connecting character consistency, spatial consistency, and symbolic/thematic consistency to the narratological taxonomy in Lin *et al.* [14], this study provides a more explicit framework for evaluating AI-generated visual learning materials in humanities education.

Pedagogically, the results suggest that GenAI tools may become more useful to teachers when they are embedded within curriculum-grounded, ethically validated workflows rather than used as general-purpose image generators. VisualLit reduces the technical burden of prompt engineering while preserving the teacher's role in guiding literary interpretation, selecting appropriate outputs, and aligning materials with lesson objectives. The framework therefore supports a human-centered model of AI-assisted material design in which technology augments, rather than replaces, teacher expertise.

VI. CONCLUSION

This study addressed a central challenge in applying generative AI to literature education: creating contextually accurate, pedagogically appropriate, and visually consistent learning materials from literary texts. To this end it proposed VisualLit, a framework that combines context-grounded prompt generation, a structured literary knowledge repository, and an Ethical-Pedagogical Validation Layer to guide and regulate the image-generation process.

The primary contribution is the reference-based prompt-generation mechanism, which enriches prompts with structured contextual parameters so that generative models can maintain narrative consistency and semantic accuracy; consistency is defined explicitly along character, spatial, and symbolic dimensions drawn from a narratological taxonomy. The second contribution is the EPVL safeguard, which integrates pedagogical and cultural constraints into the

generation process. The framework was examined through technical characterization, a teacher-based blind evaluation, and a classroom quasi-experimental study.

Across these phases, the proposed method improved the semantic accuracy, character consistency, and pedagogical suitability of the generated illustrations relative to standard prompt-based generation, and the classroom study indicated positive effects on student engagement and comprehension. These findings are promising but preliminary: they come from a single school with modest samples, wide effect-size intervals, and a comparison against the absence of AI visuals rather than conventional illustration. They nonetheless suggest that combining context-grounded prompt generation with pedagogical validation can increase the practical applicability of generative AI in humanities education. Future work will focus on multi-site replication with randomized assignment and a conventional-illustration baseline, completing and independently validating the EPVL evaluation, expanding the knowledge base to high-school texts, improving scalability, and exploring richer multimodal representations.

ETHICS STATEMENT

The study was conducted in accordance with the ethical guidelines of Hanoi National University of Education. Ethical approval was formally waived by the institution on the grounds that the study involved minimal risk and did not collect sensitive personal data: only oral-test averages from regular instruction, anonymous attitude-survey responses, and curriculum-aligned achievement-test scores were recorded, all in pseudonymized form and analyzed only in aggregate. The quasi-experimental lessons were delivered as part of the standard sixth-grade Literature curriculum, and no instructional content or assessment fell outside the routine pedagogical practice of the participating school. Consistent with this minimal-risk waiver, individual written parental or guardian consent was not collected; the study is therefore best read as an evaluation of minimal-risk educational practice, and prospective written parental/guardian consent together with a formal Institutional Review Board (IRB) review will be obtained for any larger-scale or higher-stakes replication.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

D.T. Trung conceived the study, designed the framework, supervised the work, and wrote the manuscript. N.L.T. Van, H.P.B. Trung, and D.T. Tuyen implemented the system, conducted the experiments, and prepared the initial draft. All authors approved the final version.

ACKNOWLEDGMENT

The authors thank the 17 literature teachers and 79 students for their valuable participation and feedback during the experimental phase of this study.

REFERENCES

- [1] D. T. T. Dung and N. X. Thanh, "Applying Artificial Intelligence (AI) in teaching history in a way that develops problem-solving and creative

- abilities for high school students,” in *Proc. 10th National Scientific Conference for Young Faculty Members of Teacher Training Universities*, 2023, pp. 221–236.
- [2] T. T. H. Phuong and C. T. K. Ngan, “Applying Artificial Intelligence (AI) in the teaching of literature in high schools,” *Journal of Educational Equipment: Applied Research*, vol. 1, no. 308, pp. 173–196, 2024.
- [3] D. T. T. Duong *et al.*, “Applying artificial intelligence to design digital learning resources for flipped classrooms in history teaching,” *Vinh University Journal of Science*, vol. 53, no. 308, pp. 257–266, 2024.
- [4] N. T. H. Anh, P. T. V. Nga, and N. H. P. L. Linh, “Applying Artificial Intelligence (AI) technology in lesson design to enhance teaching quality at the Faculty of English, Hanoi Open University,” *Hanoi Open University Journal of Science*, vol. 1A, 2025.
- [5] Vietnamese Ministry of Education and Training, “General education program—Comprehensive education program,” Circular No. 32/2018/TT-BGDDT, 2018.
- [6] R. N. Hafni, T. Herman, E. Nurlaelah, and L. Mustikasari, “The importance of science, technology, engineering, and mathematics (STEM) education to enhance students’ critical thinking skill in facing industry 4.0,” *Journal of Physics: Conference Series*, vol. 1521, no. 4, 042040, 2020. doi: 10.1088/1742-6596/1521/4/042040
- [7] G.-J. Hwang, H. Xie, B. W. Wah, and D. Gašević, “Vision, challenges, roles and research issues of artificial intelligence in education,” *Computers and Education: Artificial Intelligence*, vol. 1, 100001, 2020. doi: 10.1016/j.caeai.2020.100001
- [8] B. Pepin, G. Gueudet, and L. Trouche, “Refining teacher design capacity: Mathematics teachers’ interactions with digital curriculum resources,” *ZDM Mathematics Education*, vol. 49, no. 5, pp. 799–812, 2017. doi: 10.1007/s11858-017-0870-8
- [9] J. Vigna-Taglianti, “AI-generated images as a teaching tool in foreign language acquisition,” *Technology and Language*, vol. 5, no. 3, pp. 85–105, 2022.
- [10] G. Bansal *et al.*, “Revolutionizing visuals: The role of generative AI in modern image generation,” *ACM Transactions on Multimedia Computing, Communications, and Application*, vol. 20, no. 11, 356, 2024.
- [11] S. Aktay, “The usability of images generated by Artificial Intelligence (AI) in education,” *International Technology and Education Journal*, vol. 6, no. 2, pp. 51–62, 2022.
- [12] I. Kogai, “Automatic generation of illustrations for children’s fairy tales from text: AI-powered app based on DALL-E 3,” in *Proc. CONTENT 2024: The 16th International Conference on Creative Content Technologies*, 2024, pp. 14–19.
- [13] H. Kim, J.-H. Choi, and J. Y. Choi, “A novel scheme for managing multiple context transitions while ensuring consistency in text-to-image generative artificial intelligence,” *IEEE Access*, vol. 12, pp. 150468–150484, 2024. doi: 10.1109/ACCESS.2024.3476933
- [14] Z. Lin, Z. Wang, Q. Liu, X. Zhang, and J. Liu, “Narratology meets text-to-image: A survey of consistency in AI generated storybook illustrations,” *Artificial Intelligence Review*, vol. 59, 83, 2026. doi: 10.1007/s10462-025-11482-6
- [15] S. A. Chatzichristofis, A. Tsopozidis, A. Kyriakidou-Zacharoudiou, S. Evripidou, and A. Amanatiadis, “Designing an AI-supported framework for literary text adaptation in primary classrooms,” *AI*, vol. 6, no. 7, 150, 2025. doi: 10.3390/ai6070150
- [16] Y. He *et al.*, “Automated black-box prompt engineering for personalized text-to-image generation,” in *Proc. 38th Annual Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024.
- [17] Google Developers Blog. Introducing Gemini 2.5 flash image, our state-of-the-art image model. [Online]. Available: <https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/>
- [18] National Assembly of Vietnam, “Law on cybersecurity,” Law no. 24/2018/QH14, 2018.
- [19] National Assembly of Vietnam, “Children’s law,” Law no. 102/2016/QH13, 2016.
- [20] OpenAI. Usage policies. [Online]. Available: <https://openai.com/policies/usage-policies>
- [21] Google. *Generative AI Prohibited Use Policy*. [Online]. Available: <https://policies.google.com/terms/generative-ai/use-policy>
- [22] Vietnamese Ministry of Education and Training (MoET). (2017). Circular no. 33/2017/TT-BGDDT on standards, procedures for compilation, editing, and appraisal of textbooks. [Online]. Available: <https://dulieuphapluat.vn/van-ban/giao-duc-van-ban/circular-no-332017tt-bgddt-dated-december-22-2017-on-the-standards-and-procedures-for-development-and-revision-of-textbooks-standards-for-textbook-developers-organization-and-operation-of-national-textbook-appraisal-councils-1112135.html>
- [23] S. T. Luu, K. V. Nguyen, and N. L.-T. Nguyen, “A large-scale dataset for hate speech detection on Vietnamese social media texts,” in *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices, Lecture Notes in Computer Science*, vol. 12798, Cham, Switzerland: Springer, 2021, pp. 415–426. doi: 10.1007/978-3-030-79457-6_35

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).