Apollo Journals
Social Sciences and Humanities Inquiry
SSHI
Vol. 01 Issue 04 2026

Article

An AI-Mediated "Image-Text-Image" Bidirectional Training Model for Cultivating Visual Imagery in EFL Narrative Writing: A Proposed Pedagogical Framework and Classroom Case Study

Meilida Juman

Urumqi No. 36 Middle School, Urumqi 830000, China; Email: 1337047704@qq.com

Abstract: Story continuation, as a core assessment task in senior high school English, requires learners to construct coherent narratives that maintain consistency with source texts in characterization, setting, theme, and stylistic tone. Nevertheless, many EFL learners produce formulaic, hollow narratives by mechanically applying memorized sentence patterns and writing templates. This pedagogical dilemma is likely to be related, in part, to underdeveloped mental visual imagery prior to linguistic encoding: when students fail to build dynamic, concrete mental scenarios, they tend to fall back on sentence-driven rather than context-driven composition. Vocabulary knowledge, task understanding, and instructional habits may also contribute. Drawing upon dual-coding theory, scaffolding theory, and generative AI text-to-image technology, this paper proposes an "image-text-image" bidirectional training model designed to cultivate students' mental visual imagery in English narrative writing. The model operationalizes a four-step instructional workflow: (a) generating theme-oriented narrative scene scaffolds, (b) conducting layered image-to-text descriptive practice, (c) delivering visual feedback via text-to-image generation, and (d) gradually fading external visual aids to foster internal mental imagery. To examine how the model works in authentic classroom conditions, an exploratory qualitative case study was conducted at a senior high school in Urumqi, China, involving 50 Grade 11 students in Class 3 (25 male and 25 female). Data sources included 50 first drafts and 50 revised drafts, AI-generated images, classroom observation notes covering 12 class periods, and the teacher-researcher's teaching log. A thematic analysis (Braun & Clarke, 2006) of the data suggests that the intervention may strengthen learners' awareness of narrative details, improve the contextual appropriateness of linguistic output, and enhance students' sensitivity toward narrative textual clues. These findings are provisional and context-bound: they indicate the model's potential rather than its generalizable effectiveness. This study contributes a feasible technical-pedagogical framework that bridges visual cognition theory with generative AI tools, offering a practical pathway for senior high school English narrative writing instruction.

Keywords: AI Text-to-Image Technology, English Story Continuation, Mental Visual Imagery, Bidirectional Training, EFL Narrative Writing, Generative AI in Language Education, Qualitative Case Study

Received: 20 September 2026
Revised: 25 October 2026
Accepted: 10 November 2026
Published: 24 November 2026

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

1.1. Background: Story Continuation in Senior High School English Writing

The story continuation task has become a prominent component of senior high school English writing assessment, particularly in the Chinese EFL context. Unlike conventional proposition-based writing, story continuation requires learners to extend a given narrative reading material by constructing coherent plots while maintaining consistency with the source text across multiple dimensions: characterization, setting, theme, and stylistic tone (Ye et al., 2021). Source texts used in such tasks are typically rich in multisensory details, including character gestures, facial expressions, inner emotions, and environmental descriptions, all of which serve vital functions in establishing story contexts, driving plot development, and revealing thematic meanings (Shu & Xiong, 2018).

This task format reflects a broader shift in language assessment toward integrated skills testing, where reading comprehension and writing production are assessed within a single, meaning-driven activity. However, the cognitive demands of story continuation

extend well beyond surface-level linguistic proficiency. Students must first comprehend the source narrative, construct a coherent mental model of the story world, anticipate plausible plot developments, and then translate this mental representation into grammatically accurate and stylistically appropriate English prose. The task therefore places significant demands on both reading comprehension skills and the ability to generate detailed, contextually appropriate written narratives.

1.2. The Problem: Formulaic Writing and Lack of Visual Imagery

In real-world EFL classrooms, a persistent challenge emerges: many students can outline basic plot frameworks yet struggle to produce vivid, specific descriptions. In pursuit of linguistic complexity, learners frequently transplant memorized chunks and complex sentence structures regardless of situational constraints. Consequently, their writing suffers from disjointed context-language matching, flat characterization, and empty narrative content (Shu & Xiong, 2018). Template-reliant composition has become a widespread obstacle in senior high school English writing classrooms, undermining the very purpose of story continuation tasks designed to foster creative and contextualized language use.

From the perspective of writing cognition, a central problem may lie in underdeveloped mental visual imagery before linguistic encoding; limited vocabulary knowledge, task understanding, and instructional habits may also contribute. Rather than simulating story scenes in their minds, many students directly retrieve pre-stored language materials for output. Without situational cognitive scaffolding, learners tend to rely on ready-made templates, yielding vague descriptions and pragmatically inappropriate expressions. Therefore, writing instruction should not merely target vocabulary and grammar; it also needs to cultivate learners' pre-linguistic mental representation capacity (Yi, 2020).

The gap between what students can describe abstractly and what they can concretely visualize represents a critical bottleneck in EFL narrative writing development. Traditional writing pedagogy has lacked effective tools to externalize learners' invisible mental pictures, making it difficult for teachers to diagnose the gap between intended meaning and expressed language.

This challenge is particularly acute in Chinese EFL contexts, where large class sizes and high-stakes assessment pressures often lead to a focus on test-oriented writing strategies rather than on deep meaning-making. Teachers face the dilemma of needing to prepare students for standardized tests while also fostering the kind of creative, context-sensitive writing that story continuation tasks are designed to assess. The bidirectional training model proposed in this paper offers a way out of this dilemma by providing a structured, technology-supported approach that simultaneously builds descriptive language skills and deepens students' understanding of narrative context.

1.3. Purpose and Study Positioning

Drawing upon visual cognition theory and recent advances in generative AI text-to-image technology, this paper constructs an "image-text-image" bidirectional training model to address this pedagogical challenge. The study is positioned as an exploratory qualitative case study that examines the implementation of the model in a real classroom setting, grounding the instructional design in cognitive theory while treating the evidence as provisional. The primary purposes of this paper are: (a) to articulate the theoretical rationale for leveraging AI text-to-image technology in EFL narrative writing instruction; (b) to present a structured bidirectional training model that integrates visual input and verbal output in reciprocal cycles; and (c) to document the implementation of the four-step instructional workflow and to report qualitative evidence of its effects on students' writing development.

This study addresses three interrelated research questions: How can AI text-to-image technology be integrated into EFL narrative writing instruction to cultivate students' mental visual imagery? What pedagogical mechanisms underlie the "image-text-image" bidirectional training model? And what qualitative evidence does a classroom implementation of the model provide regarding students' development of visual imagery and narrative writing?

2. Literature Review

2.1. Story Continuation and EFL Narrative Writing

The story continuation task, originally introduced as an integrated reading-writing assessment format, has attracted growing research attention in Chinese EFL contexts. Ye et al. (2021) reviewed story continuation research in Chinese EFL classrooms, noting that while the task has demonstrated benefits for language development in general, significant challenges persist in helping students produce detailed, coherent narratives. Research has consistently identified that students struggle most with descriptive language and contextual appropriateness, rather than with basic plot construction.

Shu and Xiong (2018) further specified this challenge by identifying micro-skills of action description as a particularly weak area in students' story continuation writing. They observed that students tend to rely on direct emotional adjectives (e.g., happy, nervous, sad) rather than depicting emotions through concrete actions, gestures, and environmental cues. This pattern reflects a

https://doi.org/10.67728/sshi.2026.018

broader tendency toward sentence-level rather than discourse-level writing, where learners prioritize grammatical accuracy and lexical sophistication over contextual relevance and narrative vividness.

The cognitive demands of story continuation place it at the intersection of reading comprehension, writing production, and visual mental imagery. Green and Brock (2000) demonstrated that narrative transportation—the extent to which readers become immersed in a story—depends heavily on the ability to construct vivid mental representations of story events. Their research addresses narrative engagement in first-language reading; its bearing on EFL writing is inferential rather than directly demonstrated, and we draw on it only as a theoretical parallel. EFL learners with limited proficiency may struggle to achieve this transportation effect, particularly when processing texts in a second language, which constrains their ability to generate rich, coherent continuations.

Ye et al. (2021) further noted that while story continuation tasks have been widely adopted in Chinese EFL assessment, the pedagogical support for helping students succeed in these tasks has not kept pace. Many teachers focus primarily on linguistic accuracy and grammatical correctness, while giving insufficient attention to the cognitive and imagistic dimensions of narrative writing. This misalignment between assessment demands and instructional focus contributes to the persistent gap between what students are expected to produce and what they are actually able to write.

Furthermore, the story continuation task imposes unique constraints that make visual imagery particularly important. Unlike open-ended narrative writing, continuation writing requires consistency with an existing story world: characters must behave consistently with their established personalities, settings must remain coherent, and plot developments must be plausible given the established context. These constraints place a premium on the ability to mentally simulate the existing story world and to anticipate how characters and situations might evolve. Students who cannot construct vivid mental models of the source text are at a distinct disadvantage, as they must rely on superficial plot elements rather than deep understanding of character and setting.

2.2. Mental Visual Imagery in Second Language Writing

Mental visual imagery refers to learners' internal mental simulation of settings, character movements, and atmospheres when reading or composing narratives. In first-language (L1) contexts, research on reading has examined the relationship between visual imagery and narrative engagement. Brosch (2018) reviewed research on what readers "see" when reading fictional narratives, showing that visualization and the vividness of mental imagery are closely associated with reading engagement and narrative comprehension. Importantly, Brosch's review highlighted that readers vary considerably in their ability to generate detailed mental images from textual descriptions, and that this variation has measurable effects on how narratives are processed and recalled. As a review of reading research, Brosch's (2018) findings concern readers' mental imagery during L1 narrative reading; they do not by themselves demonstrate that visualization training improves EFL writing, and the connection drawn here is therefore theoretical and provisional.

In second language (L2) contexts, Yi (2020) reported on the effects of visualization training on intermediate EFL learners' narrative writing, suggesting that explicit visualization instruction can help students generate richer narrative details and move beyond rigid sentence-oriented writing. This study provided preliminary evidence that visual imagery training can transfer to L2 writing improvement, although its small sample size and short intervention duration limit the strength of the conclusions.

The dual-coding theory proposed by Paivio (1986) and extended by Mayer (2021) provides a theoretical foundation for understanding the role of visual imagery in writing. Dual-coding theory suggests that textual information processed together with visual mental representations achieves deeper cognitive engagement and richer expressive potential. When learners activate both verbal and visual mental representations simultaneously, they create multiple retrieval pathways and richer associative networks, which support more elaborate and coherent writing production.

However, traditional writing pedagogy has lacked effective tools to externalize learners' invisible mental pictures. While teachers can ask students to "imagine the scene," they cannot directly observe or evaluate the quality of those mental images. This invisibility makes it difficult to diagnose the gap between a student's intended meaning and their expressed language.

2.3. Multimodal Composing and Visual Scaffolding in L2 Education

Recent years have seen growing interest in multimodal composing as a means of enhancing L2 writing development. Lim and Kessler (2024) reviewed research on multimodal composing and second language acquisition, concluding that integrating visual and verbal modes in writing tasks can deepen learners' engagement with meaning-making processes. Multimodal composing activities encourage students to attend to how different modes work together to convey meaning, which can transfer to more nuanced and deliberate verbal writing.

Unsworth and Mills (2020) specifically examined the teaching of attitude and emotion in digital multimodal composition, noting that visual modes provide a concrete medium through which learners can explore affective dimensions of meaning that are often difficult to express purely through language. When students create visual representations alongside verbal text, they become more aware of how language choices convey emotional tone and character perspective.

https://doi.org/10.67728/sshi.2026.018

Kress and van Leeuwen (2021) established a comprehensive grammar of visual design, providing analytical tools for understanding how images construct meaning through composition, color, perspective, and framing. This theoretical framework has been influential in multimodal literacy education, though its application to L2 writing pedagogy remains relatively underexplored.

Scaffolding theory, derived from Vygotsky's (1978) sociocultural perspective and elaborated in the contingent-scaffolding literature (van de Pol et al., 2010), provides a complementary lens for understanding how visual supports can facilitate writing development. The principle of fading—gradually removing external supports as learners internalize skills—has been widely applied in educational contexts, though its application to visual scaffolding in writing instruction has been limited.

Importantly, multimodal composing research has also highlighted the social and collaborative dimensions of meaning-making. When students create visual and verbal texts together, they engage in negotiated meaning-making: discussing what to include, how to represent characters and settings, and how visual elements relate to narrative content. This social interaction around multimodal artifacts can deepen learners' understanding of narrative structure and character development. However, most existing multimodal writing research has focused on student-created visuals—such as drawings, collages, or digital stories—rather than on AI-generated visuals that serve as feedback on existing writing. The bidirectional model proposed in this paper extends multimodal writing pedagogy by using AI-generated images not as creative artifacts in themselves, but as diagnostic feedback tools that reveal gaps in students' descriptive language.

2.4. Generative AI in Language Teaching: Emerging Potentials

Recent advances in generative AI have opened new possibilities for language education. Moody et al. (2026) explored early implementations of AI-generated art in TESOL, finding that AI image generation tools can provide novel forms of visual feedback and creative prompts for language learners. Their exploratory study suggested that converting students' written descriptions into AI-generated images can help learners identify gaps in descriptive precision and detail richness.

Darvin (2025) argued that generative AI-mediated L2 writing makes critical digital literacies necessary, noting both the transformative potentials of these tools and the constraints and risks they introduce, and calling for critical, equity-oriented frameworks to guide classroom use. Kohnke et al. (2023) similarly discussed the affordances and challenges of generative AI for language teachers, emphasizing the need for pedagogically sound implementation. Building on these discussions, Chen (2026) provided empirical evidence from an EFL context showing that AI-generated imagery can reduce learners' writing anxiety and increase their engagement and motivation, which lends initial support to the visual-feedback approach examined in the present study.

The growing body of research on AI writing feedback has largely focused on generative text models—such as large language models—that provide linguistic feedback on grammar, vocabulary, and structural coherence. Comparatively little attention has been paid to multimodal AI tools that can provide visual feedback on the semantic and imagistic content of writing. This gap is significant because many of the challenges students face in narrative writing are not primarily linguistic but cognitive: they struggle to visualize the scenes they are trying to describe, and their written output reflects this impoverished mental representation. Traditional text-based feedback cannot directly address this gap, as it operates at the verbal level without access to the underlying mental imagery that drives writing.

Despite these emerging discussions, the specific application of AI text-to-image technology for cultivating visual mental imagery in narrative writing remains under-explored. Existing research on multimodal writing has largely focused on student-created visuals or static images provided by teachers, rather than on the bidirectional cycle where AI generates images from student text and students revise their text based on AI-generated visual feedback.

2.5. Research Gap and Study Positioning

The reviewed literature reveals several key gaps that the present study addresses. First, while visualization training has been shown to benefit L2 narrative writing, existing pedagogical approaches lack tools to dynamically externalize and provide feedback on students' mental imagery. Second, although generative AI is increasingly discussed in language education, few studies propose structured pedagogical models that integrate text-to-image generation into writing instruction in theoretically grounded ways. Third, the bidirectional relationship between visual input and verbal output—where images inform writing and writing informs visual feedback—has not been systematically operationalized as a training model.

This study positions itself at the intersection of these gaps, proposing an "image-text-image" bidirectional training model that leverages AI text-to-image technology to make invisible cognitive processes visible. The model is grounded in dual-coding theory and scaffolding theory, and is examined through an exploratory qualitative case study of its implementation in a senior high school English writing classroom.

https://doi.org/10.67728/sshi.2026.018

3. Theoretical Framework

3.1. Dual-Coding Theory and Visual-Verbal Integration

Dual-coding theory, originally proposed by Paivio (1986) and further developed by Mayer (2021) in the context of multimedia learning, posits that humans process information through two distinct but interconnected systems: a verbal system for language processing and a nonverbal system for visual and spatial representation. When learners encode information through both systems simultaneously, they create dual mental representations that are more durable and more readily accessible for retrieval and production.

In the context of narrative writing, dual-coding theory suggests that the ability to generate vivid visual mental images from textual descriptions, and conversely to translate visual mental images into precise verbal descriptions, depends on the strength of connections between the verbal and visual systems. EFL learners who struggle with descriptive writing may have weak cross-system connections: they can access verbal knowledge but struggle to activate corresponding visual representations, or they can visualize scenes but struggle to translate those visuals into precise language.

The "image-text-image" bidirectional training model is designed to strengthen these cross-system connections through reciprocal practice. In the image-to-text phase, learners move from visual representations to verbal descriptions, exercising their ability to translate visual impressions into language. In the text-to-image phase, they move in the opposite direction: their verbal descriptions are converted into images, allowing them to compare their intended visual representation with the AI-generated output and identify gaps in their descriptive language. This bidirectional cycling strengthens both directions of the visual-verbal connection.

Cognitive load theory provides a complementary perspective on why this dual-coding approach is effective for EFL learners. Working memory has limited capacity, and when learners must simultaneously imagine a scene, recall vocabulary, and construct grammatically correct sentences, cognitive overload can result. By providing an external visual representation, the image-to-text phase offloads the scene-construction component to the visual channel, freeing up cognitive resources for language production. Similarly, in the text-to-image phase, the AI-generated image serves as an external cognitive aid that allows learners to compare their intended meaning with their expressed language without having to mentally simulate both versions simultaneously. This distribution of cognitive load across multiple channels aligns with the multimedia learning principles identified by Mayer (2021), which demonstrate that learning is enhanced when information is presented in complementary visual and verbal formats.

3.2. Scaffolding Theory and Fading Instruction

Scaffolding theory, rooted in Vygotsky's (1978) concept of the zone of proximal development, emphasizes that learners can accomplish more with guided support than they can independently. Effective scaffolding involves providing temporary support tailored to the learner's current level, then gradually withdrawing that support as competence develops (van de Pol et al., 2010).

In the proposed model, visual images serve as cognitive scaffolding for narrative writing. When students struggle to generate detailed descriptions, a concrete visual image provides an external representation that they can observe and describe, reducing the cognitive load of simultaneously imagining the scene and finding the words to describe it. This external support allows learners to focus on developing descriptive language skills without the additional burden of constructing the mental scene from scratch.

However, the ultimate goal is not to have students depend on external images permanently. Consistent with scaffolding theory's fading principle, the model includes a gradual withdrawal of visual supports as students develop their internal mental imagery capabilities. The four-step implementation sequence described later in this paper operationalizes this fading process: starting with explicit image support, moving through guided image-based practice, then to visual feedback on self-generated text, and finally to independent mental imagery construction without external visual prompts.

3.3. From Visual Input to Expressive Output: A Bidirectional Model

Combining dual-coding theory and scaffolding theory, the proposed model conceptualizes visual imagery development as a bidirectional process. Most existing writing pedagogy focuses primarily on the ideation-to-text direction—helping students translate their ideas into written language without explicit attention to visual imagery. The proposed model adds two reciprocal directions: image-to-text practice, in which students observe and describe visual scenes, and text-to-image feedback, in which students' verbal descriptions are converted into AI-generated images for comparison and revision; the two directions play complementary roles in developing descriptive writing skills. Within this framework the three theoretical strands serve distinct purposes: dual-coding theory explains why verbal-visual integration may deepen encoding, multimedia learning principles (Mayer, 2021) specify how external visual-verbal presentation can manage cognitive load, and scaffolding theory (Vygotsky, 1978) explains why and how external supports should be faded; they are complementary rather than interchangeable.

https://doi.org/10.67728/sshi.2026.018

4. The "Image-Text-Image" Bidirectional Training Model

4.1. Model Rationale and Core Principles

The "image-text-image" bidirectional training model is built upon three core principles derived from the theoretical framework outlined above. First, visual imagery is treated as a teachable and trainable cognitive skill rather than a fixed ability. Second, the connection between visual perception and verbal expression is strengthened through reciprocal, cyclic practice rather than one-way instruction. Third, external visual supports are introduced and faded in a structured sequence to support internalization of mental imagery capabilities.

Two related constructs should be distinguished at the outset. Mental visual imagery refers to the internal, pre-linguistic representation that learners construct (for example, a scene imagined from textual clues); situational comprehension refers to understanding the story world built from the source text; descriptive ability refers to the linguistic skill of rendering that representation in words; and the AI-generated image is an external artifact that approximates, but never fully captures, any of these three constructs. The model treats the AI image strictly as an external proxy used for comparison, not as a measure of the learner's internal imagery.

Throughout the paper, the development of students' internal mental imagery is treated as a proposed mechanism that the model is designed to foster. It was not directly measured in this study; what the classroom evidence can document is change in the external artifacts of students' writing and in their engagement with generated images, from which changes in imagery are inferred cautiously.

The model addresses a key limitation of traditional writing instruction: the invisibility of mental processes. By using AI text-to-image technology to externalize students' intended meanings as visual outputs, the model creates a feedback loop that makes the gap between intended meaning and expressed language visible to both students and teachers. This externalization enables targeted diagnosis and revision, moving beyond generic grammar correction to focus on semantic precision and descriptive completeness.

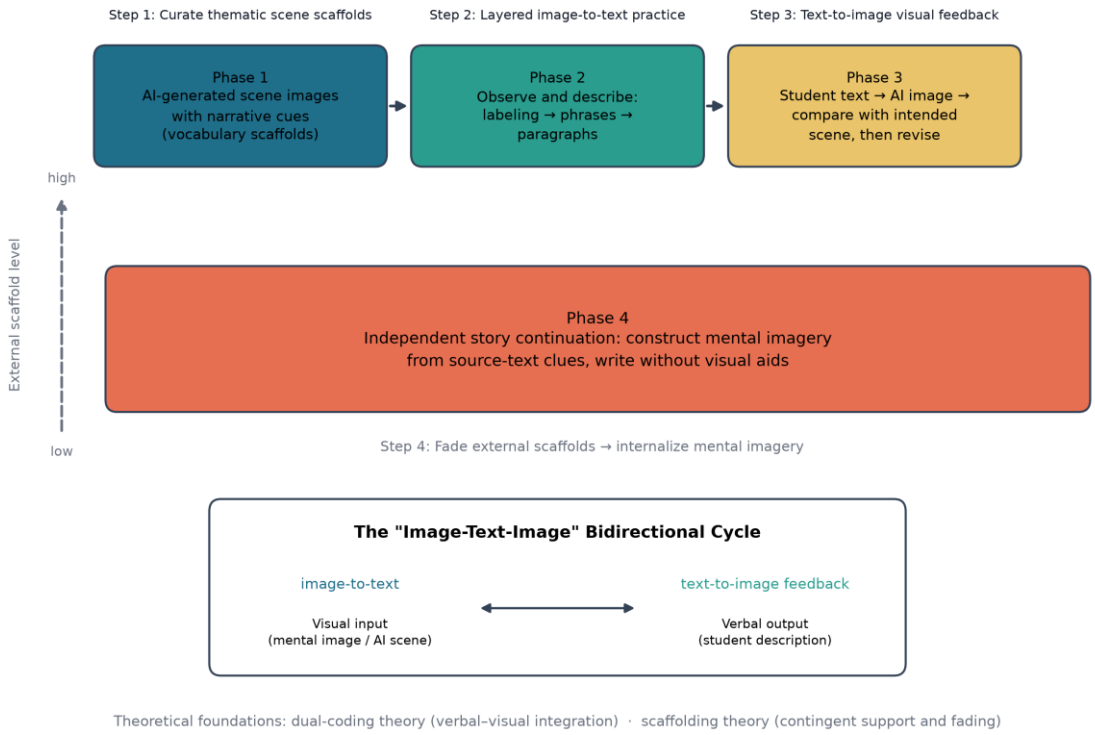


Figure 1. Conceptual diagram of the "image-text-image" bidirectional training model. Source: author.

Figure 1 presents a conceptual diagram of the proposed model, showing the four phases, the corresponding four implementation steps, the bidirectional cycle, and the scaffold-fading sequence.

4.2. Phase 1: Image-to-Text (Visual Input → Verbal Description)

The image-to-text phase involves students observing AI-generated narrative scenes and producing detailed English descriptions. This phase trains students' sensitivity toward multi-dimensional narrative details by directing their attention to specific visual elements: character appearance and posture, facial expressions, spatial relationships, lighting and color, and overall emotional atmosphere.

https://doi.org/10.67728/sshi.2026.018

The pedagogical rationale for starting with image-to-text is twofold. First, observing a concrete image reduces the cognitive load of simultaneously imagining a scene and finding words to describe it, allowing students to focus on developing descriptive vocabulary and sentence patterns. Second, structured observation of visual details helps students become more attentive to narrative elements in general, which transfers to their reading comprehension of source texts and their ability to extract narrative clues for story continuation.

This phase is implemented through layered description activities, progressing from labeling individual elements (characters, actions, environment) to expanding those fragments into coherent descriptive paragraphs. Guiding questions help orient observation and ensure that students attend to multiple dimensions of the scene rather than focusing on superficial elements.

4.3. Phase 2: Text-to-Image (Verbal Output → Visual Feedback)

The text-to-image phase reverses the direction: students' written texts are fed into AI image generators, and the resulting visuals are compared against expected story scenes. Learners identify descriptive deficiencies by examining mismatches between the AI-generated image and what they intended to describe, then revise their compositions accordingly.

This phase serves as a form of visual feedback that complements traditional written feedback. While teachers can comment on grammar, vocabulary choice, and overall coherence, they cannot directly assess whether a student's description successfully conveys the intended visual scene. AI text-to-image generation offers a useful visual heuristic for making the semantic content of student writing observable. It must be emphasized, however, that the AI-generated image is not an objective diagnostic of writing quality: image generation is probabilistic, and a mismatch between the generated image and the intended scene may reflect prompt ambiguity, tool limitations, or stylistic variation in the AI model rather than deficiencies in the student's writing. AI outputs should therefore be treated as discussion starting points for comparison and revision, not as authoritative verdicts.

The pedagogical value of this phase lies in its immediacy and specificity. Students receive instant visual feedback on their writing, and the feedback is concrete rather than abstract. Rather than being told "your description is vague," students can see for themselves that the AI-generated image lacks the character's nervous expression, or that the spatial relationship between characters is unclear. This concrete feedback leads to more meaningful revision than generic written comments.

Equally important is the converse case: a text can be descriptively adequate even when the AI-generated image differs from the expected scene. For example, a student who writes "The boy crept cautiously toward the door, one hand reaching for the handle" may receive an image in which the boy appears startled rather than cautious, because the generator foregrounded the hand gesture and misread the body posture. In such cases, the discrepancy signals a limitation of the AI tool or of the prompt wording rather than a writing deficiency, and teachers should guide students to consider multiple generated outputs before deciding whether revision is warranted. Treating any single image mismatch as an error would be pedagogically misleading; the comparison exercise is valuable precisely because it invites discussion of what the language does and does not specify.

4.4. Scaffold Fading and Internalization

The bidirectional training model incorporates a structured fading process, consistent with scaffolding theory. In the early stages of training, students rely heavily on externally provided images as observation prompts and feedback tools. As they develop descriptive skills and visual awareness, external supports are gradually withdrawn.

The fading process follows a developmental sequence: (a) teacher-provided images for guided observation and description, (b) AI-generated images based on student writing for comparison and revision, (c) student-generated mental imagery without external visual aids, and (d) independent application of visual imagery skills to authentic story continuation tasks. The ultimate objective is to help students internalize the habit of constructing situational context before selecting linguistic resources, rather than relying on external picture prompts.

5. Instructional Implementation: Four-Step Workflow

Table 1 summarizes the four steps of the workflow, specifying the teacher's actions, the students' activities, the role of AI, the level of external scaffolding at each step, and the indicative criteria for fading the scaffolds.

https://doi.org/10.67728/sshi.2026.018

Table 1. The four-step instructional workflow and its components. Source:

Step	Teacher Action	Student Activity	Role of AI	Scaffold Level	Indicative Fading Criteria
Step 1: Curate thematic narrative scene scaffolds	Select or prompt AI-generated scene images rich in narrative cues; provide thematic vocabulary and phrase scaffolds	Observe and label characters, actions, settings, and atmosphere	Generates scene images from teacher-designed prompts	High (image plus vocabulary scaffolds)	Move on when students can identify and label most key narrative elements without prompting
Step 2: Layered image-to-text practice	Model labeling-to-phrase-to-paragraph expansion; guide observation with questions	Produce descriptive phrases, then coherent paragraphs	Provides concrete visual input for description	High to medium	Move on when students produce multi-clause descriptive paragraphs with specific details without scaffolding
Step 3: Text-to-image visual feedback	Select representative drafts; input them into the AI tool; lead comparison and revision discussions	Compare the generated image with the intended scene; revise the draft; iterate	Converts student text into images for comparison	Medium	Fade when students can self-identify missing details before viewing the image
Step 4: Fade scaffolds and internalize imagery	Withdraw external images; prompt mental-visualization strategies	Construct mental imagery from source-text clues; write independently	No longer used for feedback	Low	Withdraw fully when students consistently construct detailed mental scenes before writing

5.1. Step 1: Curate Thematic Narrative Scene Scaffolds

The first implementation step involves teachers creating or selecting narrative scene images that match unit themes or common story-continuation contexts. Suitable contexts include wilderness adventures, interpersonal conflicts, campus events, family interactions, and other scenarios commonly encountered in EFL narrative writing tasks.

Generated images should be carefully selected or prompted to contain informative narrative cues rather than plain static figures. Key visual elements to include are: lighting conditions, color tones, character postures and gestures, facial expressions, spatial relationships between characters and objects, and overall emotional atmosphere. These multi-dimensional cues provide rich material for descriptive practice and help students learn to attend to multiple narrative details simultaneously.

In addition to visual scaffolds, teachers provide thematic vocabulary lists and functional phrase scaffolds to support descriptive tasks. These language supports are tailored to the proficiency level of students and the specific narrative context, and are gradually reduced as students develop greater descriptive vocabulary. The combination of visual and linguistic scaffolds provides comprehensive support that addresses both the cognitive and linguistic dimensions of descriptive writing.

When designing AI image prompts, teachers should be explicit about the narrative elements they want students to practice describing. For example, rather than prompting the AI to generate a generic "school scene," a more effective prompt might specify: "Two high school students standing in a school hallway, one looking excited and gesturing, the other looking hesitant and looking at their watch, lockers on both sides, fluorescent lighting, after-school atmosphere." This level of specificity ensures that the generated image contains rich descriptive material that targets the particular narrative elements students are learning to describe. Over time, teachers can increase the complexity of the prompts, adding more characters, more nuanced emotional expressions, and more complex spatial arrangements.

5.2. Step 2: Layered Image-to-Text Description Practice

Under teachers' guidance, students engage in layered description activities that progress from basic element identification to coherent paragraph construction. Initially, students label core elements in short phrases: identifying characters, their actions, environmental features, and notable details. This initial labeling stage ensures that students have carefully observed the image and can identify key narrative elements before attempting to produce extended description.

Guiding questions help orient observation and direct students' attention to dimensions they might otherwise overlook: Who are the characters? Where are they situated? What gestures and facial reactions do they exhibit? How do environmental details shape emotions and foreshadow plots? These questions scaffold students' observation by modeling what to look for in narrative scenes.

Fragmented notes from the labeling stage are then expanded into coherent descriptive paragraphs. Students learn to connect individual elements into flowing prose, using appropriate conjunctions and discourse markers to establish relationships between characters, actions, and settings. This step translates abstract narrative-writing strategies into observable, operable observation-expression tasks that students can practice systematically.

To illustrate this layered process, consider a classroom example centered on a wilderness adventure scene. At the labeling stage, students might produce short phrases such as: "two hikers," "dense forest," "sunlight through trees," "worried expression,"

"map in hand." These fragments are then expanded into a paragraph: "Two hikers stood at the edge of a dense forest, peering at a crumpled map in one of their hands. Slanting sunlight filtered through the tall pine trees, casting long shadows across the path. The older hiker's brow was furrowed with worry, while the younger one glanced nervously over her shoulder as if uncertain about what lay ahead." This progression from isolated elements to integrated description mirrors the cognitive process that skilled narrative writers employ intuitively, but which novice writers rarely execute explicitly.

The layered structure—from words to phrases to paragraphs—ensures that students build descriptive skills incrementally rather than being asked to produce full descriptions immediately. This incremental approach aligns with scaffolding principles, providing appropriate challenge at each stage while ensuring that students experience success and build confidence.

5.3. Step 3: Visual Feedback through Text-to-Image Generation

In the third step, teachers select representative student writing samples and input them into AI text-to-image tools to generate corresponding visuals. The original reference image (or the expected scene) and the AI-produced output are displayed side by side for comparison and discussion.

Classroom discussion focuses on semantic clarity and descriptive precision rather than grammatical accuracy. Key discussion questions include: Are character identities and action agents unambiguous? Are spatial relationships sufficiently described? Is critical environmental or emotional information missing? Does the generated image convey the mood the writer intended? These questions direct students' attention to the gap between what they intended to say and what their language actually conveys.

Students then revise their texts with the goal of improving semantic precision and descriptive completeness, rather than only correcting grammatical errors. Where conditions permit, students iterate the "write-generate-compare-revise" cycle multiple times, each time refining their descriptions and observing improvements in the correspondence between their writing and the AI-generated visuals.

To make this feedback process concrete, consider an example from classroom practice. A student writes: "The man was angry and shouted at his son." When this description is fed into an AI image generator, the resulting image might show a generic adult male shouting, with little indication of the relationship between the characters, the setting, or the specific nature of the anger. Through guided comparison, the student identifies that their description lacks key details: How old is the man? How old is the son? Where are they? What is the man's body language? Is he pointing a finger? Is his face red with rage? Does the son look afraid, defiant, or ashamed? The student then revises: "The middle-aged father stood in the middle of their small kitchen, his face flushed with anger, his index finger jabbing at the air as he shouted at his teenage son, who stood with his shoulders hunched, staring at the floor." When this revised version is re-input to the AI generator, the resulting image is significantly more specific and closely matches the intended scene. This iterative process makes abstract feedback about "being more specific" concrete and visible.

This iterative cycle provides a powerful form of formative assessment. Both teachers and students can see concretely which aspects of description need improvement: if the AI consistently fails to represent emotional tone, students know they need to add more affective description; if spatial relationships are unclear, students practice prepositional phrases and location descriptions. The visual nature of the feedback makes these gaps immediately obvious, in contrast to traditional written feedback that may be more abstract.

To prevent generation artifacts from being misread as writing deficiencies, the comparison procedure distinguishes three cases before any decision to revise. First, if the generated image omits or distorts an element that the student's wording specifies (e.g., 'the old man smiled' yields an image without a smile), the wording may be under-specified or ambiguous, and revision is appropriate. Second, if the wording intentionally leaves an element implicit (e.g., the student deliberately withholds a detail for narrative effect), the image should not be taken as evidence that the writing is deficient; teacher and peers discuss the intended meaning rather than demanding explicitness. Third, if the wording is clear and complete yet the image still diverges, the discrepancy is attributed to the AI tool's limitations or to unsupported additions, and no revision of the writing is warranted; generating multiple outputs or re-prompting can help confirm this diagnosis. Agreement among several generated images does not by itself validate an interpretation of the student's writing; the source text, the student's stated intention, and teacher judgment remain central. Section 4.3 illustrates the converse case with a student text that is descriptively adequate even though the generated image differs from the expected scene.

5.4. Step 4: Fade External Visual Scaffolds and Internalize Mental Imagery

After sufficient practice with visual prompts and feedback, teachers gradually remove externally provided images. When completing story-continuation assignments, students are asked to independently construct mental imagery from source-text clues: they imagine story locations, lighting, sounds, character spatial positions, gestures, and potential subsequent actions before they begin writing.

Teachers can use prompting strategies to encourage this internal visualization: asking students to close their eyes and "watch the movie" of the story in their minds before writing, or asking them to draw quick mental sketches of key scenes before putting

them into words. These metacognitive strategies help students become more aware of their own mental imagery processes and more intentional about constructing vivid mental scenes.

The ultimate objective is to help students form a sustainable cognitive habit: construct situational context → observe details → generate intended meanings → select appropriate linguistic resources. Rather than relying on external picture prompts, students internalize the visual imagery construction process and apply it independently to new writing tasks. This internalization represents the long-term goal of the training: not just better description with visual aids, but improved narrative writing ability that persists when visual aids are removed.

5.5. A Complete Worked Example of the "Image-Text-Image" Cycle

To illustrate how the four steps are integrated in practice, this subsection presents a complete worked example, from source-text clues to a revised continuation. The example draws on a real task from the case-study classroom; the excerpts are anonymized and lightly normalized for spelling and grammar.

Step 1—Source text and clues. The teacher selected a story-continuation prompt in which a Kazakh girl, Ahema, finds an injured Han botanist, Mr. Li, near her family's yurt at the Tianshan foothills; her family shelters him, and the two form a friendship symbolized by the red box of seeds he gives her before leaving. Students were guided to identify narrative clues such as Ahema's actions of rushing to help and caring for Mr. Li, the yurt-and-stream setting, Mr. Li's pale, pained face, and the emotional significance of the red box of seeds as a token of friendship.

Step 2—Initial draft. A focal student produced the following initial continuation: "Ahema opened it, found it was filled with seeds. With curious eyes, Ahema looked at the seeds, then at Mr. Li, who smiled warmly and explained, 'These are special seeds from Jiangsu.'".

Step 3—Generation prompt and visual feedback. The draft was converted into a text-to-image prompt of the form A Kazakh girl named Ahema kneeling beside a stream at the foot of the Tianshan Mountains, opening a small red box filled with seeds, an elderly Han man with a pale, pained face smiling warmly at her, a yurt with smoke curling from the chimney in the background, crisp autumn morning light; the AI tool Jimeng AI v4, accessed in September 2026, generated the image, which was displayed side by side with the expected scene described in the source text. In the comparison discussion, the student noticed that the generated image showed only a generic box of seeds with no facial expression on either character, no clear spatial relationship between Ahema and Mr. Li, and none of the yurt-and-stream setting that the source text had foregrounded.

Step 4—Revision. After the discussion, the student revised the draft as follows: "Mr. Li smiled warmly, explaining, 'These special seeds from Jiangsu—plant them come spring. They'll grow into something beautiful—a symbol of our friendship.' Ahema held the box tightly, her heart swelling with gratitude." The revised text was re-generated, and the resulting image now showed both characters' facial expressions and the yurt-and-stream setting that had previously been missing, indicating that the added detail had become concrete enough to specify a scene. Table 2 summarizes the complete cycle for this example.

This revision preserved characterization, plot plausibility, theme, and tone rather than merely making the image more specific. Characterization: Mr. Li retains his role as a gentle, grateful botanist, and Ahema's curiosity gives way to understanding. Plot plausibility: the revised dialogue explains the origin of the seeds (Jiangsu) and their future use, which makes the earlier gift concrete and credible. Theme and tone: the friendship symbolism of the red box of seeds, previously latent, is now voiced by Mr. Li and felt by Ahema, while the register stays simple and age-appropriate for the narrator. The example is drawn from anonymized classroom material; minor spelling and grammar were normalized in transcription, and it is presented to illustrate typical patterns rather than as statistically representative evidence.

https://doi.org/10.67728/sshi.2026.018

Table 2. Summary of the complete worked example (compiled from the real classroom materials).

Cycle Stage	Student Text/Action	Notes and Evidence
Source-text clues	Kazakh girl Ahema; injured Han botanist Mr. Li; yurt and stream setting; red box of seeds as a token of friendship	Character, setting, and emotional cues identified by the student
Initial draft	Ahema opened it, found it was filled with seeds. With curious eyes, Ahema looked at the seeds, then at Mr. Li, who smiled warmly and explained, 'These are special seeds from Jiangsu.'	e.g., vague emotion, missing spatial or environmental detail
Generation prompt	A Kazakh girl named Ahema kneeling beside a stream at the foot of the Tianshan Mountains, opening a small red box filled with seeds, an elderly Han man with a pale, pained face smiling warmly at her, a yurt with smoke curling from the chimney in the background, crisp autumn morning light	AI tool: Jimeng AI v4, September 2026
Generated image and feedback discussion	The image rendered the box but omitted the characters' expressions and the setting; the discussion centered on adding facial-expression, posture, and environmental cues before regenerating.	Mismatch observed: the generated image omitted the warm smile and the yurt-and-stream setting; the spatial relationship between the two characters was unclear
Revised draft	Mr. Li smiled warmly, explaining, 'These special seeds from Jiangsu—plant them come spring. They'll grow into something beautiful—a symbol of our friendship.' Ahema held the box tightly, her heart swelling with gratitude.	Improvement: abstract emotion replaced by concrete, observable detail; friendship symbolism made explicit.

6. Research Design and Methods

Because the reviewers requested explicit methodological documentation, this section specifies the design, context, participants, intervention, data, and analysis procedures of the study. The account is deliberately transparent about what was done and what remains to be reported, so that the findings in Section 7 can be evaluated against the evidence actually collected.

6.1. Research Design

The study employed an exploratory qualitative case study design (Yin, 2018), chosen because the research questions concern how the "image-text-image" model operates in an authentic classroom and what evidence of student development it affords. A single instrumental case was examined: one intact class implementing the model over a full instructional cycle. The design prioritizes depth of description and ecological validity over statistical generalizability.

6.2. Context and Participants

The study was conducted at Urumqi No. 36 Middle School, a public senior high school in Xinjiang, China. Participants were 50 Grade 11 students (25 male and 25 female) in Class 3, whose English proficiency was mid-to-upper intermediate. The class was selected because it was a regular parallel class and the homeroom teacher was highly cooperative, which allowed the intervention to run in an authentic, non-selective classroom. The teacher-researcher, who was also the first author and has 10 years of English teaching experience, implemented the model in the regular English writing curriculum.

Judgements about change in students' writing were based on systematic comparison of the 50 first drafts and 50 revised drafts, triangulated with classroom observation notes and the teacher-researcher's reflective log; the analytic procedure is described in Section 6.5. The excerpts and the worked example in Sections 5.5 and 7 were drawn from the participating students' anonymized drafts. Items were selected because they illustrate patterns that recurred across the corpus; they are not claimed to be statistically representative.

6.3. The Intervention

The intervention followed the four-step workflow described in Section 5. It was implemented over six weeks, comprising 12 (two per week) regular 40-minute class periods, with one complete write-generate-compare-revise cycle per unit. All instructional materials, prompts, and scaffold-fading decisions were documented in a teaching log.

6.4. Data Sources and Collection

Four main data sources were collected: (a) students' first and revised drafts for each continuation task (50 first drafts and 50 revised drafts); (b) the AI-generated images produced from students' texts, together with the prompts used (Jimeng AI v4, accessed in September 2026); (c) classroom observation notes and audio-recorded feedback discussions (covering 12 class periods); and (d)

the teacher-researcher's reflective log. No formal semi-structured interviews were conducted in this cycle; classroom observation notes and the reflective log served as the complementary qualitative sources.

6.5. Data Analysis

Drafts, images, and observation notes were analyzed thematically following Braun and Clarke's (2006) six-phase procedure. In phase one, the researcher read all texts and viewed all generated images repeatedly to become familiar with the data. In phase two, initial codes were generated at the level of descriptive episodes (for example, "concrete gesture replaces emotion adjective", "spatial detail added after image comparison"). In phase three, codes were collated into candidate themes, which were refined in phases four and five against the full data set. Coding was performed by the teacher-researcher, and the candidate themes were reviewed and refined through peer debriefing with a colleague experienced in qualitative EFL research. The analysis was primarily interpretative and iterative, moving between individual texts and the complete corpus.

6.6. Trustworthiness

Trustworthiness was addressed through triangulation of the multiple data sources (first and revised drafts, AI-generated images, classroom observation notes, and the teacher-researcher's reflective log), peer debriefing with a colleague familiar with EFL writing instruction, and a documented audit trail of the analytic decisions. No formal member checking was conducted in this cycle; credibility rests primarily on data-source triangulation and the audit trail. The study does not claim statistical generalization; instead, it offers a thick, context-rich account that allows readers to judge transferability to comparable classrooms.

6.7. Ethical Considerations

Ethical approval and informed consent were obtained through the school's local procedures, with written informed consent from the school administration and the students' parents or guardians. All student names were replaced with pseudonyms, and any identifying information was removed from excerpts. Because students' written texts were processed by third-party AI image-generation services, the purpose and implications of this processing were explained to students and their guardians, and separate consent to use anonymized samples for AI processing was obtained from the students' parents/guardians and the school administration. Raw data are stored securely and are accessible only to the research team.

7. Findings

This section reports the findings of the case study as three interrelated themes. The themes synthesize evidence across drafts, generated images, classroom discussions, and the teaching log; representative evidence is quoted where available. All excerpts are presented in the students' original wording, with minor spelling normalization.

The themes below combine what was directly observed in the classroom data (drafts, AI-generated images, observation notes, and the teaching log) with the teacher-researcher's interpretation of those observations; interpretive claims are phrased as such (e.g., 'appeared to', 'was associated with'). The reported changes cannot be attributed solely to the image-text-image cycle: vocabulary scaffolds, teacher questioning, and repeated revision may also have contributed, and the single-case design does not permit causal attribution. In particular, the development of internal mental imagery is treated as a proposed mechanism, not as a directly measured outcome; what the data document is change in students' written texts and in their engagement with generated images.

7.1. Theme 1: Enhanced Awareness of Narrative Detail

The first theme concerns students' growing tendency to depict emotions and actions through concrete, observable detail rather than through isolated emotion adjectives. In early drafts, students frequently ended a scene with an isolated emotion adjective, naming an emotional state without any observable cue; by the later cycles, the same students produced descriptions such as "Seeing the little trees swaying gently, his face brightened with a wide smile", in which the emotion is carried by observable detail. This shift from telling to showing was noted by the teacher-researcher across most of the 50 participants' draft pairs during the analysis described in Section 6.5.

The text-to-image feedback loop appeared to reinforce this shift. When a student's use of a vague emotion adjective produced an image in which the intended emotional state was absent, the ensuing discussion directed attention to observable cues (facial expression, posture, gesture, environment) that the AI tool could render. In the first comparison session, the observation notes record that students pointed out that the AI-generated image rendered only the box and the seeds, with no facial expression, posture, or setting cues; the discussion that followed prompted students to name such observable cues before revising. Over successive cycles, students increasingly added such cues at the drafting stage rather than only after feedback.

This theme is consistent with Brosch's (2018) discussion of visualization and vividness in narrative reading, and it extends that discussion to the production side: under the conditions of this case study, repeated image-to-text and text-to-image practice was

associated with more vivid, detail-rich descriptions. The direction of the relationship cannot be established from a single case, and the theme should be read as an observed pattern rather than a demonstrated effect.

7.2. Theme 2: Improved Contextual Appropriateness of Language Use

The second theme concerns a change in how learners deployed memorized vocabulary and sentence patterns. Rather than abandoning language accumulation, students became more selective: expressions that clashed with character identity or story context decreased in frequency over the intervention. For example, a student who initially expressed the story's symbolic meaning in grammatically garbled, translation-like phrasing produced, after AI-assisted revision, a passage in which the language was contextually appropriate and idiomatic (draft: 'This gift from the Tianshan foothills was not merely a seed, but it was also a symbol...'; revised: 'This gift from the Tianshan foothills was not merely seeds; it symbolized the friendship between Han and Kazakh peoples, showing how love between different groups grows strong like trees in one forest').

The AI feedback cycle made such expression mismatches concrete: the garbled, literal-translation phrasing in the draft was corrected and re-expressed in the revised text, and the discussion that followed focused on why the language and the intended meaning had diverged. For example, when the draft sentence was generated as an image, the result showed only a single seed packet with no suggestion of trees growing in one forest; the discussion focused on making the friendship-and-forest metaphor explicit, and the student revised the sentence accordingly. Students then revised toward contextually suitable language.

This theme speaks to a common critique of template-based instruction—that learners apply formulas without understanding when they fit. Within the boundaries of this single case, the model appeared to support students in connecting language choice to concrete imagined scenarios; the extent to which this transfer persists beyond the intervention remains an open question.

7.3. Theme 3: Heightened Sensitivity to Narrative Textual Clues

The third theme concerns students' reading behavior. Over the intervention, students showed increased attention to setting, object, and behavioral cues in the source texts of continuation tasks. In the final feedback discussion, when asked how they approached the continuation task, several students reported that they first constructed the scene in their minds from the source-text clues (e.g., the yurt, the stream, and the red box of seeds) before writing; no formal interviews were conducted in this cycle. Rather than extracting plot facts only, students described constructing holistic mental story worlds that included sensory details and emotional atmosphere.

This apparent transfer—from training on images to reading narrative text—reflects the interconnected nature of reading and writing. As students practiced building mental models from images, they reported applying the same attentional strategies when reading source texts. In the later continuation tasks, students' drafts showed increasing use of setting and object cues from the source text (e.g., the yurt-and-stream setting and the red box of seeds), and the observation notes record students rereading the source text to locate such clues before drafting.

This theme aligns with Green and Brock's (2000) work on narrative transportation, which links vivid mental modeling to narrative engagement. In this case study, the model appeared to strengthen students' deliberate visualization of story scenes, whether the scene was presented as an image or described in text. As with the other themes, this pattern is suggestive rather than conclusive.

7.4. Caveats and Practical Considerations

It is important to acknowledge several caveats and practical considerations when implementing this model. First, AI image generation carries inherent randomness. Discrepancies between generated visuals and expected scenes are not automatically evidence of poor student writing. Teachers are advised to generate multiple image outputs from the same student prompt for comprehensive evaluation, and to treat AI-generated visuals as discussion starters rather than definitive assessments of writing quality; it should also be noted that multiple consistent outputs do not constitute independent verification of writing quality—they show only that the tool was not confused by the prompt.

Second, the model requires access to AI text-to-image tools, which may present practical challenges in some educational contexts. Schools with limited technology infrastructure may need to adapt the model, for example by using teacher-generated images rather than student-specific AI outputs, or by using storyboarding activities as a bridge between text and mental imagery.

Third, teachers need appropriate training to facilitate the visual comparison and revision discussions effectively. Simply showing students AI-generated images is not sufficient; teachers need to guide students to identify specific gaps between their intended meaning and the generated visuals, and to connect those gaps to specific revision strategies. The model works best when embedded within a systematic writing curriculum rather than used as an isolated activity.

Fourth, the findings reported above are exploratory and derive from a single qualitative case study without a control group. They represent promising indications of the model's potential under specific classroom conditions, but they cannot support causal claims; further research with rigorous designs is needed to confirm these outcomes and quantify their magnitude.

https://doi.org/10.67728/sshi.2026.018

Fifth, ethical and practical considerations related to student data deserve consideration. When students' writing samples are input into AI image generation tools, their written work becomes input to third-party AI systems. Teachers should be transparent with students and parents about how their writing samples will be used, and should ensure compliance with relevant data protection regulations. Where possible, teachers can use their own anonymized samples rather than individual student work when demonstrating the text-to-image feedback process, thereby protecting student privacy while still providing instructive examples.

Sixth, AI image generation can reproduce or even amplify social and cultural stereotypes, and the images produced from students' descriptions may reflect biased associations in the underlying model. Teachers should discuss this risk explicitly with students, examine generated images critically for stereotypical representation, and vary prompts and tools to reduce the influence of any single model's biases. This ethical dimension is part of the broader responsibility of integrating generative AI into language classrooms in a critical, equitable way.

8. Discussion

8.1. Reconceptualizing Writing Instruction: From Sentence-Driven to Context-Driven

The "image-text-image" bidirectional training model represents a reconceptualization of narrative writing instruction, shifting the focus from sentence-level accuracy to context-driven composition. Rather than treating writing primarily as the production of correct sentences, the model begins with the construction of a coherent mental model of the story world and lets linguistic choices follow from it (Section 3.3).

This perspective aligns with Yi's (2020) argument that writing instruction should target pre-linguistic mental representation capacity, not just linguistic forms. By making mental imagery explicit and trainable, the bidirectional model addresses a root cause of formulaic writing: when students lack a clear mental model of what they want to write about, they fall back on memorized templates and sentence patterns. The visual scaffolding and feedback provided by the model help students build that mental model before they begin to write.

The model also speaks to the distinction between "writing to learn" and "learning to write." While traditional instruction often focuses on the latter—teaching students how to write correctly—the bidirectional model emphasizes that improving writing also involves developing the thinking and imagining processes that precede writing. This is consistent with broader shifts in writing pedagogy toward process-oriented and meaning-centered approaches.

8.2. AI as a Cognitive Mirror, not a Gradebook

A key contribution of this model is its framing of AI text-to-image technology as a cognitive mirror rather than an automated grading tool. The model does not use AI to evaluate student writing or assign scores; instead, it uses AI to externalize the semantic content of student writing, making it visible for reflection and revision. This approach addresses a common concern about generative AI in education: that it may encourage over-reliance on technology or replace human judgment.

By positioning AI as a diagnostic and reflective aid rather than an evaluative tool, the model keeps students at the center of the learning process. Students are the ones who compare the generated image to their intended meaning, identify gaps, and decide how to revise. The AI serves as a window into how well their language conveys their ideas, not as an authority that judges whether their writing is good or bad.

This framing is consistent with the observation of Kohnke et al. (2023) that generative AI in writing classrooms works best when it is used as a supportive tool that enhances rather than replaces teacher-student interaction. In the proposed model, AI-generated visuals serve as discussion catalysts that enrich classroom dialogue about writing quality, rather than as automated feedback that replaces teacher guidance.

Importantly, this model also implies a shift in the teacher's role. Rather than being the primary source of feedback on writing quality, the teacher becomes a facilitator of visual comparison and revision dialogue. The teacher's expertise is directed toward helping students interpret the mismatches between their intended meaning and the AI-generated image, and toward guiding them in selecting appropriate revision strategies. This shift can be empowering for teachers who have long struggled to provide timely, individualized feedback to large classes: the AI handles the initial generation of visual feedback, freeing the teacher to focus on higher-order instructional tasks such as guiding discussion, facilitating peer review, and designing increasingly challenging tasks that push students toward greater descriptive sophistication.

8.3. Alignment with Multimodal Pedagogical Principles

The bidirectional model also aligns with broader multimodal pedagogical principles that emphasize the interconnectedness of visual and verbal meaning-making. Lim and Kessler (2024) argued that multimodal composing experiences deepen learners' understanding of how meaning is constructed across different modes. The "image-text-image" model operationalizes this principle

within a narrative writing context, showing how alternating between visual and verbal modes can strengthen students' understanding of how language choices convey visual and sensory meaning.

The model also extends Kress and van Leeuwen's (2021) grammar of visual design into the writing classroom. While Kress and van Leeuwen focused on analyzing how images make meaning, the proposed model uses AI-generated images as a practical tool for helping students understand how their written language contributes to meaning-making. By seeing how their words are translated into images, students gain insight into the relationship between linguistic choices and visual meaning.

This multimodal perspective also helps explain why the model appears to transfer beyond writing to improved reading comprehension. When students learn to attend to visual details in images, and to connect those details to language, they also become more attentive to how language creates visual imagery in reading. The same cross-modal connections that support writing production also support reading comprehension, which is particularly valuable for integrated tasks like story continuation.

8.4. Practical Implications for EFL Writing Pedagogy

The proposed model carries several practical implications for EFL writing pedagogy more broadly. First, it suggests that writing instruction should begin earlier in the cognitive process than traditionally assumed. Rather than starting with topic sentences, paragraph structure, or grammatical patterns, effective writing instruction should begin with helping students construct clear, vivid mental representations of what they want to write about. The bidirectional model provides a concrete, operationalized way to make this pre-writing visualization stage explicit and trainable.

Second, the model demonstrates how technology can serve as a bridge between cognitive theory and classroom practice. Dual-coding theory and scaffolding theory have been influential in educational psychology for decades, but their practical application in writing pedagogy has often remained abstract. AI text-to-image technology provides a tangible tool that makes these theoretical principles immediately actionable in the classroom. Teachers do not need to be cognitive scientists to implement the model; they only need to follow the four-step workflow and guide students through the visual comparison and revision process.

Third, the model emphasizes the importance of iterative revision cycles. Traditional writing instruction often treats revision as a final step focused on correcting grammar and spelling. The bidirectional model reframes revision as a meaning-making process: students revise not just to fix errors, but to improve the correspondence between their intended meaning and their expressed language. This shift in perspective can help students develop a more sophisticated understanding of writing as a recursive, meaning-driven process rather than a linear, accuracy-focused task.

The model also has implications for differentiated instruction. Students at different proficiency levels may benefit from different amounts and types of visual scaffolding. Beginner-level learners may require more extensive image-to-text practice with highly detailed prompts and more explicit vocabulary support, while more advanced learners can move more quickly to the text-to-image feedback phase and can handle more complex and ambiguous scene descriptions. Teachers can adjust the pacing and intensity of the bidirectional training to meet the needs of students at different proficiency levels, ensuring that the model is accessible and effective across a range of classroom contexts.

9. Conclusions

9.1. Summary of the Study

This paper has proposed an "image-text-image" bidirectional training model that leverages AI text-to-image technology to cultivate students' mental visual imagery in English narrative writing. Grounded in dual-coding theory, multimedia learning principles, and scaffolding theory, the model addresses a persistent challenge in EFL writing instruction: students' tendency to produce formulaic, contextually inappropriate narratives, to which underdeveloped mental imagery may contribute alongside limited vocabulary knowledge and instructional habits.

The model operationalizes a four-step instructional workflow: curating thematic narrative scene scaffolds, conducting layered image-to-text description practice, delivering visual feedback through text-to-image generation, and gradually fading external visual supports to foster internal mental imagery. The qualitative evidence gathered in the case study suggests that the approach may enhance students' awareness of narrative details, improve the contextual appropriateness of their language use, and increase their sensitivity toward narrative textual clues in source texts; these indications are provisional and context-bound.

The core contribution of this study lies in proposing how generative AI can be meaningfully integrated into EFL writing pedagogy not as an automated grading tool, but as a cognitive mirror that makes invisible mental processes visible. By externalizing the semantic content of students' writing through AI-generated images, the model creates a feedback loop that supports targeted revision and helps students develop the mental imagery capabilities that underpin vivid, context-driven narrative writing.

https://doi.org/10.67728/sshi.2026.018

9.2. Limitations and Future Research Directions

This study has several limitations that should be acknowledged. First, the evidence base is a single qualitative case study conducted in one school, without a control group; the findings therefore describe patterns observed under specific classroom conditions rather than demonstrated effects. Future research should employ quasi-experimental or mixed-methods designs with larger and more diverse samples to examine the model's impact on writing quality, descriptive sophistication, and narrative coherence.

It is also important to distinguish two outcome dimensions for future evaluation: image-supported descriptive performance during training, and performance on a new story-continuation task completed without visual support; these address different constructs, and the present study provides evidence mainly for the former, in the form of observed patterns in drafts and classroom engagement.

Second, the model has been primarily implemented in senior high school English contexts in China. Its applicability to other proficiency levels, age groups, and educational contexts remains to be investigated. Future research should explore how the model can be adapted for beginning-level learners, advanced university students, and instructional settings in different cultural and linguistic contexts.

Third, the rapid evolution of generative AI technology means that the specific tools and capabilities described in this paper may change rapidly. Future research should track how advances in text-to-image generation (e.g., improved text understanding, better control over visual elements) create new pedagogical possibilities, while also addressing emerging ethical and practical considerations such as data privacy and equity of access.

Finally, future research should explore the long-term effects of the model on students' writing development. While the present findings are promising, it remains to be seen whether the gains observed in the short term persist after students leave the intervention, and whether the internalization of visual imagery skills transfers to other types of writing tasks beyond narrative continuation.

Despite these limitations, the proposed model offers a theoretically grounded and practically feasible approach to addressing a significant pedagogical challenge in EFL writing instruction. As generative AI technology becomes increasingly accessible, frameworks like the "image-text-image" bidirectional training model can help ensure that these tools are used in pedagogically sound ways that support meaningful language learning rather than serving as technological novelties. The model represents a step toward bridging the gap between cognitive theories of writing and practical classroom implementation, providing a concrete operationalization of how visual mental imagery can be cultivated through carefully structured, technology-supported instructional sequences.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki. Ethical review and approval were obtained through the school's local administrative procedures at Urumqi No. 36 Middle School; because the study was a classroom-based pedagogical inquiry using anonymized student writing, no formal ethics protocol number was assigned.

Informed Consent Statement: Written informed consent was obtained from the school administration and from the parents or guardians of all participating students. Separate written consent for the anonymized processing of student texts by AI image-generation services was also obtained. Written informed consent for publication of the anonymized excerpts was obtained from the students' parents or guardians.

Data Availability Statement: The data presented in this study are not publicly available because they consist of student writing samples and generated images that could compromise participant privacy. Anonymized excerpts are available from the corresponding author on reasonable request.

Acknowledgments: The author thanks the participating students and the school administration of Urumqi No. 36 Middle School for their cooperation.

Conflicts of Interest: The author declares no conflicts of interest.

References

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brosch, R. (2018). What we "see" when we read: Visualization and vividness in reading fictional narratives. *Cortex, 105*, 135–143. https://doi.org/10.1016/j.cortex.2017.08.020

Chen, D. (2026). Transforming uncertainty into exploration: Effects of text-to-image AI on EFL learners' writing anxiety, motivation, and performance. *Frontiers in Psychology, 17*, Article 1866688. https://doi.org/10.3389/fpsyg.2026.1866688

Darvin, R. (2025). The need for critical digital literacies in generative AI-mediated L2 writing. *Journal of Second Language Writing, 67*, Article 101186. https://doi.org/10.1016/j.jslw.2025.101186

Green, M. C., & Brock, T. C. (2000). The role of transportation in the persuasiveness of public narratives. *Journal of Personality and Social Psychology, 79*(5), 701–721. https://doi.org/10.1037/0022-3514.79.5.701

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal, 54*(2), 537–550. https://doi.org/10.1177/00336882231162868

Kress, G., & van Leeuwen, T. (2021). *Reading images: The grammar of visual design* (3rd ed.). Routledge. https://doi.org/10.4324/9781003099857

Lim, J., & Kessler, M. (2024). Multimodal composing and second-language acquisition. *Language Teaching, 57*(2), 183–202. https://doi.org/10.1017/S0261444823000125

Mayer, R. E. (2021). *Multimedia learning* (3rd ed.). Cambridge University Press. https://www.cambridge.org/gb/universitypress/subjects/psychology/educational-psychology/multimedia-learning-3rd-edition

Moody, M. J., Lopez, V. A., & Johnson, P. R. (2026). Early implementation of AI-generated art in TESOL: An exploration of AI-generated art for language acquisition. *Studies in Linguistics and Language Education, 2*(1), 209–228. https://asjp.cerist.dz/en/downArticle/1009/2/1/298794

Paivio, A. (1986). *Mental representations: A dual coding approach*. Oxford University Press. https://global.oup.com/academic/product/mental-representations-9780195066661

Shu, L. P., & Xiong, B. Z. (2018). Micro-skills of action description in story continuation. *Journal of Foreign Language Teaching in Schools (Middle-School Edition), 41*(7), 30–33. http://lib.cqvip.com/Qikan/Article/Detail?id=675547811

Unsworth, L., & Mills, K. A. (2020). English language teaching of attitude and emotion in digital multimodal composition. *Journal of Second Language Writing, 47*, Article 100712. https://doi.org/10.1016/j.jslw.2020.100712

van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher–student interaction: A decade of research. *Educational Psychology Review, 22*(3), 271–296. https://doi.org/10.1007/s10648-010-9127-6

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press. https://www.hup.harvard.edu/catalog.php?isbn=9780674576292

Ye, W., Ren, W., & Zheng, S. (2021). The construct of the story continuation writing task: Insights from the China's Standards of English Language Ability. *Chinese Journal of Applied Linguistics, 44*(3), 382–398. https://doi.org/10.1515/CJAL-2021-0024

Yi, Y. (2020). Using visualization to enhance intermediate EFL learners' narrative writing. *International Journal of TESOL Studies, 2*(1), 34–49. https://doi.org/10.46451/ijts.2020.06.03

Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). SAGE. https://study.sagepub.com/yin6e

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of Apollo and/or the editor(s). Apollo and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.