AUREUM ACADEMIC JOURNAL
VOL. 3, NO. 4, October 2026
https://doi.org/10.48161/aaj.v3n4a100

Using Visual Storytelling to Improve Learning Outcomes in Primary and Secondary Education

Roseline Sama[1*]

[1] Science Education, Faculty of Education, Federal University, Ndufu-Alike, Ebonyi State, Abakaliki, Nigeria;

* **Corresponding author:** roselineanochiwa@gmail.com.

ABSTRACT: Visual storytelling combines visual representation, narrative organization, academic language, and revision to support curriculum learning. This study examined the effects of a curriculum-aligned visual-storytelling program on comprehension and academic achievement (co-primary outcomes) and on visual literacy and communication (prioritized secondary outcomes). We conducted a cluster-randomized mixed-methods trial in six public schools from one mixed urban-peri-urban district, involving 24 intact primary and secondary classes and 384 students. Twelve classes received visual-storytelling instruction, and twelve received active conventional instruction, for 12 weeks. Baseline-adjusted multilevel models accounted for class clustering and the finite number of clusters. At post-test, the adjusted advantage was 5.4 points for comprehension (95% CI [2.5, 8.3]; g = 0.44) and 4.9 points for achievement (95% CI [2.0, 7.8]; g = 0.40). At the eight-week follow-up, the corresponding differences were 4.1 points (95% CI [1.1, 7.1]; g = 0.32) and 3.6 points (95% CI [0.7, 6.5]; g = 0.29). We assessed visual literacy, communication, imagination, engagement, and motivation at baseline, post-test, and follow-up; visual literacy and communication showed the largest secondary effects. Multiple imputation, complete-case, and cluster-robust sensitivity analyses yielded comparable conclusions. School-stage interactions were exploratory and imprecise. Artifact and qualitative evidence provided interpretive support for program implementation, particularly sequence coherence, image-concept alignment, audience interpretation, and revision. The main limitation was the sample's concentration in six schools, which limited subgroup precision and broader generalization. The findings indicate that visual storytelling can strengthen curriculum learning when multimodal production is explicitly connected to disciplinary content, assessment criteria, and individual academic explanation.

Keywords: visual storytelling; language and literacy education; multimodal composition; visual literacy; cluster-randomized trial.

I. INTRODUCTION

Language and literacy education requires learners to interpret and produce meaning across written language, speech, images, diagrams, symbols, and spatial organization. Visual storytelling addresses this demand by asking learners to transform curriculum content into multimodal compositions that coordinate visual evidence with disciplinary vocabulary, oral explanation, and written response. Digital storytelling is one technology-mediated form of this broader practice, while paper-based storyboards, illustrated explanations, and sequential visual arguments share the same requirement: connecting modes for a communicative purpose [1].

In language-related classrooms, this coordination can support reading comprehension by making relationships, causes, and perspectives visible; disciplinary writing by requiring students to translate a visual artifact into an individually assessable explanation; and oral academic language by creating occasions for rehearsal, questioning, and clarification. Classroom studies have reported gains in language performance,

visual memory, writing, and learning motivation when story production includes explicit curriculum alignment and revision [2]–[4].

Collaborative visual-narrative work also alters classroom interaction. Students compare evidence, negotiate the adequacy of a representation, test whether an audience can reconstruct the intended meaning, and revise ambiguous elements. Research on co-creativity and transmedia learning links these practices to distributed responsibility and audience awareness [5], [6]. Related studies connected digital storytelling with active listening and creative thinking [7], [8]. Speaking and writing research identified gains in selected communication outcomes, although effects were not uniform across all dimensions [9], [10].

Visual literacy provides a related but distinct foundation. Familiarity with images does not guarantee the capacity to evaluate visual claims, identify evidential relevance, or design representations for a disciplinary purpose. A controlled visual-literacy program was associated with improved reading comprehension [11]. Multimodal analyses demonstrated that image, language, identity, values, and audience expectations interact in student-created texts [12], [13]. Conceptual work further defined visual literacy as interpretation, evaluation, and purposeful production rather than image exposure alone [14]. These findings support assessment criteria that distinguish conceptual accuracy and communicative clarity from artistic polish.

Quasi-experimental language research has also reported positive speaking outcomes, while showing that the magnitude and distribution of benefits can vary across learners and assessed dimensions [15]. The wider evidence therefore supports a cautious interpretation: visual storytelling is most defensible when the intervention, outcome hierarchy, classroom context, and analytical unit are reported precisely.

The researchers selected active conventional instruction as the comparison because the research question concerned the added value of visual-storytelling routines beyond ordinary curriculum teaching, not the contrast between instruction and no instruction. Both conditions covered the same units, objectives, contact time, and assessment schedule. This design reduced the likelihood that observed differences reflected simple exposure to teaching or additional curriculum time.

The study contributes a within-school class-randomized estimate of comprehension and academic achievement, an eight-week assessment of persistence, and an integrated account of outcome, artifact, fidelity, and classroom-process evidence. These contributions strengthen causal estimation while retaining a bounded interpretation of the representational processes observed during implementation.

1. RESEARCH PROBLEM

The research problem comprised three related limitations in the available evidence. First, class-randomized evaluations of curriculum-based visual storytelling remained uncommon, particularly across more than one school stage. Second, delayed assessment was infrequent, which limited conclusions about whether immediate learning differences persisted after instruction. Third, the study rarely integrated artifact quality with academic outcomes in a way that linked representational work to measured learning without treating qualitative processes as proven mediators. The study addressed these limitations through within-school class randomization, repeated academic assessment, and an embedded analysis of artifacts, implementation, and classroom interaction.

2. RESEARCH AIM

This study estimated the effects of curriculum-aligned visual-storytelling instruction on academic and multimodal outcomes and examined classroom and artifact evidence relevant to its implementation. The analysis focused on four objectives: estimating effects on comprehension and academic achievement; assessing persistence at eight weeks; examining school-stage differences as exploratory interactions; and interpreting artifact and classroom evidence in relation to the observed outcome pattern.

3. RESEARCH QUESTIONS

- RQ1. To what extent did visual-storytelling instruction improve comprehension and academic achievement relative to conventional instruction?
- RQ2. Were the estimated comprehension and achievement effects sustained eight weeks after the program ended?
- RQ3. What exploratory differences were observed between primary and secondary school-stage estimates?

- RQ4. How did artifact characteristics and classroom interactions inform interpretation of the observed outcome patterns?

II. LITERATURE REVIEW

1. VISUAL STORYTELLING, DIGITAL STORYTELLING, AND MULTIMODAL COMPOSITION

Visual storytelling organizes meaning through visual sequence and narrative relations. Digital storytelling is one implementation that combines images, narration, written text, audio, or video through digital production, whereas multimodal composition refers more broadly to the purposeful orchestration of semiotic modes. The traditions overlap but emphasize different features: technology-mediated authorship, sequential visual representation, and the design relationship among modes [1], [12], [13]. Across all three, planning, audience testing, revision, and explanation matter analytically because they reveal the learner's decisions rather than treating the final artifact as the sole outcome.

Studies differ in intervention duration, school stage, product requirements, and outcome selection. Some assess writing or speaking, others emphasize motivation or artifact quality, and relatively few connect collaborative production to an individually scored academic response. This variation limits direct comparison and makes precise definitions of treatment essential. The evidence supports visual storytelling as a family of instructional practices rather than a single standardized intervention.

2. LANGUAGE, LITERACY, AND ACADEMIC OUTCOMES

Controlled and quasi-experimental language studies have linked story creation with writing, speaking, listening, and engagement [2], [7], [9]. Additional EFL studies reported improvements in speaking and writing performance [15]–[17]. Qualitative and multimedia-supported investigations further connected digital storytelling with collaborative literacy practices and oral-language development [18], [19]. A systematic review found that positive conclusions were common, but reporting on reliability and methodological support for some claims remained inconsistent [20].

The evidence was not uniformly conclusive. Effects varied by teacher support, task complexity, assessment alignment, learner achievement, and the outcome analyzed [15]–[17]. The analytical synthesis is therefore conditional: visual and digital storytelling supported learning most consistently when tasks required sustained content processing, explicit criteria, revision, and accountable explanation [18]–[20].

3. VISUAL LITERACY AND MULTIMODAL ASSESSMENT

Visual literacy involves interpreting, evaluating, and creating visual communication in context. Curriculum-based studies link explicit visual analysis to comprehension and critical thinking, while teacher-education research emphasizes guided questioning, contextual interpretation, and source evaluation [21], [22]. Studies of academic image use further indicate that visually experienced learners still require instruction in evidential relevance, attribution, and ethical communication [23].

Assessment is therefore a central methodological issue. Rubrics that reward artistic polish can confound drawing confidence with conceptual representation. More defensible dimensions include symbol relevance, sequence coherence, image-caption consistency, evidential linkage, audience clarity, and the quality of subsequent oral or written explanation. This literature supports measuring visual literacy and academic achievement separately and using artifact evidence as an interpretive record of disciplinary representation.

4. CONTROLLED EVIDENCE AND METHODOLOGICAL LIMITATIONS

Controlled studies provide stronger estimates than single-group classroom reports, yet recurring limitations include small numbers of intact classes, omission of intraclass correlations, limited delayed assessment, and weak integration of artifact and outcome evidence. Several language studies used quasi-experimental or classroom-based designs without within-school class randomization [15]–[17]. Related research varied substantially in sample, duration, outcome selection, and statistical treatment [18]–[20].

The literature therefore reveals a tension between pedagogical promise and uneven methodological precision. More informative evaluations require class-aware estimation, prioritized outcomes, delayed

measurement, transparent scoring, implementation monitoring, and integration that identifies convergence and divergence rather than treating qualitative evidence as proof of mediation.

III. RESEARCH GAP

Systematic and controlled evidence identified four persistent, although not absolute, gaps [1], [20]. Class-level evaluations across multiple school stages remained limited; delayed academic measurement was uncommon; artifact quality was seldom connected to academic outcomes through explicit mixed-methods integration; and developmental comparisons were typically underpowered. The present design strengthened evidence in each area through within-school class randomization, repeated assessment, exploratory school-stage interactions, and an embedded process evaluation. It did not fully resolve questions of long-term maintenance, developmental moderation, or statistical mediation.

IV. CONCEPTUAL AND THEORETICAL FRAMEWORK

1. COMPLEMENTARY THEORETICAL LENSES

The framework centers on three complementary components. Multimedia learning explains how to select, organize, and integrate verbal and visual information [24]–[26]. Narrative and signaling research explains how sequence and cues make temporal and causal relations explicit [27], [28]. Spatial-integration research clarifies how proximity between representations can reduce unnecessary search demands [29], [30]. Dual coding provides a foundational account of coordinated representation and later retrieval [31], while narrative theory explains how learners organize experience through temporal and causal structure [32]. Research on visual displays emphasizes instructional support for representational comparison [33]. Cooperative-learning research further emphasizes peer questioning and individual accountability [34].

The components corresponded to distinct analytical roles. Multimedia and dual-coding principles were linked to comprehension, achievement, and delayed persistence; narrative and representational organization were linked to sequence coherence, image-concept alignment, and visual literacy; collaborative interpretation and revision were linked to communication, audience awareness, and implementation fidelity. Together, they formed a representational-cycle interpretive model in which students externalized content, interpreted it, revised it, and translated it into academic language. The model organized interpretation but did not establish a causal mediation chain.

2. VARIABLES AND ANALYTICAL ROLES

The independent variable was assigned classroom condition: visual-storytelling instruction or active conventional instruction. Comprehension and academic achievement were the co-primary outcomes. Eight-week scores represented delayed persistence. We prioritized visual literacy and communication as secondary outcomes, while imagination, engagement, and motivation were exploratory. We examined sequence coherence, image-concept alignment, audience interpretation, and revision as interpretive processes. School stage was prespecified as an exploratory moderator.

3. ANALYTICAL EXPECTATIONS

- E1. The prespecified expectation concerned higher baseline-adjusted comprehension and achievement in visual-storytelling classes at post-test.
- E2. Positive comprehension and achievement differences were expected to remain at the eight-week follow-up, but with smaller magnitudes than at post-test.
- E3. Visual literacy and communication were expected to show positive post-test differences because both competencies were directly practiced.
- E4. We designated school-stage contrasts as exploratory, and artifact and classroom evidence as interpretive rather than mediational.

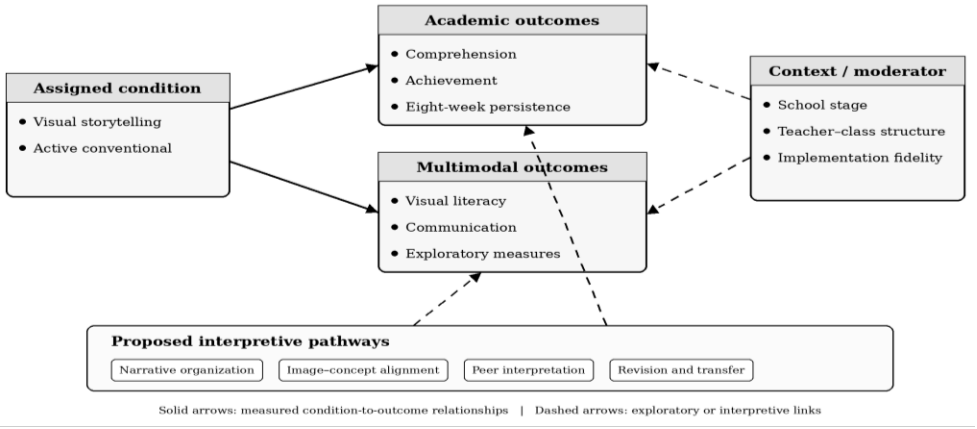


FIGURE 1. Analytical framework of measured outcomes and proposed interpretive pathways.

Note. Solid connectors represent measured condition-to-outcome relationships. Dashed connectors represent proposed interpretive pathways.

V. STUDY CONTEXT AND INTERVENTION

1. CLASSROOM AND CURRICULUM CONTEXT

The study was conducted in six public schools within one mixed urban-peri-urban education district: three primary schools and three secondary schools. Each school contributed four intact classes, for a total of 24 participating classes. Primary classes were drawn from Grades 5–6 and secondary classes from Grades 8–9. Within each school, we assigned two classes to visual-storytelling instruction and two to active conventional instruction. Student instructional contact time, curriculum objectives, subject content, and assessment schedules were identical across conditions; the contrast concerned the representational, peer-interpretation, and revision routines used to support learning.

Table 1. Classroom and curriculum characteristics by instructional condition.

Characteristic	Visual-storytelling condition	Active conventional condition
Classes and baseline sample	12 classes; 194 students; mean class size 16.2	12 classes; 190 students; mean class size 15.8
Schools and allocation	Six schools; two classes per school	Six schools; two classes per school
School stages and grades	6 primary classes (Grades 5–6); 6 secondary classes (Grades 8–9)	6 primary classes (Grades 5–6); 6 secondary classes (Grades 8–9)
Subject distribution	Language, humanities, and social-studies classes	Matched language, humanities, and social-studies classes
Programme dose	12 weeks; two 50-minute lessons weekly; 20 contact hours	12 weeks; two 50-minute lessons weekly; 20 contact hours
Teachers	12 certified teachers; one established class each	12 certified teachers; one established class each
Teacher experience	Condition staff drawn from 24-teacher sample; overall mean 10.4 years (SD 5.8)	Condition staff drawn from 24-teacher sample; overall mean 10.4 years (SD 5.8)
Teacher preparation	Two 3-hour workshops, manual, and brief logistics/fidelity meetings	90-minute assessment/scheduling briefing and ordinary departmental support

Characteristic	Visual-storytelling condition	Active conventional condition
Core learning activity	Visual interpretation, storyboarding, peer critique, revision, presentation, and individual transfer writing	Teacher explanation, textbook work, guided questioning, oral response, written exercises, and standard revision
Unit of assignment and analysis	Intact class; student outcomes modeled with class clustering	Intact class; student outcomes modeled with class clustering

Note. Six public schools contributed four classes each. We established classes and teachers before randomization; one teacher taught each class, and no teacher delivered both conditions. Teacher and class effects were therefore not statistically separable.

2. *VISUAL-STORYTELLING PROGRAMME*

The program consisted of five linked phases. Foundation activities established visual-noticing and image-questioning routines. Construction activities required students to organize sequence, select evidence, and draft captions or visual metaphors. Development activities emphasized collaborative artifact production and oral rehearsal. Revision activities used peer interpretation and teacher conferences to identify ambiguity and missing conceptual links. Transfer activities required a presentation and an individually assessed written explanation. Rotating roles, including image selector, sequence organizer, caption writer, peer questioner, visual editor, and explanation leader, reduced dependence on drawing confidence and preserved individual accountability.

Table 2. Visual-narrative components, purposes, observable activity, and fidelity evidence.

Component and dose	Instructional purpose	Student activity and responsible person	Fidelity criterion and record
Visual noticing; Weeks 1–2	Direct attention to relevant visual information	Students identified salient features and missing information; classroom teacher facilitated	Routine documented in lesson log and noticing sheet completed
Storyboard construction; Weeks 3–5	Externalize sequence, causality, and relationships	Students ordered panels, selected evidence, and drafted captions; classroom teacher facilitated	Storyboard and peer-question record present for the scheduled phase
Concept visualization; Weeks 6–8	Translate abstract content into visible form	Students used symbols and image-concept links; classroom teacher facilitated	Artifact met curriculum-alignment criterion and rubric record was completed
Peer interpretation; Weeks 8–9	Test audience understanding and expose ambiguity	Students explained panels and responded to questions; peers and teacher facilitated	Peer-interpretation sheet documented meaning-focused feedback
Revision and resequencing; Weeks 9–10	Improve conceptual and explanatory coherence	Students revised symbols, captions, order, or missing steps; classroom teacher conferred	Revision sheet identified at least one content-related change
Presentation and transfer; Weeks 11–12	Translate the artifact into oral and written disciplinary language	Students presented and completed an individual explanation; classroom teacher assessed	Presentation record and individual transfer response completed

3. INSTRUCTIONAL PROGRESSION

The twelve-week progression separated interpretation, construction, development, revision, and transfer. We reduced labels and activities to five instructional phases to support replication without treating every classroom action as a distinct component. The final phase required students to explain how visual choices represented subject knowledge and to complete an individually assessed response.

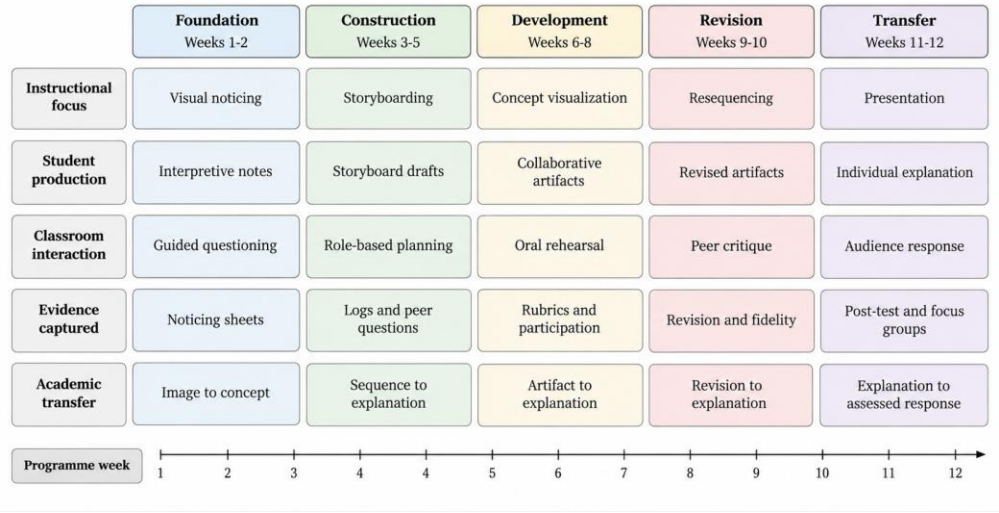


FIGURE 2. Week-by-week instructional matrix for the twelve-week visual-storytelling program.

Note. The matrix aligns instructional emphasis, student production, classroom interaction, evidence capture, and academic transfer across the programme.

4. CONVENTIONAL INSTRUCTION

Comparison classes covered the same units, learning objectives, student contact time, and assessment schedule through teacher explanation, textbook materials, guided questioning, written exercises, brief oral responses, and standard revision tasks. This active comparison represented ordinary district instruction and controlled for curriculum exposure, assessment timing, and instructional contact. It did not control for the additional teacher preparation required to implement the visual-storytelling programme.

5. IMPLEMENTATION MONITORING AND CONTAMINATION CONTROL

Twenty-four certified teachers participated, one per class; 17 were women, and 7 were men, with a mean of 10.4 years of teaching experience (SD = 5.8). Intervention teachers completed two three-hour workshops and received a structured manual. Brief weekly meetings addressed scheduling, record completion, and adherence to the manual; they did not add student contact time or provide outcome-specific coaching. Comparison teachers received a 90-minute assessment and scheduling briefing and continued ordinary departmental support. We monitored fidelity through teacher logs, attendance records, revision sheets, structured checklists, and 26 classroom observations, equally divided between conditions. Mean intervention fidelity was 89.2% (SD = 5.7; range = 78%–97%); teachers implemented noticing and storyboard routines most consistently, while protected revision time varied most. Two comparison observations included isolated visual-organizer use, but no comparison class used storyboards, peer visual interpretation, or the complete representational cycle. The difference in teacher preparation remained an implementation-related alternative explanation.

VI. METHODOLOGY

1. DESIGN AND PARTICIPANTS

The study used a parallel cluster-randomized classroom trial with an embedded explanatory mixed-methods process evaluation. The classroom was the unit of assignment and the student was the unit of outcome measurement. Twenty-four intact classes and 384 students completed baseline assessment, with equal allocation to visual-storytelling and active conventional instruction. Classes were eligible when they were scheduled to complete the selected language, humanities, or social-studies units, had an established class teacher, and could complete baseline, post-test, and follow-up assessment. Students entered the research sample after parental permission and student assent; students who transferred before baseline or were absent throughout the baseline window did not enter the trial. The protocol designated comprehension and academic achievement as co-primary outcomes, visual literacy and communication as prioritized secondary outcomes, and school stage as exploratory.

2. PROSPECTIVE SAMPLE-SIZE DETERMINATION

The prospective calculation used the cluster-trial design effect:

$$DE = 1 + [m(1 + CV^2) - 1] \times ICC$$

where m was the anticipated mean class size and CV represented cluster-size variation. The calculation assumed g = 0.425, ICC = .04, m = 16, CV = .10, equal allocation, 80% power, and a two-sided alpha of .025 for the two co-primary outcomes. The anticipated effect and ICC were informed by the instrument pilot and prior classroom studies [2], [4], [11]. Allowing for approximately 10% student attrition yielded a target of 24 classes and about 380 students; 24 classes and 384 consented students completed the baseline assessment.

3. ELIGIBILITY, RECRUITMENT, AND PARTICIPANT PROGRESSION

School leaders identified 24 intact classes meeting the curriculum and scheduling criteria. Of 401 eligible students, 384 enrolled after parental permission and student assent, yielding a 95.8% recruitment rate. Seventeen students did not enter the trial: nine lacked parental permission, five transferred before baseline assessment, and three were absent throughout the baseline window. The enrolled sample comprised 197 girls and 187 boys, with a mean age of 12.9 years (SD = 2.1; range = 10–16). Primary and secondary school stages each contributed 192 students. At post-test, we assessed 365 students, and 345 completed the eight-week follow-up.

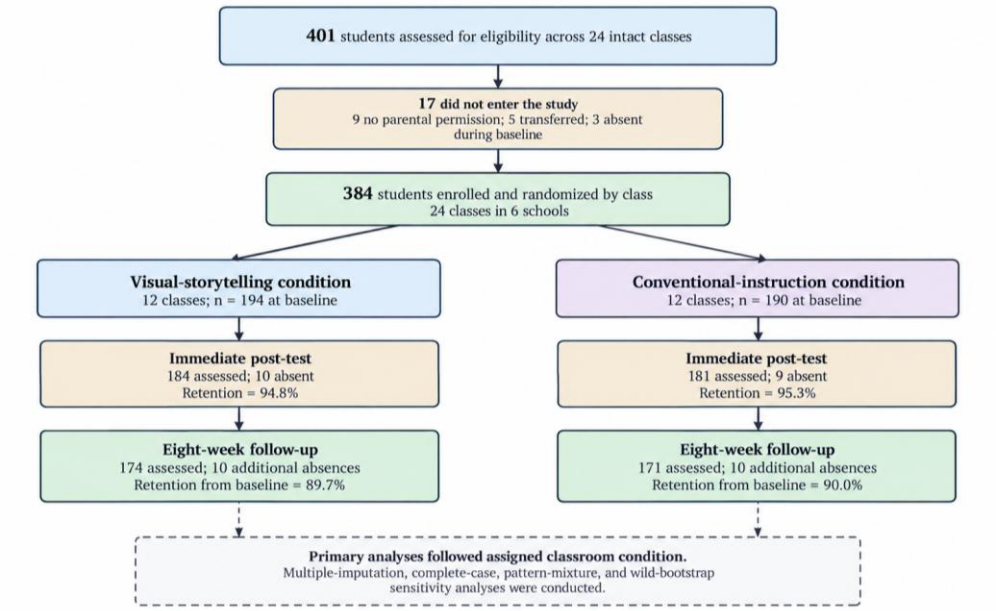


FIGURE 3. Cluster-trial participant progression from enrollment to eight-week follow-up.

Note. The diagram reports eligibility, exclusions, randomized classes, condition-specific assessment counts, and the analysis populations.

4. ALLOCATION AND ASSESSMENT MASKING

Randomization occurred after consent and baseline assessment. We established classes and teachers before allocation. Within each school, an independent statistician used a reproducible permuted block of four to assign two classes to each condition. The sequence was stored in a password-protected file and released to the project coordinator after baseline data were locked. Teachers and students could not be blinded to instruction. Outcome booklets and artifact records were labeled with numeric codes, and three outcome raters remained unaware of condition until scoring and data cleaning were complete. In a post-scoring masking check, raters identified the condition correctly for 51% of sampled records, which did not differ from chance ($p = .78$). Qualitative analysts were aware of the condition because the instructional content was evident in the data. Because one teacher taught one participating class, teacher and class effects remained statistically inseparable.

5. OUTCOME MEASUREMENT

The measurement strategy separated co-primary academic outcomes from prioritized secondary and exploratory measures. We developed grade-band-specific comprehension and achievement forms for Grades 5–6 and Grades 8–9 from common construct blueprints, shared content specifications, and anchor-item formats. We converted raw scores to 0–100 scales, allowing interpretation on a common metric while preserving grade-appropriate content. Parallel forms assessed the same academic domains at eight weeks. Visual literacy and communication used four-dimension analytic rubrics; raw totals from the 0–4 dimension scores were divided by the maximum score and multiplied by 100. We assessed imagination, engagement, and motivation at baseline, post-test, and follow-up. We piloted all instruments with 62 students from non-participating classes, and stage-specific reliability patterns were consistent with the pooled coefficients.

Table 3. Outcome definitions, scoring procedures, and measurement quality.

Domain and role	Instrument, scoring, and direction	Assessment timing and assessor	Validation, missingness, and interpretation
Comprehension—co-primary	28-item grade-band test; 0–100; higher = better; 3-point difference treated as curriculum-relevant	Baseline, post-test, 8-week follow-up; individual administration	KR-20 .83; S-CVI .92; difficulty .34–.77; discrimination .29–.58; missing 0%, 4.9%, 10.2%
Achievement—co-primary	30-item grade-band unit test; 0–100; higher = better; 3-point difference treated as curriculum-relevant	Baseline, post-test, 8-week follow-up; individual administration	KR-20 .81; S-CVI .89; difficulty .31–.79; discrimination .27–.55; missing 0%, 4.9%, 10.2%
Visual literacy—prioritized	Four dimensions × five levels; total rescaled to 0–100; higher = better	Baseline, post-test, follow-up; three masked raters	ICC(2,k) .85, 95% CI [.80, .89]; 25% double-scored
Communication—prioritized	Four dimensions × five levels; total rescaled to 0–100; higher = better	Baseline, post-test, follow-up; three masked raters	ICC(2,k) .82, 95% CI [.76, .87]; 25% double-scored
Imagination—exploratory	12-item curriculum-bounded scale; 0–100; higher = better	Baseline, post-test, follow-up; student self-report	Omega .86; CFI .95; RMSEA .052; metric invariance

Domain and role	Instrument, scoring, and direction	Assessment timing and assessor	Validation, missingness, and interpretation
Engagement—exploratory	10-indicator observation checklist; 0–100; higher = greater observed engagement	Baseline, post-test, follow-up sampling; trained observers	Kappa .79; expert review and observer calibration
Motivation—exploratory	8-item five-point scale; higher = greater motivation	Baseline, post-test, follow-up; student self-report	Omega .88; metric invariance
Participation—implementation	Role and participation records; descriptive only	Throughout programme; teacher and observer records	Cross-checked against observations; not a treatment outcome

Content validity was evaluated by six subject and measurement specialists. Item-level content-validity indices ranged from .83 to 1.00, and scale-level indices ranged from .89 to .93. Test-item difficulty ranged from .31 to .79 and discrimination from .27 to .58. Three trained raters completed six hours of calibration using benchmark responses; 25% of visual-literacy and communication records were independently double-scored. Inter-rater reliability was ICC(2,k) = .85 (95% CI [.80, .89]) for visual literacy and ICC(2,k) = .82 (95% CI [.76, .87]) for communication. Confirmatory analyses supported the intended factor structures of the imagination and motivation scales, and metric invariance across school stages met the change-in-CFI criterion of less than .01. Appendix A reports scoring anchors, representative items, administration instructions, and score-transformation rules.

6. ARTIFACT AND QUALITATIVE EVIDENCE

The process evaluation included 84 group storyboard drafts, 92 final visual artifacts, 49 individual reflection journals, 121 revision sheets, 116 peer-interpretation sheets, 12 intervention-teacher logs, 26 classroom observations, eight focus groups, and 30 student interviews. Each focus group included six students and was balanced across condition and school stage. We selected interview participants through maximum-variation sampling across condition, school stage, sex, attendance, and outcome profile; the sample included 15 primary and 15 secondary students, 16 from the intervention condition and 14 from the comparison condition. No student participated in both an interview and a focus group. The 26 observations represented 13 lessons per condition; 19 contained detailed participation-field notes used in the confidence-inclusive theme. Artifact counts represented group drafts, final group products, or individual records as specified in Appendix D.

7. ETHICS, CONSENT, AND PRIVACY PROTECTION

The Institutional Committee for Arts and Education Research approved the study on 18 November 2021. Written authorization was obtained from the education authority and participating principals. Parents or guardians provided written consent and students provided age-appropriate assent before baseline assessment and randomization. Participation in research data collection was voluntary; students who declined or withdrew remained in their ordinary assigned classroom instruction without academic penalty. Names were replaced with coded identifiers. An institutional data manager, not involved in outcome scoring, maintained the re-identification key separately from the assessment, artifact, interview, and observation files. The team stored paper records in locked cabinets and digital records on encrypted, access-controlled institutional servers. Quotations and student-created materials were de-identified before analysis and reporting. We retained consent records and research data separately for five years, in accordance with the approved protocol, consent documents, and institutional retention policy, after which secure destruction was authorized.

VII. STATISTICAL AND MIXED-METHODS ANALYSIS

1. PRIMARY ESTIMAND AND MULTILEVEL MODEL

The primary estimand was the adjusted mean difference between assigned classroom conditions at immediate post-test for comprehension and academic achievement. Analyses were conducted in R 4.3.2. Each outcome was modeled with fixed effects for condition, baseline score, school stage, and school, together with a class random intercept. School fixed effects were used because randomization occurred within six schools and the estimand concerned the average within-school treatment contrast; estimating school as a random effect with only six units produced an unstable variance component. A sensitivity model with school as a random intercept and class nested within school yielded the same substantive conclusions. CR2 cluster-robust covariance estimation and Satterthwaite degrees of freedom addressed finite-cluster inference. Results were reported as adjusted mean differences, 95% confidence intervals, exact p values, and Hedges g.

All primary models converged without singular-fit warnings. The class-level ICCs were .026 for post-test comprehension, .026 for post-test achievement, .030 for follow-up comprehension, and .027 for follow-up achievement. Residual plots showed no material departure from normality or homoscedasticity, and variance-inflation factors were below 1.4. CR2-corrected effective degrees of freedom ranged from 18.7 to 19.8. A class-level wild cluster bootstrap with 9,999 replications and the school-random-intercept sensitivity model produced conclusions consistent with the primary analysis. Appendix B reports coefficient estimates, variance components, diagnostics, and the reproducibility specification.

2. REPEATED MEASUREMENT, EXPLORATORY CONTRASTS, AND EFFECT SIZES

Immediate and eight-week outcomes were estimated separately in the primary analysis. A longitudinal sensitivity model included condition, occasion, baseline score, school stage, school, and a condition-by-occasion interaction with repeated observations nested within students and classes. The interaction indicated attenuation between post-test and follow-up but retained positive adjusted contrasts at both occasions. School-stage interactions were interpreted through adjusted marginal contrasts. Hedges g was calculated by dividing the adjusted condition difference by the pooled within-condition standard deviation and applying the small-sample correction based on the effective cluster degrees of freedom. Interaction estimates remained exploratory because only six classes per condition were available within each school stage.

3. MISSING DATA, MULTIPLICITY, AND SENSITIVITY

The primary analysis followed assigned classroom condition. Missing outcomes were addressed through multilevel multiple imputation by chained equations using 40 datasets and 30 iterations. Condition, school, class, school stage, baseline and available follow-up scores, attendance, and secondary outcomes were included in the imputation models. CR2-corrected coefficient estimates and robust within-imputation variances were obtained in each dataset and pooled using Rubin's rules with small-sample degrees-of-freedom adjustment. Trace plots stabilized before the final iterations, observed and imputed distributions overlapped closely, Monte Carlo error was small relative to model standard errors, and fractions of missing information were low to moderate. Complete-case and pattern-mixture analyses produced comparable estimates. The two co-primary outcomes used a two-sided alpha of .025. Visual literacy and communication were prespecified prioritized secondary outcomes and controlled with the Benjamini-Hochberg false-discovery rate at 5%; remaining outcomes were exploratory.

4. QUALITATIVE AND ARTIFACT ANALYSIS

Qualitative analysis followed reflexive thematic analysis through familiarization, initial coding, candidate-theme construction, comparison across conditions and school stages, and explicit review of negative cases [35]. The team comprised an education researcher responsible for classroom-process analysis and a qualitative methodologist who had no teaching or assessment role in the participating schools. Neither analyst held evaluative authority over students. Both researchers read the full corpus, compared interpretations, and maintained reflexive memos documenting assumptions, positionality, decisions, and alternative explanations. Coding was managed in NVivo 14. Artifact analysis proceeded in parallel using

sequence coherence, symbolism, image-concept alignment, caption precision, audience awareness, and revision. Theme frequency was not treated as prevalence; negative cases modified theme boundaries and identified conditions under which visual production did not support disciplinary meaning.

5. MIXED-METHODS INTEGRATION

Integration occurred at design, analysis, and interpretation stages. The process evaluation was embedded in the cluster trial, artifact and qualitative findings were compared with quantitative results during analysis, and a joint display supported meta-inference. Artifact rubric scores were linked to participant outcomes only when an individual identifier was available; group artifacts were used for class-level interpretation rather than participant-level association. Convergence required compatible direction and construct relevance across evidence streams; complementarity required non-duplicative explanatory detail; divergence required a documented contradiction or boundary condition. Qualitative evidence was not treated as proof of mediation. Joint-display methodology informed the integration procedure [36]. Reporting followed CONSORT, the cluster-trial extension, and TIDieR guidance [37]–[39].

VIII. RESULTS

1. BASELINE PROFILE, PARTICIPANT PROGRESSION, AND IMPLEMENTATION EXPOSURE

The two conditions were closely matched at baseline on age, sex, school stage, teacher experience, mean class size, and all measured outcomes. At immediate post-test, 184 students in the visual-storytelling condition and 181 in the comparison condition were assessed; follow-up included 174 and 171 students, respectively. Missingness was 4.9% at post-test and 10.2% at follow-up, with nearly identical condition-specific rates. All intervention classes completed the specified sequence. Mean implementation fidelity was 89.2% (SD = 5.7; range = 78%–97%), with the greatest between-class variation in protected revision time. Two isolated uses of a visual organizer occurred in comparison lessons, but no comparison class implemented the full visual-storytelling sequence.

Table 4. Baseline outcome profile by instructional condition.

Characteristic	Visual-storytelling baseline	Comparison baseline	Absolute standardized difference
Age, years	12.8 (2.0)	12.9 (2.1)	0.05
Girls	100 (51.5%)	97 (51.1%)	0.01
Primary-stage students	97 (50.0%)	95 (50.0%)	0.00
Mean class size	16.2	15.8	0.05
Teachers, n	12	12	0.00
Teacher experience, years	Overall mean 10.4 (5.8)	Overall mean 10.4 (5.8)	—
Comprehension (0–100)	61.4 (12.2)	61.9 (12.4)	0.04
Achievement (0–100)	60.8 (12.6)	61.2 (12.7)	0.03
Communication (0–100)	58.6 (13.1)	58.9 (13.0)	0.02
Visual literacy (0–100)	56.8 (14.0)	57.0 (13.8)	0.01
Imagination (0–100)	59.7 (12.8)	59.2 (12.9)	0.04
Motivation (1–5)	3.08 (0.61)	3.10 (0.63)	0.03
Engagement (0–100)	60.2 (12.0)	60.5 (12.2)	0.02

Note. Continuous variables are reported as mean (SD), categorical variables as n (%), and standardized mean differences as absolute values.

Table 5. Participant progression and attrition by school level and instructional condition.

School stage and condition	Classes	Baseline n	Post-test n	Follow-up n	Baseline-to-follow-up retention
Primary stage—visual storytelling	6	97	92	88	90.7%
Primary stage—conventional instruction	6	95	91	87	91.6%
Secondary stage—visual storytelling	6	97	92	86	88.7%
Secondary stage—conventional instruction	6	95	90	84	88.4%
Total	24	384	365	345	89.8%

2. *CO-PRIMARY ACADEMIC OUTCOMES*

Comprehension increased from 61.4 to 76.2 in the visual-storytelling condition and from 61.9 to 70.8 in the comparison condition. The CR2-corrected baseline-adjusted difference was 5.4 points (SE = 1.39; 95% CI [2.5, 8.3]; df = 19.8; $p < .001$; $g = 0.44$). Achievement increased from 60.8 to 74.7 and from 61.2 to 69.5, respectively; the adjusted difference was 4.9 points (SE = 1.38; 95% CI [2.0, 7.8]; df = 19.5; $p = .002$; $g = 0.40$). Both co-primary outcomes met the prespecified alpha of .025.

At the eight-week follow-up, the adjusted difference was 4.1 points for comprehension (SE = 1.43; 95% CI [1.1, 7.1]; df = 18.7; $p = .010$; $g = 0.32$) and 3.6 points for achievement (SE = 1.39; 95% CI [0.7, 6.5]; df = 18.9; $p = .018$; $g = 0.29$). Multiple-imputation estimates were 5.3, 4.8, 4.0, and 3.5 points for post-test comprehension, post-test achievement, follow-up comprehension, and follow-up achievement, respectively. Complete-case, pattern-mixture, school-random-intercept, longitudinal, and wild-bootstrap analyses retained the same directional conclusions.

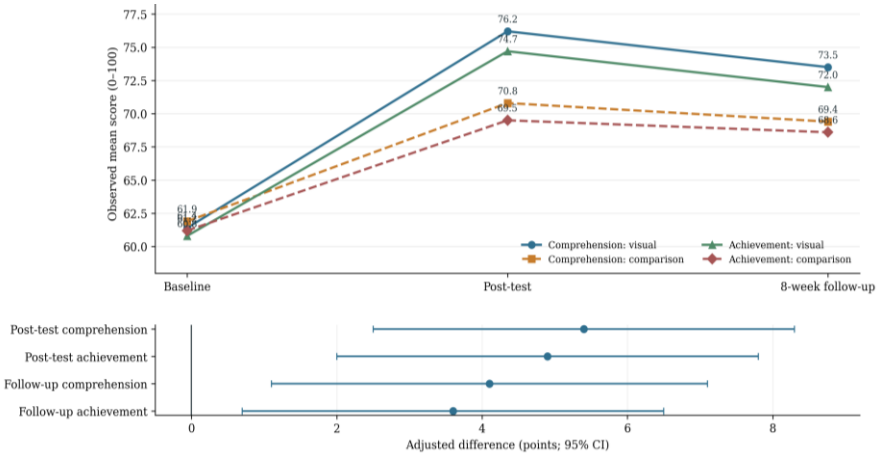


FIGURE 4. Observed score trajectories with baseline-adjusted contrasts and 95% confidence intervals.

Note. *The upper panel displays observed condition means. The lower panel displays adjusted condition differences from the reported multilevel models.*

Table 6. Baseline-adjusted estimates for co-primary academic outcomes.

Outcome	Primary estimate (SE)	95% CI	df	p	g / MI estimate
Post-test comprehension	5.4 (1.39)	[2.5, 8.3]	19.8	< .001	0.44 / 5.3
Post-test achievement	4.9 (1.38)	[2.0, 7.8]	19.5	.002	0.40 / 4.8
Follow-up comprehension	4.1 (1.43)	[1.1, 7.1]	18.7	.010	0.32 / 4.0
Follow-up achievement	3.6 (1.39)	[0.7, 6.5]	18.9	.018	0.29 / 3.5

3. SECONDARY AND CLASSROOM-PROCESS OUTCOMES

Visual literacy showed the largest prioritized secondary post-test estimate, followed by communication. Both remained significant after false-discovery-rate adjustment at post-test and follow-up. Imagination, engagement, and motivation also favored visual-storytelling instruction, but these outcomes were exploratory. Results are presented separately by outcome hierarchy and assessment occasion.

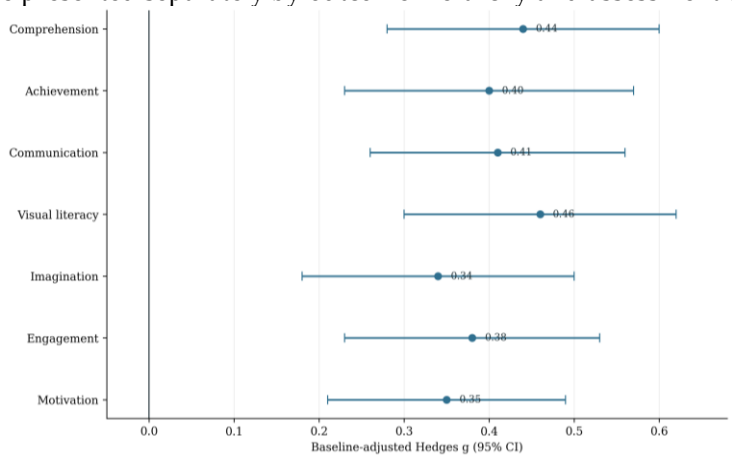


FIGURE 5. Baseline-adjusted post-test effect estimates across academic and multimodal domains.

Note. Confidence intervals summarize precision around standardized post-test effects. The plot excludes the generic retention row because persistence was reported separately for comprehension and achievement.

Table 7. Immediate and delayed estimates for secondary outcomes.

Outcome and status	Post-test adjusted estimate	Eight-week adjusted estimate
Visual literacy — prioritized	7.4 [4.0, 10.8]; g = 0.46; q < .001	4.8 [1.7, 7.9]; g = 0.30; q = .012
Communication — prioritized	6.8 [3.6, 10.0]; g = 0.41; q < .001	4.2 [1.2, 7.2]; g = 0.28; q = .015
Imagination — exploratory	5.9 [2.7, 9.1]; g = 0.34	3.4 [0.6, 6.2]; g = 0.25
Engagement — exploratory	6.2 [3.0, 9.4]; g = 0.38	3.7 [0.9, 6.5]; g = 0.27
Motivation — exploratory	0.28 [0.11, 0.45]; g = 0.35	0.19 [0.03, 0.35]; g = 0.24

4. EXPLORATORY SCHOOL-STAGE CONTRASTS

Primary and secondary school stages showed positive effects in the same direction. The condition-by-stage interaction coefficient was 0.8 points for comprehension (95% CI [−1.5, 3.1]; p = .48), 0.5 for achievement (95% CI [−1.6, 2.6]; p = .62), 1.3 for visual literacy (95% CI [−1.3, 3.9]; p = .31), and 0.7 for communication (95% CI [−1.7, 3.1]; p = .54). The intervals were wide and included zero; school-stage findings were therefore exploratory.

5. ARTIFACT AND QUALITATIVE FINDINGS

Four interpretive patterns were prominent. Sequence coherence organized causal and procedural relationships before writing; image-concept alignment made abstract content available for discussion; peer interpretation exposed ambiguity and prompted audience-oriented revision; and inclusive role allocation broadened participation beyond drawing confidence. Negative cases refined these interpretations: decorative detail sometimes displaced curriculum meaning, appearance-focused feedback reduced conceptual revision, limited lesson time constrained reworking, and drawing anxiety narrowed participation. These patterns described implementation and did not establish causal mediation.

Table 8. Integrated artifact dimensions, qualitative evidence, and boundary conditions.

Theme or process	Evidence denominator and coded account	Interpretive relevance	Negative or divergent case
Narrative organization	84 group storyboard drafts; 49 individual journals. "Putting the pictures in order helped me see what caused the change" (PS-I12).	Externalized causal and procedural relations before writing.	Unclear transitions produced fragmented explanations until prompts were added.
Concept visibility	92 final group artifacts; 26 observations. "The symbol showed the idea faster than my first sentence" (SS-I07).	Made abstract curriculum ideas available for discussion.	Some polished symbols were disconnected from curriculum content.
Audience awareness and revision	116 peer sheets; 8 focus groups. "They did not understand my panel, so I changed the caption" (PS-FG3).	Exposed ambiguity and strengthened audience clarity.	Appearance-focused feedback did not improve conceptual accuracy.
Reworking and academic transfer	121 revision sheets; 12 teacher logs. "The second version explained the idea more clearly" (SS-I21).	Connected group production with individual oral and written explanation.	Limited revision time reduced the depth of reworking.
Confidence-inclusive participation	30 interviews; detailed participation notes in 19 of 26 observations. "I was not good at drawing, but I could explain the meaning" (PS-I04).	Broadened contribution through sequencing, questioning, and explanation.	Participation narrowed when participants interpreted the task as an art-quality competition.

6. MIXED-METHODS INTEGRATION

The strongest convergence concerned comprehension, visual literacy, and communication: positive estimates corresponded with clearer sequence and image-concept alignment in artifacts and with participant accounts of explanation and revision. Delayed persistence showed complementarity because quantitative effects attenuated while participants described recurring images and story structures as retrieval cues. Divergence occurred when visually polished artifacts inaccurately represented curriculum content. School-stage evidence was directionally consistent but statistically imprecise. Individual artifact scores were connected to quantitative outcomes only where identifiers permitted; group artifacts informed class-level interpretation.

Table 9. Condensed integrated evidence matrix.

Analytical focus	Quantitative inference	Qualitative or artifact interpretation	Meta-inference
Academic learning	Moderate adjusted gains in comprehension and achievement.	Sequence and concept alignment organized content before explanation.	Convergence: representational quality was consistent with the academic pattern.
Persistence	Smaller positive effects remained at eight weeks.	Images and narrative structures were described as recall cues.	Complementarity: process evidence informed interpretation but did not establish mediation.
Multimodal competence	Visual literacy and communication had the strongest secondary effects.	Peer interpretation and revision improved captions, symbols, and clarity for the audience.	Convergence: outcome and process evidence addressed related instructional functions.
Divergence	No quantitative estimate represented decorative detail.	Some visually polished artifacts were conceptually inaccurate.	Aesthetic quality could conflict with disciplinary meaning.
School stage	Interaction estimates were small and imprecise.	Primary artifacts used more concrete sequence; secondary work required stronger evidential linkage.	Qualified convergence: common direction without reliable differential effectiveness.

IX. DISCUSSION

1. ACADEMIC OUTCOMES (RQ1)

The adjusted post-test estimates indicated moderate advantages in comprehension and academic achievement. The findings were consistent with controlled research in which learners improved when story production required active organization and explanation [2], [4], [15]. Related studies also reported benefits for EFL writing, speaking, and collaborative literacy [16]–[18]. The present study extended that evidence through within-school class randomization, an active comparison condition, prospective sample-size planning, finite-cluster correction, and prespecified co-primary outcomes. The effects were moderate in standardized terms; their practical importance is best interpreted against curriculum-specific benchmarks and the active nature of the comparison condition.

The pattern aligned with multimedia learning accounts because students selected, organized, and integrated visual and verbal information rather than passively viewing multimedia [24]–[26]. Dual coding and research on instructional visual displays provide complementary explanations for retrieval and representational comparison [31], [33]. Alternative explanations remain relevant. Intervention teachers received more preparation than comparison teachers, the program introduced novel production routines, and the outcome tests were aligned with the taught curriculum. Teacher and class effects were inseparable, and the small number of schools limited the ability to separate district, school, and teacher influences. These factors limit attribution of the complete observed difference to the representational cycle alone.

2. PERSISTENCE AT EIGHT WEEKS (RQ2)

Comprehension and achievement differences remained positive at eight weeks but declined in magnitude. This pattern was compatible with dual-coding and narrative accounts in which coordinated representations provide retrieval cues while remaining subject to ordinary forgetting [31], [32]. Multiple-imputation, longitudinal, and cluster-robust sensitivity analyses supported the direction of the delayed

estimates. The eight-week period did not establish maintenance across a full academic year, and the design did not isolate whether continued reinforcement accounted for variation in persistence.

3. EXPLORATORY SCHOOL-STAGE PATTERNS (RQ3)

Primary classes showed somewhat larger descriptive estimates for visual literacy, but interaction coefficients were small, imprecise, and statistically unreliable. Concrete sequence may have offered younger learners an accessible bridge from image to oral explanation, whereas secondary tasks required stronger links among visual narrative, disciplinary evidence, and written argument. This interpretation remained exploratory because only six classes per condition were available within each school stage.

4. ARTIFACT AND CLASSROOM PROCESSES (RQ4)

The convergent evidence was consistent with a role for coordinated representation in the instructional process. Sequence coherence corresponded to signaling principles [27], [28]. Image-concept alignment was also consistent with spatial-integration research [29], [30]. Audience interpretation and revision were consistent with structured cooperative learning [34]. We did not estimate these processes in a mediation model; therefore, they provided evidence-informed explanations rather than established causal pathways. A meaningful divergence occurred when visual polish coexisted with conceptual weakness, reinforcing the need to assess disciplinary meaning separately from artistic appearance.

5. METHODOLOGICAL AND THEORETICAL CONTRIBUTION

The study contributes methodologically through a prospectively planned cluster comparison, delayed academic measurement, validated artifact assessment, implementation monitoring, and explicit mixed-methods integration. Joint displays supported analysis and meta-inference rather than serving only as reporting devices [36]. The theoretical contribution is a bounded and testable representational-cycle model in which curriculum content is externalized, interpreted by an audience, revised, and translated into academic language. The present design supported this model's coherence but did not test mediation or isolate each component's independent effect.

X. EDUCATIONAL IMPLICATIONS

1. PRIMARY CLASSROOMS

In primary classrooms, the observed implementation pattern supported concrete prompts, short visual sequences, oral rehearsal, and explicit explanation of how images represented relationships. These practices were consistent with the intervention logic and require evaluation in other curricula before broader adoption.

2. SECONDARY CLASSROOMS

In secondary classrooms, the observed pattern supported tasks that connected visual narration with disciplinary evidence, precise captions, academic vocabulary, and individual written transfer. The results did not establish that any single component was independently effective.

3. IMPLEMENTATION BOUNDARIES

Across school stages, assessment criteria that separated conceptual alignment, sequence coherence, audience clarity, and revision from artistic polish were consistent with the strongest artifact evidence. Protected revision time, inclusive roles, and individual explanatory work were reasonable implementation principles, but teacher preparation, novelty, and contextual conditions remained possible contributors to the observed effects.

XI. METHODOLOGICAL STRENGTHS AND LIMITATIONS

Methodological strengths included prospective cluster-level sample-size planning, within-school class randomization, an active comparison condition, repeated academic assessment, masked outcome scoring,

formal finite-cluster correction, multiple imputation, validated rubric scoring, artifact-based evidence, and explicit mixed-methods integration. The 24 clusters supported the prespecified co-primary analysis but did not provide broad precision for teacher effects, school-stage interactions, or school-level variance estimation.

The study was conducted in six schools within one education district, limiting geographic, institutional, cultural, and linguistic generalizability. With one teacher per class, teacher and class effects were confounded. Intervention teachers also received more preparation than comparison teachers, creating an alternative explanation based on teacher support and expectancy. Teachers and students could not be blinded, and isolated visual-organizer use in two comparison lessons indicated minor contamination.

Pilot validation, reliability, and measurement-invariance evidence were acceptable, but transfer to other curricula and language contexts is not guaranteed. Curriculum alignment strengthened instructional validity and increased the likelihood of assessment alignment. Motivation and parts of the process evaluation relied on self-report or observer judgment and remained susceptible to expectancy and response bias. The false-discovery-rate procedure reduced, but did not eliminate, the risk associated with multiple secondary outcomes.

The follow-up lasted eight weeks and did not establish long-term maintenance. Qualitative maximum-variation sampling supported explanatory depth rather than prevalence estimation, and no statistical mediation analysis was conducted. Variation in revision time, confidence, teacher facilitation, and school context may have influenced the magnitude of the effect. These limitations bound the findings to a district-based, structured implementation of visual storytelling rather than a universal effect across educational settings.

XII. CONCLUSION

Visual-storytelling instruction produced moderate baseline-adjusted advantages in comprehension and academic achievement, with smaller but still positive differences at the eight-week follow-up. Visual literacy and communication showed the strongest secondary effects, indicating that the intervention supported not only curriculum attainment but also students' capacity to interpret, organize, and communicate disciplinary meaning through coordinated visual and verbal modes. These findings suggest that curriculum-aligned visual narration can strengthen academic and multimodal learning beyond active conventional instruction when students are required to connect images, sequence, captions, peer interpretation, revision, and individual explanation to explicit learning objectives.

Artifact and qualitative evidence were consistent with an interpretive representational cycle involving externalizing content, testing with an audience, revising ambiguous representations, and transferring into oral and written academic language. These processes clarified how the program operated in practice, although they were not tested as formal mediators and therefore should not be interpreted as established causal pathways. The six-school district setting, teacher–class confounding, unequal teacher preparation, locally adapted measures, possible expectancy effects, and the limited follow-up period constrain generalization. Within these boundaries, the study credibly contributes to research on multimodal language and curriculum learning and offers a practical basis for further controlled evaluation across more diverse educational settings.

XIII. DATA AVAILABILITY AND SUPPLEMENTARY MATERIALS

De-identified data, analysis code, scoring rubrics, and supporting research materials are available through the corresponding author's institution under controlled access. Requests are subject to ethics and data-protection review. We do not publicly share qualitative transcripts, student artifacts, consent records, or re-identification files.

Author Contributions

The author conducted the conceptualization, methodology, data analysis, investigation, writing, review, editing, and final approval of the manuscript.

Funding

This research received no external funding.

Data Availability

The dataset will be available from the author upon reasonable request.

Conflicts of Interest

The author declares no conflict of interest.

REFERENCES

[1]. J. Wu and D. T. V. Chen, "A systematic review of educational digital storytelling," Computers & Education, vol. 147, Art. no. 103786, 2020, doi: 10.1016/j.compedu.2019.103786.

[2]. K.-P. Liu, S.-J. D. Tai, and C.-C. Liu, "Enhancing language learning through creation: The effect of digital storytelling on student learning motivation and performance in a school English course," Educational Technology Research and Development, vol. 66, no. 4, pp. 913–935, 2018, doi: 10.1007/s11423-018-9592-z.

[3]. N. Parsazadeh, P.-Y. Cheng, T.-T. Wu, and Y.-M. Huang, "Integrating computational thinking concept into digital storytelling to improve learners' motivation and performance," Journal of Educational Computing Research, vol. 59, no. 3, pp. 470–495, 2021, doi: 10.1177/0735633120967315.

[4]. H. C. Sarica and Y. K. Usluel, "The effect of digital storytelling on visual memory and writing skills," Computers & Education, vol. 94, pp. 298–309, 2016, doi: 10.1016/j.compedu.2015.11.016.

[5]. A. Schmoelz, "Enabling co-creativity through digital storytelling in education," Thinking Skills and Creativity, vol. 28, pp. 1–13, 2018, doi: 10.1016/j.tsc.2018.02.002.

[6]. M. Palioura and C. Dimoulas, "Digital storytelling in education: A transmedia integration approach for the non-developers," Education Sciences, vol. 12, no. 8, Art. no. 559, 2022, doi: 10.3390/educsci12080559.

[7]. A. A. S. Tabieh, M. M. Al-Hileh, H. M. J. Abu Afifa, and H. Y. Abuzagha, "The effect of using digital storytelling on developing active listening and creative thinking skills," European Journal of Educational Research, vol. 10, no. 1, pp. 13–21, 2021, doi: 10.12973/eu-jer.10.1.13.

[8]. N. E. Ozen and E. Duran, "Contribution of digital storytelling to creative thinking skills," Turkish Journal of Education, vol. 10, no. 4, pp. 297–318, 2021, doi: 10.19128/turje.909865.

[9]. K. Ramalingam, Y. K. Jiar, and S. Mathiyazhagan, "Speaking skills enhancement through digital storytelling among primary school students in Malaysia," International Journal of Learning, Teaching and Educational Research, vol. 21, no. 3, pp. 22–35, 2022, doi: 10.26803/ijlter.21.3.2.

[10]. P. R. Lim and N. Md Noor, "Digital storytelling as a creative teaching method in promoting secondary school students' writing skills," International Journal of Interactive Mobile Technologies, vol. 13, no. 7, pp. 117–128, 2019, doi: 10.3991/ijim.v13i07.10798.

[11]. D. G. Nikleva and F. J. Rodriguez-Munoz, "Effects of a visual literacy program for the improvement of reading comprehension in primary and secondary school students," L1-Educational Studies in Language and Literature, vol. 22, no. 1, pp. 1–19, 2022, doi: 10.21248/l1esll.2022.22.1.394.

[12]. D. Kim, D. K. Yatsu, and Y. Li, "A multimodal model for analyzing middle school English language learners' digital stories," International Journal of Educational Research Open, vol. 2, Art. no. 100067, 2021, doi: 10.1016/j.ijedro.2021.100067.

[13]. H. Dahlstrom and U. Damber, "Meanings made in students' multimodal digital stories: Resources, popular culture, and values," Designs for Learning, vol. 12, no. 1, pp. 45–55, 2020, doi: 10.16993/dfl.145.

[14]. J. Kedra, "What does it mean to be visually literate? Examination of visual literacy definitions in a context of higher education," Journal of Visual Literacy, vol. 37, no. 2, pp. 67–84, 2018, doi: 10.1080/1051144X.2018.1492234.

[15]. K. Hava, "Exploring the role of digital storytelling in student motivation and satisfaction in EFL education," Computer Assisted Language Learning, vol. 34, no. 7, pp. 958–978, 2021, doi: 10.1080/09588221.2019.1650071.

[16]. J. S. Fu, S.-H. Yang, and H.-C. Yeh, "Exploring the impacts of digital storytelling on English as a foreign language learners' speaking competence," Journal of Research on Technology in Education, vol. 54, no. 5, pp. 679–694, 2022, doi: 10.1080/15391523.2021.1911008.

[17]. D. Y. Guzmán Gámez and J. A. Moreno Cuellar, "The use of Plotagon to enhance the English writing skill in secondary school students," Profile: Issues in Teachers' Professional Development, vol. 21, no. 1, pp. 139–153, 2019, doi: 10.15446/profile.v21n1.71721.

[18]. P. A. González Mesa, "Digital storytelling: Boosting literacy practices in students at A1-level," HOW, vol. 27, no. 1, pp. 83–104, 2020, doi: 10.19183/how.27.1.505.

[19]. W.-Y. Hwang, R. Shadiev, J.-L. Hsu, Y.-M. Huang, G.-L. Hsu, and Y.-C. Lin, "Effects of storytelling to facilitate EFL speaking using Web-based multimedia system," Computer Assisted Language Learning, vol. 29, no. 2, pp. 215–241, 2016, doi: 10.1080/09588221.2014.927367.

[20]. N. Z. L. Lim, A. Zakaria, and V. Aryadoust, "A systematic review of digital storytelling in language learning in adolescents and adults," Education and Information Technologies, vol. 27, pp. 6125–6155, 2022, doi: 10.1007/s10639-021-10861-0.

[21]. C. K. Cheung and A. D. Jhaveri, "Developing students' critical thinking skills through visual literacy in the new secondary school curriculum in Hong Kong," Asia Pacific Journal of Education, vol. 36, no. 3, pp. 379–389, 2016, doi: 10.1080/02188791.2014.959470.

[22]. V. Ozsoy and S. Saribas, "Developing visual literacy skills in teacher education: Different ways of looking at visual images," Educational Policy Analysis and Strategic Research, vol. 16, no. 3, pp. 67–88, 2021, doi: 10.29329/epasr.2021.373.5.

[23]. K. K. Matusiak, C. Heinbach, A. Harper, and M. Bovee, "Visual literacy in practice: Use of images in students' academic work," College & Research Libraries, vol. 80, no. 1, pp. 123–139, 2019, doi: 10.5860/crl.80.1.123.

[24]. R. E. Mayer, "Cognitive theory of multimedia learning," in The Cambridge Handbook of Multimedia Learning, 3rd ed., R. E. Mayer and L. Fiorella, Eds. Cambridge, U.K.: Cambridge University Press, 2021, pp. 57–72, doi: 10.1017/9781108894333.008.

[25]. M. Noetel, S. Griffith, O. Delaney, N. R. Harris, T. Sanders, P. Parker, B. del Pozo Cruz, and C. Lonsdale, "Multimedia design for learning: An overview of reviews with meta-meta-analysis," Review of Educational Research, vol. 92, no. 3, pp. 413–454, 2022, doi: 10.3102/00346543211052329.

[26]. B. Ceken and N. Taskin, "Multimedia learning principles in different learning environments: A systematic review," Smart Learning Environments, vol. 9, Art. no. 19, 2022, doi: 10.1186/s40561-022-00200-2.

[27]. S. Schneider, M. Beege, S. Nebel, and G. D. Rey, "A meta-analysis of how signaling affects learning with media," Educational Research Review, vol. 23, pp. 1–24, 2018, doi: 10.1016/j.edurev.2017.11.001.

[28]. J. Richter, K. Scheiter, and A. Eitel, "Signaling text-picture relations in multimedia learning: A comprehensive meta-analysis," Educational Research Review, vol. 17, pp. 19–36, 2016, doi: 10.1016/j.edurev.2015.12.003.

[29]. N. L. Schroeder and A. T. Cenkci, "Spatial contiguity and spatial split-attention effects in multimedia learning environments: A meta-analysis," Educational Psychology Review, vol. 30, no. 3, pp. 679–701, 2018, doi: 10.1007/s10648-018-9435-9.

[30]. N. L. Schroeder and A. T. Cenkci, "Do measures of cognitive load explain the spatial split-attention principle in multimedia learning environments? A systematic review," Journal of Educational Psychology, vol. 112, no. 2, pp. 254–270, 2020, doi: 10.1037/edu0000372.

[31]. J. M. Clark and A. Paivio, "Dual coding theory and education," Educational Psychology Review, vol. 3, no. 3, pp. 149–210, 1991, doi: 10.1007/BF01320076.

[32]. J. Bruner, "The narrative construction of reality," Critical Inquiry, vol. 18, no. 1, pp. 1–21, 1991, doi: 10.1086/448619.

[33]. A. Renkl and K. Scheiter, "Studying visual displays: How to instructionally support learning," Educational Psychology Review, vol. 29, no. 4, pp. 599–621, 2017, doi: 10.1007/s10648-015-9340-4.

[34]. R. M. Gillies, "Cooperative learning: Review of research and practice," Australian Journal of Teacher Education, vol. 41, no. 3, pp. 39–54, 2016, doi: 10.14221/ajte.2016v41n3.3.

[35]. V. Braun and V. Clarke, "Reflecting on reflexive thematic analysis," Qualitative Research in Sport, Exercise and Health, vol. 11, no. 4, pp. 589–597, 2019, doi: 10.1080/2159676X.2019.1628806.

[36]. T. C. Guetterman, S. Fàbregues, and R. Sakakibara, "Visuals in joint displays to represent integration in mixed methods research: A methodological review," Methods in Psychology, vol. 5, Art. no. 100080, 2021, doi: 10.1016/j.metip.2021.100080.

[37]. K. F. Schulz, D. G. Altman, and D. Moher, "CONSORT 2010 statement: Updated guidelines for reporting parallel group randomized trials," BMJ, vol. 340, Art. no. c332, 2010, doi: 10.1136/bmj.c332.

[38]. M. K. Campbell, G. Piaggio, D. R. Elbourne, and D. G. Altman, "CONSORT 2010 statement: Extension to cluster randomized trials," BMJ, vol. 345, Art. no. e5661, 2012, doi: 10.1136/bmj.e5661.

[39]. T. C. Hoffmann et al., "Better reporting of interventions: Template for intervention description and replication (TIDieR) checklist and guide," BMJ, vol. 348, Art. no. g1687, 2014, doi: 10.1136/bmj.g1687.

APPENDIX A. INSTRUMENT SCORING, ADMINISTRATION, AND VALIDATION EVIDENCE

Instrument/source	Representative content and scoring	Administration and transformation	Validation evidence
Comprehension test	Identify relationships between events and justify interpretation; correct = 1, incorrect/blank = 0.	28 grade-band items; 35 minutes; raw total ÷ 28 × 100; parallel follow-up form.	KR-20 .83; S-CVI .92; difficulty .34–.77; discrimination .29–.58.
Achievement test	Selected-response and constructed-response items aligned with the unit blueprint.	30 items; 40 minutes; scored points ÷ maximum × 100; parallel follow-up form.	KR-20 .81; S-CVI .89; difficulty .31–.79; discrimination .27–.55.
Visual-literacy rubric	Symbol relevance, sequence clarity, image-concept alignment, audience readability; anchors 0–4.	Three masked raters; dimension total ÷ 16 × 100; 25% double-scored.	ICC(2,k) .85, 95% CI [.80, .89].
Communication rubric	Audience clarity, academic vocabulary, evidence linkage, explanatory coherence; anchors 0–4.	Three masked raters; dimension total ÷ 16 × 100; 25% double-scored.	ICC(2,k) .82, 95% CI [.76, .87].
Imagination scale	Twelve curriculum-bounded items assessing alternative representation and interpretive flexibility.	Student self-report at baseline, post-test, and follow-up; total rescaled to 0–100.	Omega .86; CFI .95; RMSEA .052; metric invariance.
Engagement checklist	Ten indicators covering attention, contribution, peer response, and persistence.	Trained observers sampled all three occasions; total rescaled to 0–100.	Kappa = .79; expert review and observer calibration.
Motivation scale	Eight items on interest, value, effort, and willingness to continue.	Student self-report at three occasions; item mean retained on 1–5 scale.	Omega .88; metric invariance.
Qualitative guides	Prompts addressed representation, explanation, revision, barriers, and transfer.	Semi-structured administration; de-identified transcription and reflexive thematic analysis.	Audit trail, negative-case review, and cross-source comparison.

APPENDIX B. STATISTICAL MODELS, IMPUTATION DIAGNOSTICS, AND SENSITIVITY OUTPUTS

Analysis	Specification or diagnostic	Result	Interpretation
Post-test comprehension	Class random intercept; school fixed effects; CR2/Satterthwaite	Class variance 4.08; residual 153.12; ICC .026; 5.4 (SE 1.39), CI [2.5, 8.3]	Positive co-primary effect
Post-test achievement	Same primary specification	Class variance 4.31; residual 160.08; ICC .026; 4.9 (SE 1.38), CI [2.0, 7.8]	Positive co-primary effect

Analysis	Specification or diagnostic	Result	Interpretation
Follow-up comprehension	Baseline-adjusted delayed model	Class variance 4.86; residual 158.42; ICC .030; 4.1 (SE 1.43), CI [1.1, 7.1]	Smaller positive delayed effect
Follow-up achievement	Baseline-adjusted delayed model	Class variance 4.55; residual 162.10; ICC .027; 3.6 (SE 1.39), CI [0.7, 6.5]	Smaller positive delayed effect
Multiple imputation	40 datasets; 30 iterations; robust variance pooled with Rubin's rules	Stable chains; close observed/imputed distributions; low-to-moderate missing information	Matched complete-case conclusions
School random-intercept sensitivity	School random intercept; class nested within school	Condition estimates remained consistent with fixed-school models	Inference not dependent on school specification
Wild cluster bootstrap	9,999 class-level replications	Intervals retained direction and significance pattern	Finite-cluster robustness supported
Longitudinal sensitivity	Condition × occasion model with repeated observations	Post-test effects attenuated but remained positive at follow-up	Consistent with separate-occasion models
Diagnostics	Residual, influence, convergence, and collinearity checks	No singular fit; assumptions acceptable; VIF < 1.4	No major model-instability signal

XIV. SUPPLEMENTARY APPENDIX C. EXPLORATORY SCHOOL-STAGE RESULTS

Outcome	Primary-stage effect	Secondary-stage effect	Interaction coefficient (95% CI)	p
Visual literacy	8.2	6.9	1.3 [−1.3, 3.9]	.31
Communication	7.1	6.4	0.7 [−1.7, 3.1]	.54
Comprehension	5.6	4.8	0.8 [−1.5, 3.1]	.48
Achievement	5.1	4.6	0.5 [−1.6, 2.6]	.62

APPENDIX D. QUALITATIVE SAMPLING, PARTICIPANT CODES, AND EVIDENCE-SOURCE MATRIX

Evidence source	Quantity and coding	Selection or unit represented	Analytical use
Student interviews	30; PS-I/SS-I codes; 15 primary, 15 secondary; 16 intervention, 14 comparison	Maximum variation by condition, stage, sex, attendance, and outcome profile	Individual experience, transfer, barriers, and negative cases
Focus groups	8 groups × 6; PS-FG/SS-FG codes; balanced by condition and stage	Purposive classroom-level selection; one session per participant	Peer interpretation, audience response, and implementation contrast
Classroom observations	26 lessons; 13 per condition; O01–O26	Coverage across phases and stages; 19 records had detailed participation notes	Fidelity, teacher practice, participation, and contamination

https://doi.org/10.48161/aaj.v3n4a100

Evidence source	Quantity and coding	Selection or unit represented	Analytical use
Storyboard drafts	84 group drafts	Samples across schools, stages, and performance levels	Sequence coherence and causal organization
Final artifacts	92 group products	Linked to class and group identifiers	Image-concept alignment, symbolism, and audience clarity
Reflection journals	49 individual journals	Purposive sample across stage and outcome profile	Individual interpretation and transfer
Revision sheets	121 records	Group or individual revision events	Nature and depth of content-related revision
Peer-interpretation sheets	116 records	Peer-response episodes linked to class and group	Audience understanding and ambiguity detection
Teacher logs	12 intervention-class logs	One per intervention teacher/class	Dose, fidelity, revision time, and implementation variation