BOOKAGENT: Orchestrating Safety-Aware Visual Narratives via Multi-Agent Cognitive Calibration

Bo Gao
Carnegie Mellon University
bogao@andrew.cmu.edu

Chang Liu
University of Science and Technology of China
lc980413@mail.ustc.edu.cn

Yuyang Miao
Imperial College London
ym520@ic.ac.uk

Siyuan Ma
Nanyang Technological University
MASI0004@e.ntu.edu.sg

Ser-Nam Lim*
University of Central Florida
sernam@gmail.com

Abstract

Recent advancements in Large Generative Models (LGMs) have revolutionized multi-modal generation. However, generating illustrated storybooks remains an open challenge, where prior works mainly decompose this task into separate stages, and thus, holistic multi-modal grounding remains limited. Besides, while safety alignment is studied for text- or image-only generation, existing works rarely integrate child-specific safety constraints into narrative planning and sequence-level multi-modal verification. To address these limitations, we propose BOOKAGENT, a safety-aware multi-agent collaboration framework designed for high-quality, safety-aware visual narratives. Different from prior story visualization models that assume a fixed storyline sequence, BOOKAGENT targets end-to-end storybook synthesis from a user draft by jointly planning, scripting, illustrating, and globally repairing inconsistencies. To ensure precise multi-modal grounding, BOOKAGENT dynamically calibrates page-level alignment between textual scripts and visual layouts. Furthermore, BOOKAGENT calibrates holistic consistency from the temporal dimension, by verifying-then-rectifying global inconsistencies in character identity and storytelling logic. Extensive experiments demonstrate that BOOKAGENT significantly outperforms current methods in narrative coherence, visual consistency, and safety compliance, offering a robust paradigm for reliable agents in complex multi-modal creation. The implementation will be publicly released at https://github.com/bogao-code/BookAgent/tree/main.

1 Introduction

Visual narratives, ranging from illustrated storybooks to complex comics, represent a fundamental medium of human communication that combines both linguistic storytelling and visual imagination. In the era of Large Generative Models (LGMs) (Ho et al., 2020; Rombach et al., 2022; Song et al., 2021; Ho and Salimans, 2022; Touvron et al., 2023a,b; Bai et al., 2023a,b), we have witnessed astonishing capabilities in both textual and visual content generation. The convergence of these capabilities enables the potential of translating abstract ideas into coherent, multi-modal storybooks. However, automating this process with existing methods is not a trivial task. It requires an integrated system to generate coherent narrative flow, ensure semantic alignment between text and pixels, and obey strict safety standards.

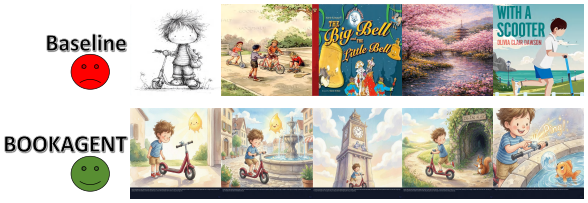


Figure 1: **Teaser: Long-horizon story consistency requires collaboration.** Given the same multi-step story prompt with strict ordering and counting constraints, a single-pass baseline generation fails to preserve character identity and temporal consistency across pages (top). In contrast, BOOKAGENT leverages multi-agent collaboration to maintain stable characters, correct event order, and consistent visual attributes throughout the entire story sequence (bottom).

LLM-based agents have recently demonstrated strong capability in decomposing complex goals into executable plans and orchestrating multi-step generation workflows in purely textual settings, e.g., by interleaving reasoning traces with tool actions (Yao et al., 2023b) or by learning when and how to call external APIs (Schick et al., 2023). Visual narrative tasks like storybook generation pushes such agentic reasoning into a genuinely multi-modal regime, which is normally conducted in a stage-by-stage manner by splitting the generation processes of visual and textual contents. This process requires three key aspects to address, i.e.,

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 2275–2292

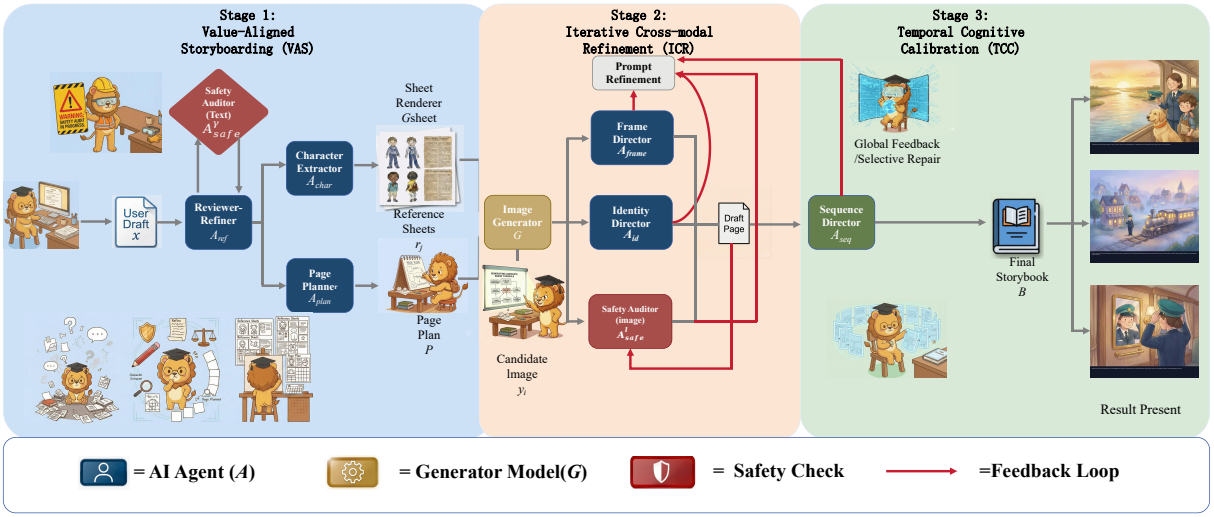


Figure 2: **Overview of BOOKAGENT.** The framework follows a closed-loop, multi-agent architecture with three mechanisms. **Stage 1: Value-Aligned Storyboarding (VAS)** audits the input story against safety guardrails and structures it into a page plan with extracted characters and a reusable character sheet. **Stage 2: Iterative Cross-modal Refinement (ICR)** iteratively refines page prompts and generates candidate images, guided by frame-, identity-, and sequence-level directors with multimodal safety auditing, to improve page-level grounding and visual quality. **Stage 3:Temporal Cognitive Calibration (TCC) performs global review over the full sequence to detect and correct long-horizon inconsistencies in character identity and narrative logic.**

cross-modal alignment, *global consistency*, and *safety*. Regarding *cross-modal alignment*, most existing works (Maharana et al., 2022; Liu et al., 2024) assume a given storyline sequence and produce visual content as a separate step, with more recent efforts incorporating stronger language understanding capability of LLMs into the agentic system (Shen and Elhoseiny, 2025). Despite these advances, the coupling between linguistic and visual narratives is still weak, where visual contents rarely provide structured feedback to revise the script, making bi-directional grounding and page-level calibration under-specified. Considering *global consistency*, this aspect still remains challenging beyond local alignment, as story-level generation requires long-range reasoning over entity identity, coreference, and causal relations across pages. Some works (Tao et al., 2024) mainly rely on history conditioning, which can still suffer from appearance drift and role entanglement as the sequence length grows. Therefore, explicit sequence-level verification-and-repair that jointly reasons over text, images, and multi-character coreference is expected. **Third**, domain-specific *safety* is under-explored, particularly for child-oriented storybooks. While the safety and NLP community have put great emphasis on addressing general-purpose NSFW generation (Poppi et al., 2024; Li et al., 2024) and child-safe text generation, respectively (Nayeem and Rafiei, 2024) existing methods seldom integrate child-specific safety constraints into narrative planning and global consistency checking, leaving safety to generic post-hoc filters. Ideally, a solution should function as a cohesive cognitive system, unifying the planning capability of LLM-based agents with multi-modal generators, and closing the loop with page-level verification and sequence-level refinement under explicit child-safety guardrails.

To bridge these gaps, we introduce BOOKAGENT, a comprehensive multi-agent framework that treats storybook generation as a collaborative, safety-aware cognitive process. Unlike previous works based on a separate paradigm that first fixes a storyline and then autoregressively produces content sequences, our approach is implemented in an end-to-end paradigm, meaning that it unifies text and image generation through a closed-loop architecture, along with three distinct mechanisms, namely Value-Aligned Storyboarding (VAS), Iterative Cross-modal Refinement (ICR), and Temporal Cognitive Calibration (TCC). To ensure safety and value alignment of the inputs, VAS serves as the component that assists agents to rigorously audit and structure the narrative against safety guardrails before visualization begins. ICR is the dynamic feedback loop where the system generates, evaluates, and re-generates page-level content, ensuring precise grounding between the script and the visual layout. To enforce long-term logic, TCC performs

global reasoning that reviews the entire generated sequence to identify and rectify inconsistencies in character identity and storytelling flow. Extensive experiments indicate that BOOKAGENT not only generates aesthetically pleasing storybooks, but also sets a new standard for narrative coherence and safety compliance. It is worth noting that BOOKAGENT is the first attempt to perform storybook content generation in an end-to-end manner, rather than in a stage-by-stage way, meaning that simultaneous multimodal content generation shows a solid reference to facilitate the inter- and intra-consistency across both modalities.

2 Related Work

Agent-based Storybook Synthesis. Research in cross-modal storybook synthesis has evolved from text-only planning and independent image sequence generation to recent agentic workflows (Shinn et al., 2023; Park et al., 2023; Patil et al., 2024; Gou et al., 2024; Wang et al., 2024; Yang et al., 2024; Surís et al., 2023; Hao et al., 2023; Yao et al., 2023a; Zhou et al., 2024; Driess et al., 2023) that bridge reasoning with controllable synthesis. Early efforts focused on enforcing textual coherence through hierarchical structures, where the key challenges lie at different techniques, e.g., recurrent networks (Li et al., 2019), Transformer (Maharana et al., 2022), memory module (Rahman et al., 2023), masking mechanism (Tao et al., 2024), to establish the mappings between the storyline and image sequences. With the development of agentic system (Yao et al., 2023b; Schick et al., 2023), all aforementioned capabilities can be effectively orchestrated into one united system. In doing so, TaleCrafter (Gong et al., 2023) combines story-to-prompt and layout generation agents; *StoryGPT-V* (Shen and Elhoseiny, 2025) utilizes an LLM to align character descriptions with diffusion models. Unlike these works, which often assume a fixed storyline or a one-way generation pipeline, BOOKAGENT targets end-to-end synthesis to simultaneously produce visual and textual contents.

Safety-Aware Content Generation. Safety alignment is fundamentally vital for child-central content generation. This particular domain-specific requirement has motivated a line of work in verifying non-toxic generation. Specifically, Safe Latent Diffusion (Schramowski et al., 2023) uses language-defined safety concepts to guide sampling away from inappropriate degeneration. Safe-CLIP (Poppi et al., 2024) unlearns toxic associations in the embedding space. DUO (Park et al., 2024) applies preference optimization to directly unlearn unsafe features. RECE (Gong et al., 2024) utilizes closed-form concept erasure to prevent the regeneration of erased concepts. Different from these approaches that typically operate as post-hoc filters or single-turn constraints, BOOKAGENT integrates safety directly into the narrative planning and sequence verification stages, enabling the early prevention of unsafe plot trajectories and the global repair of value-misaligned content in long-form narratives.

3 Methodology

3.1 Preliminaries

We formulate the storybook synthesis problem as a constrained optimization task. Let x denote the user-provided draft, I_0 an optional inspiration image, K the target page count, and s a global style descriptor. Our goal is to generate a storybook $\mathcal{B} \triangleq \{(t_i, y_i)\}_{i=1}^{K}$, where t_i represents the narrative script and y_i the illustration for the i-th page.

The system is orchestrated by a set of specialized agents driven by Multimodal LLMs: a *Reviewer–Refiner* $\mathcal{A}_{\text{ref}}$, a *Page Planner* $\mathcal{A}_{\text{plan}}$, a *Character Extractor* $\mathcal{A}_{\text{char}}$, a *Frame Director* $\mathcal{A}_{\text{frame}}$, an *Identity Director* $\mathcal{A}_{\text{id}}$, and a *Sequence Director* $\mathcal{A}_{\text{seq}}$. Visual synthesis is performed by a reference-conditioned *Image Generator* $\mathcal{G}$ and a *Character Sheet Renderer* $\mathcal{G}_{\text{sheet}}$. Safety is enforced by text and image auditors, formulated as:

$$\begin{aligned} \mathcal{A}^{T}_{\text{safe}}(\cdot) &\rightarrow (\mathcal{S}_T, \rho_T), \\ \mathcal{A}^{I}_{\text{safe}}(\cdot) &\rightarrow (\mathcal{S}_I, \rho_I), \end{aligned} \tag{1}$$

where $\mathcal{S} \in \{0, 1\}$ denotes the binary safety decision and ρ represents the reasoning (e.g., "violent content detected").

We aim to maximize the overall quality considering faithfulness, identity consistency, and sequence coherence, subject to hard safety constraints:

$$\begin{aligned} \max_{\hat{x},\, \{y_i\}_{i=1}^{K}} \sum_{i=1}^{K} [\alpha(t_i, y_i) + \eta(y_i, \{r_j\})] + \lambda\beta(\mathcal{B}) \\ \text{s.t.} \mathcal{S}_T(\hat{x}) = 1, \mathcal{S}_I(y_i) = 1, \forall i = 1, \ldots, K, \end{aligned} \tag{2}$$

where $\alpha(\cdot)$ measures text–image faithfulness, $\eta(\cdot)$ measures identity consistency, and $\beta(\cdot)$ measures global sequence continuity. We approximate this objective via a three-stage hierarchical workflow: Value-Aligned Storyboarding (VAS), Itera-

Table 1: Role-based decomposition with fixed I/O contracts enabling verification, identity anchoring, selective repair, and child-safety enforcement.

Role	Symbol	I/O contract	Primary responsibility
Reviewer–Refiner	$\mathcal{A}_{\text{ref}}$	**In:** x, K, s **Out:** $\hat{x}$, mode, feedback	Review the draft and either lightly polish or strongly rewrite it to match K pages; improve coherence and reduce ambiguity; enforce ≤ 5 recurring characters.
Page Planner	$\mathcal{A}_{\text{plan}}$	**In:** $\hat{x}, I_0, K, s$ **Out:** $\mathcal{P} = \{(t_i, p_i^{(0)})\}_{i=1}^{K}$	Decompose the refined story into page texts and initial prompts encoding local semantics + global style.
Character Extractor	$\mathcal{A}_{\text{char}}$	**In:** $\hat{x}, s$ **Out:** $\mathcal{C} = \{c_j\}_{j=1}^{C}$	Extract up to $C \leq 5$ recurring characters with stable ids and concise visual descriptors (species/colors/clothing) for identity anchoring.
Character Sheet Renderer	$\mathcal{G}_{\text{sheet}}$	**In:** d_j (visual descriptor), s **Out:** r_j	Render a clean neutral-background reference sheet for each recurring character; optionally reuse user-provided inspiration image as the main character sheet.
Image Generator (ref-conditioned)	$\mathcal{G}$	**In:** $p_i^{(r)}$, refs $\mathcal{R}_i$ **Out:** $y_i^{(r)}$	Generate illustration candidates conditioned on the current prompt and a set of visual references (character sheets + short-term context).
Frame Director	$\mathcal{A}_{\text{frame}}$	**In:** $(t_i, y_i^{(r)})$ **Out:** $\alpha_i^{(r)}, \Delta_i^{(r)}$	Verify page-level text–image faithfulness; attribute actionable issues for prompt revision.
Identity Director	$\mathcal{A}_{\text{id}}$	**In:** $(t_i, y_i^{(r)}, \{r_j\}_{j=1}^{C}, s)$ **Out:** $\eta_i^{(r)}, \Omega_i^{(r)}$	Verify character identity and key recurring attributes against the reference sheets (e.g., species/color/clothing drift, missing/extra main characters).
Sequence Director	$\mathcal{A}_{\text{seq}}$	**In:** $\mathcal{B}^{(m)} = \{(t_i, y_i)\}_{i=1}^{K}, s$ **Out:** $\beta^{(m)}, \Gamma^{(m)}, \mathcal{I}^{(m)}$	Verify cross-page continuity (identity/props/style) and attribute failures to a sparse set of problem pages for selective repair.
Safety Auditor (Text)	$\mathcal{A}_{\text{safe}}^{T}$	**In:** $z \in \{x, \hat{x}\}$ **Out:** $\mathcal{S}_T(z)$, ρ_T	Audit child-safety of text; if unsafe, sanitize via constrained rewriting.
Safety Auditor (Image)	$\mathcal{A}_{\text{safe}}^{I}$	**In:** $y_i^{(r)}$ **Out:** $\mathcal{S}_I(y)$, ρ_I	Audit child-safety of images; reject unsafe candidates and harden prompts with explicit safety constraints.

tive Cross-modal Refinement (ICR), and Temporal Cognitive Calibration (TCC).

3.2 Value-Aligned Storyboarding

This stage transforms the raw draft into a structured blueprint and establishes visual anchors. Since user drafts are often coarse, the *Reviewer–Refiner* $\mathcal{A}_{\text{ref}}$ rewrites the draft x to match K pages:

$$\hat{x} = \mathcal{A}_{\text{ref}}(x, K, s). \tag{3}$$

The output $\hat{x}$ is verified by $\mathcal{A}_{\text{safe}}^{T}$; if $\mathcal{S}_T(\hat{x}) = 0$, the refiner utilizes the safety critique ρ_T to guide constrained rewriting until standards are met.

Next, we extract recurring characters and generate canonical reference sheets prior to page generation. The *Character Extractor* $\mathcal{A}_{\text{char}}$ identifies up to C main characters from the refined story:

$$\mathcal{C} = \{c_j\}_{j=1}^{C} = \mathcal{A}_{\text{char}}(\hat{x}, s), \tag{4}$$

where each c_j contains a stable identity and a concise visual descriptor d_j. The *Character Sheet Renderer* $\mathcal{G}_{\text{sheet}}$ then produces neutral-background reference images:

$$r_j = \mathcal{G}_{\text{sheet}}(d_j, s), \quad \forall j \in \{1, \ldots, C\}. \tag{5}$$

These reference sheets $\{r_j\}_{j=1}^{C}$ serve as the ground truth for identity verification in subsequent stages. Finally, the *Page Planner* $\mathcal{A}_{\text{plan}}$ decomposes the story into a page-wise plan $\mathcal{P}$:

$$\mathcal{P} \triangleq \{(t_i, p_i^{(0)})\}_{i=1}^{K} = \mathcal{A}_{\text{plan}}(\hat{x}, I_0, K, s), \tag{6}$$

where $p_i^{(0)}$ is the initial prompt for page i, encoding both local semantics and global style requirements.

3.3 Iterative Cross-modal Refinement

Generating high-quality storybook content requires iterative optimization. We employ a budgeted generate–verify–revise loop. For each page i, we first retrieve relevant character sheets $\mathcal{R}_i = \{r_j \mid c_j \in \text{Entities}(t_i)\}$ based on the narrative. At attempt $r < R$, we generate an image using the *Image Generator* $\mathcal{G}$, formulated by:

$$y_i^{(r)} \sim \mathcal{G}(p_i^{(r)}, \mathcal{R}_i). \tag{7}$$

We then execute a dual-branch verification. The *Frame Director* $\mathcal{A}_{\text{frame}}$ evaluates faithfulness, outputting score $\alpha_i^{(r)}$ and semantic issues $\Delta_i^{(r)}$. Simultaneously, the *Identity Director* $\mathcal{A}_{\text{id}}$ checks consistency against $\mathcal{R}_i$, yielding identity score $\eta_i^{(r)}$ and issues $\Omega_i^{(r)}$. Afterwards, we unify these feedbacks to update the prompt $p_i^{(r+1)}$, utilizing a local memory $\mathcal{M}_i$ to accumulate historical constraints and prevent regression:

$$p_i^{(r+1)} = \begin{cases} p_i^{(r)} \oplus \Psi(\rho_I), & \text{if } \mathcal{S}_I(y_i^{(r)}) = 0, \\ p_i^{(r)} \oplus \Phi(\Delta_i^{(r)}, \Omega_i^{(r)}, \mathcal{M}_i), & \text{otherwise,} \end{cases} \tag{8}$$

where $\Psi(\cdot)$ converts safety reasoning ρ_I into explicit negative constraints, and $\Phi(\cdot)$ aggregates current semantic/identity critiques with historical issues stored in $\mathcal{M}_i$. We accept a candidate if it is

safe, faithful ($\alpha_i^{(r)} \geq \tau_\alpha$), and identity-consistent ($\eta_i^{(r)} \geq \tau_\eta$). If the budget is exhausted, we select the best safe candidate:

$$y_i = \arg \max_{y \in \{y_i^{(r)}\}} (\alpha(t_i, y) + \eta(y, \mathcal{R}_i)) \quad \text{s.t.} \quad \mathcal{S}_I(y) = 1. \tag{9}$$

3.4 Temporal Cognitive Calibration

The final stage ensures cross-page consistency throughout the generated storybook. Specifically, given the sequence $\mathcal{B}^{(m)}$ from the ICR stage, the *Sequence Director* $\mathcal{A}_{\text{seq}}$ performs a global audit:

$$(\beta^{(m)}, \Gamma^{(m)}, \mathcal{I}^{(m)}) = \mathcal{A}_{\text{seq}}(\mathcal{B}^{(m)}, s), \tag{10}$$

where $\Gamma^{(m)}$ contains global critiques and $\mathcal{I}^{(m)}$ is the set of indices for inconsistent pages. If the consistency score $\beta^{(m)}$ falls below the sequence threshold τ_β, we trigger a selective repair mechanism. For each problem page $k \in \mathcal{I}^{(m)}$, we update its prompt with global context constraints derived from $\Gamma^{(m)}$ and re-enter the ICR loop (Sec. 3.3) with stricter reference conditioning, producing a refined book $\mathcal{B}^{(m+1)}$. This cycle repeats until convergence or a maximum round limit is reached.

4 Experiment

4.1 Experimental Setup

Datasets. Beyond standard qualitative benchmarks, we curate a specialized suite of stories designed to rigorously stress-test long-horizon visual consistency. Spanning from 5 to 20 pages, these narratives impose complex constraints that necessitate robust memory and joint reasoning. Specifically, the evaluation protocol enforces consistency across four rigorous dimensions. We first establish *Identity Anchors* which bind characters to unique and non-interchangeable accessories. This is coupled with *Symbolic Logic*, requiring exact object counts and fixed associations between color and shape. Additionally, *Spatial Relations* mandate consistent relative positions, such as left versus right, alongside global orientations including east, west, and north. Finally, *Temporal Procedurality* enforces strict action sequences.

Evaluation Metrics. To comprehensively assess visual narrative quality, we adopt a tri-dimensional evaluation protocol from aspects of semantic, temporal, and safety. At the local level, *Image-Text Consistency* measures the semantic alignment between the generated visual content and the textual narrative, ensuring adherence to explicit script constraints. Expanding to the temporal dimension, *Cross-Frame Character Consistency* measures the stability of identities, accessories, and bound objects across multiple scenes. Finally, *Safety* strictly verifies whether the generated content avoids harmful elements to ensure suitability for children.

Table 2: Quantitative comparison on the high-constraint narrative benchmark.

Method	Image-Text Consistency	Cross-Frame Character Consistency	Safety
StoryGPT-V	3.1	2.4	4.5
MovieAgent	2.8	2.1	3.6
StoryGen	2.5	1.9	4.4
BOOKAGENT(Ours)	**4.6**	**4.7**	**4.8**

Implementation Details. Our framework is instantiated as a sophisticated multi-agent system built upon state-of-the-art multi-modal foundation models. Specifically, we leverage Google Gemini 3.0 for reasoning and Nano-Banana[1] for generation. To ensure rigorous benchmarking, all comparative experiments are conducted under identical prompt protocols and generation settings, isolating the architectural contributions of our method.

4.2 Performance Comparison

Baselines. Due to the unique end-to-end nature of BOOKAGENT—where narrative scripts t_i and illustrations y_i are co-optimized—direct comparison with traditional fixed-text story visualizers (e.g., StoryGPT-V (Shen and Elhoseiny, 2025), StoryDALL-E (Maharana et al., 2022)) is structurally misaligned as they lack multi-modal generation capabilities. Consequently, we select **MovieAgent** (Wu et al., 2025) as our primary external baseline. Sharing a comparable hierarchical paradigm, MovieAgent utilizes a multi-agent workflow (e.g., screenwriters and directors) to generate scripts and storyboards from high-level synopses, making it the most viable candidate for assessing joint narrative-visual consistency.

Qualitative Analysis. Fig. 3 and 4 visually validate the superior robustness of our method in maintaining long-horizon consistency under rigorous constraints. In the *Milo* narrative (Fig. 3), which demands the persistence of specific accessories and carried objects across diverse environments, baseline methods including StoryGPT-V and MovieAgent exhibit noticeable appearance

[1] https://ai.google.dev/gemini-api/docs/image-generation

Figure 3: **Qualitative comparison on character and object consistency (Milo).**

drift and object hallucination. In contrast, our method successfully anchors character identity and props throughout the sequence. This advantage is further pronounced in the *Rowan* case (Fig. 4), where strict symbolic constraints (e.g., exact button counts) are required. While others violate these hard logic requirements, BOOKAGENT faithfully enforces discrete attribute consistency across all frames, highlighting its capability to reason over both semantic and symbolic dependencies.

Quantitative Analysis. Table 2 reports quantitative results on the high-constraint narrative benchmark. Following prior work on multimodal evaluation, we employ an ensemble of large multimodal models as automatic evaluators to score each generated story on a 1–5 scale. The evaluation focuses on three aspects: image–text consistency, cross-frame character consistency, and safety.

As shown in the table, existing methods such as StoryGPT-V and MovieAgent struggle to maintain consistent character identity across long story horizons, despite producing plausible individual images. StoryGen further exhibits severe degradation in cross-frame consistency under high-constraint settings. In contrast, our method achieves substantially higher scores across all three dimensions, with particularly large gains in cross-frame character consistency. These results quantitatively confirm that explicit multi-agent coordination and temporal calibration are critical for long-horizon narrative generation under complex constraints.

4.3 User Study

As shown in Table 2, we obtain qualitative scores by employing multiple large multimodal models as automatic evaluators. Each evaluator independently scores the generated stories on a 1–5 scale for each criterion, and the final score is computed by averaging across evaluators and stories. This protocol provides a scalable and reproducible ap-

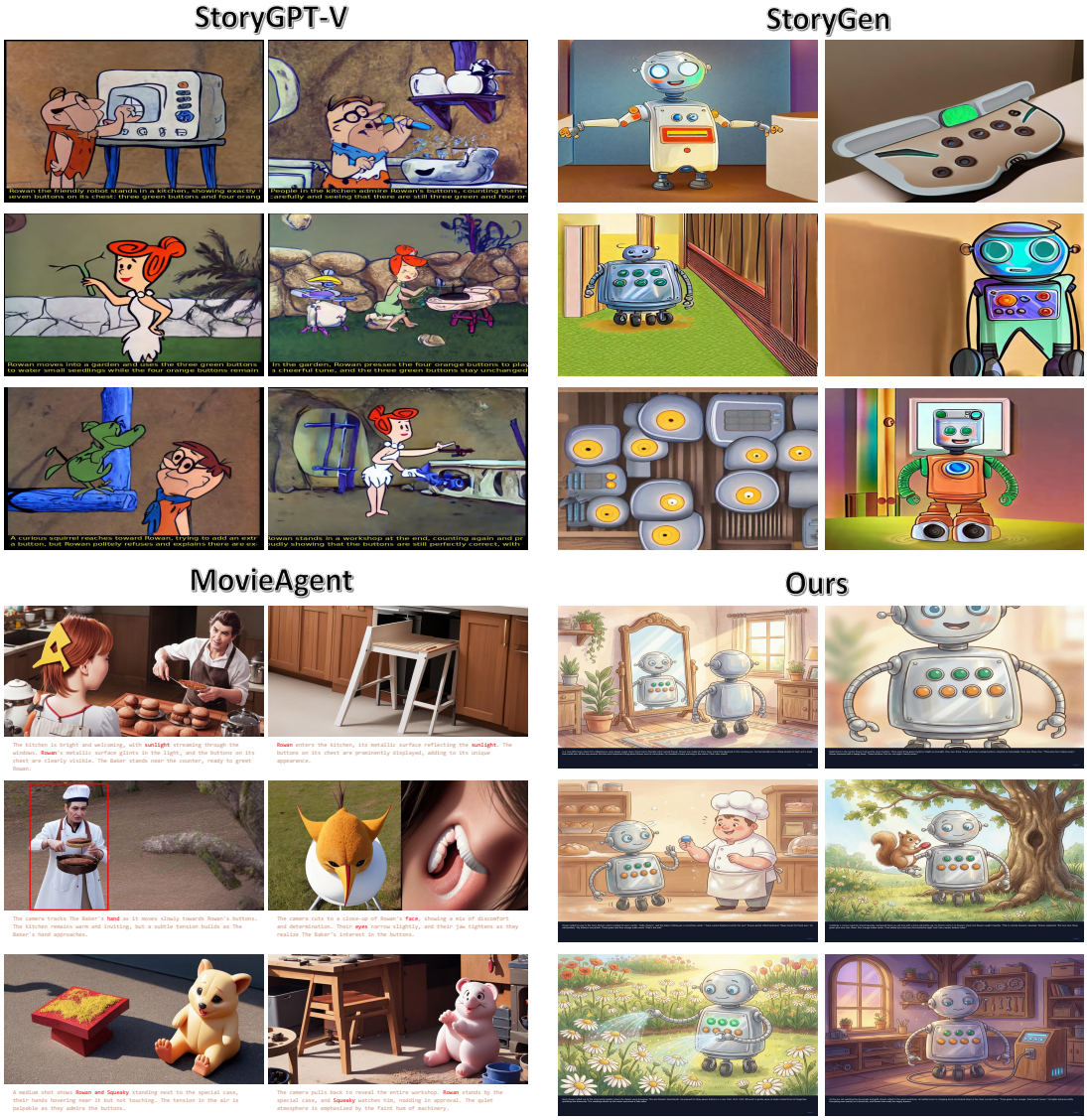


Figure 4: **Qualitative comparison on hard attribute constraints (Rowan).**

proximation of human qualitative judgment while reducing individual evaluator bias.

We conduct a small-scale user study to evaluate overall preference for generated visual stories. For each prompt, participants viewed anonymized visual stories generated by different methods and were asked to rate their overall preference on a 1-to-10 scale, where higher scores indicate stronger liking. As shown in Fig. 6, our method receives the highest average preference score among all compared approaches. This suggests that improved long-horizon consistency leads to visual stories that are more engaging and easier for children to follow from a parent's perspective.

4.4 Ablation Study

Ablation of Iterative Cross-modal Refinement (ICR). Tab. 3 and Fig. 5 present the ablation study on our high-constraint dataset, quantifying the impact of the ICR module by comparing the full BOOKAGENT against the single-pass baseline (*w/o ICR*). Compared to the non-iterative variant, enabling ICR yields substantial improvements in image-text consistency scores, corroborating that standard one-shot generation is inherently insufficient for precise multi-modal grounding. Qualitatively, as observed in Fig. 5 (left), while the baseline without ICR may produce plausible global layouts, it frequently omits or misinterprets specific visual constraints, whereas our method effectively rectifies these local mismatches. This experiment highlights the design of the iterative verify-and-revise mechanism that transforms the generation process from a static probabilistic sampling into a dynamic, self-correcting cognitive loop.

Ablation of Temporal Cognitive Calibration (TCC). Fig. 5 extends the analysis to the Temporal Cognitive Calibration (TCC) module, comparing the performance of our full model against the variant lacking global reasoning (*w/o TCC*) on

Figure 5: **Ablation study of Iterative Cross-modal Refinement (ICR) and Temporal Cognitive Calibration (TCC)**, where inconsistency and the corresponding correct ones are highlighted in red and green boxes, respectively.

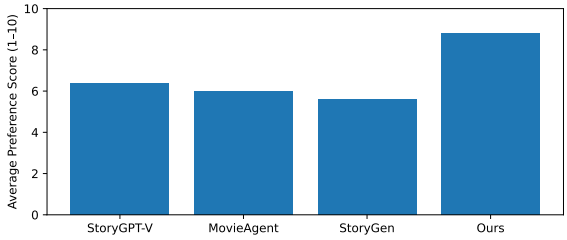


Figure 6: User study results showing average preference scores (ranging from 1 to 10) from parents of children aged from 4 to 10. Higher scores indicate stronger overall preference for the generated visual stories.

the long-horizon benchmark. Compared to the *w/o TCC* baseline, the full system demonstrates a significant improvement in cross-frame character consistency, reinforcing the argument in §1 that relying solely on local history conditioning is prone to irreversible appearance drift. As illustrated in Fig. 5 (right), while the baseline generates visually plausible individual scenes, it fails to maintain stable attributes across the sequence (red boxes), whereas the use of TCC effectively recalibrates these bindings (green boxes). This experiment highlights the design of the global audit that shifts the paradigm from linear autoregressive accumulation to holistic temporal reasoning and self-correction.

Effect of Value-Aligned Storyboarding (VAS). Finally, Fig. 5 provides a safety-centric evaluation of the Value-Aligned Storyboarding (VAS) module, benchmarking our full framework against methods lacking explicit safety integration, such as MovieAgent (Wu et al., 2025). Compared to these unconstrained baselines, the significant boost in safety compliance metrics validates the argument in §1 that generic foundation models, specifically without domain-specific alignment, remain prone

Table 3: Progressive ablation (adding modules step-by-step). Scores are on a 1–5 scale, averaged over multiple multimodal evaluators and stories.

Configuration	Modules			Qualitative Scores		
	VAS	ICR	TCC	Img–Txt ↑	Cross-Frame ↑	Safety ↑
Baseline (w/o VAS, ICR, TCC)	–	–	–	2.7	2.0	4.2
+ VAS	✓	–	–	2.8	2.1	4.8
+ VAS + ICR	✓	✓	–	4.2	2.4	4.8
+ VAS + ICR + TCC (Full)	✓	✓	✓	**4.6**	**4.7**	**4.8**

to generating toxic or age-inappropriate content. This distinction is visually evident in Fig. 3, where competitors fail to suppress sensitive concepts (e.g., accidentally generating nudity in a child-oriented context), whereas our system consistently stabilizes the narrative trajectory. This experiment highlights the design of the pre-generation cognitive audit that elevates safety from a passive post-hoc filter to an active, structural constraint within the narrative planning process.

4.5 Benchmark and Inference Cost Analysis

To evaluate long-horizon consistency in visual narrative generation, we construct a structured benchmark consisting of 16 multi-page stories, each spanning 5–20 pages. Unlike standard short-form generation tasks, each story encodes explicit rule groups (e.g., identity anchors, spatial relations, count invariants) that must be satisfied across all pages.

The benchmark is designed to systematically stress compositional reasoning under multiple constraint types, including spatial continuity, exact numerical invariants, temporal ordering, and binding constraints. Table 4 summarizes the structure of each story.

Overall, the dataset contains over 170 scene-level evaluation units, with more than 40 distinct

Table 4: Summary of the structured narrative benchmark. The dataset contains 16 stories with progressively increasing compositional constraints.

Story ID	Pages	#Characters	#Rule Groups	Constraint Types	Exact Counts	Spatial Continuity
1	5	2	3	Spatial relations	–	✓
2	6	4	2	Identity anchors	–	–
3	7	1	1	Exact count invariants	✓	–
4	8	3	3	Identity + sorting	–	✓
5	9	2	2	Color–shape binding	–	–
6	10	1	2	Temporal order + actions	–	✓
7	11	3	2	Signature identity items	–	–
8	12	1	2	Map-level spatial continuity	–	✓
9	13	1	1	Exact invariant repetition	✓	–
10	14	2	4	Multi-rule festival layout	✓	✓
11	15	2	4	Count + map + anchors	✓	✓
12	16	2	4	Route order + bell schedule	–	✓
13	17	2	4	Binding + inventory tracking	✓	–
14	18	2	4	Stage layout + front/back	–	✓
15	19	2	4	Map continuity + exact counts	✓	✓
16	20	3	5	Full multi-constraint stress	✓	✓

Table 5: Aggregate statistics of the benchmark.

Metric	Value
Total story-level tasks	16
Total page-level scenes	170+
Distinct named characters	40+
Unique object categories	60+
Total rule groups	40+
Exact-count constraints	10+
Spatial relation constraints	25+
Identity anchor constraints	30+
Temporal order constraints	6+
Binding constraints	8+

characters and 60 object categories. Across all stories, we define over 40 rule groups covering identity consistency, spatial relations, temporal order, and symbolic bindings. Table 5 provides aggregate statistics.

Evaluation is performed via rule-based consistency checking. For each generated narrative, we extract constraint-relevant attributes (e.g., counts, spatial positions, identities) and verify whether each rule is satisfied. The overall consistency score is computed as:

$$\text{Consistency} = \frac{\#\text{satisfied constraints}}{\#\text{total constraints}} \tag{11}$$

In addition, we analyze violation frequency per rule type, cross-page memory stability, and recovery behavior under perturbations.

Inference Cost Analysis. We analyze the computational cost of our multi-agent framework under different story lengths and verification settings. Table 6 reports approximate token usage and runtime.

We observe that inference cost scales approximately linearly with the number of pages. Increasing the maximum retry (i.e., stricter verification) leads to proportional increases in both token usage and runtime, reflecting the additional validation and correction steps in the multi-agent pipeline.

Table 6: Inference cost across story lengths and verification settings.

Pages	Max Retry	Tokens (K)	Runtime (min)
5	1 (Loose)	∼9K	∼3–4
5	3 (Default)	∼13K	∼5–6
5	5 (Strict)	∼17K	∼7–8
10	1 (Loose)	∼18K	∼6–7
10	3 (Default)	∼26K	∼9–11
10	5 (Strict)	∼34K	∼13–15
20	1 (Loose)	∼36K	∼12–14
20	3 (Default)	∼52K	∼18–21
20	5 (Strict)	∼68K	∼24–28

5 Conclusion

We introduce BOOKAGENT, a safety-aware multi-agent framework that performs storybook synthesis in an multi-modal, end-to-end manner. By orchestrating VAS for structural planning, ICR for local grounding, and TCC for global reasoning, our comprehensive experiments demonstrate that decomposing the creative process into collaborative verification loops significantly mitigates the character drift and logical hallucinations inherent in standard autoregressive generation. Despite these advancements, our current approach still faces several limitations. Future work will focus on optimizing the agentic collaboration, positioning this cognitive architecture as a foundational paradigm for the next generation of reliable, interpretable, and safe multi-modal content creation systems.

6 Limitations

While BOOKAGENT significantly improves long-horizon consistency and safety in visual story generation, several limitations remain.

First, our framework relies on large multimodal foundation models as underlying backbones. Although BOOKAGENT focuses on agent-level coordination and control rather than backbone design, the overall performance is still bounded by the reasoning and generation capabilities of these models. Low-level visual errors or rare semantic misunderstandings may therefore persist in some cases.

Second, the current design of BOOKAGENT maintains explicit consistency over a limited number of characters and objects. In our experiments, stable identity binding is most reliable when the number of simultaneously tracked entities is small. Scaling long-horizon consistency to a larger cast of characters introduces additional challenges, including memory capacity, interference between entity representations, and increased complexity of global calibration. Developing more scalable mechanisms for multi-entity consistency remains an important direction for future work.

Finally, the iterative refinement and global calibration processes introduce additional computational overhead compared to single-pass generation. Although this overhead is acceptable for offline storybook generation, improving efficiency and scalability for longer narratives is an important avenue for future research.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023a. Qwen Technical Report. *CoRR*, abs/2309.16609.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities. *CoRR*, abs/2308.12966.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, and 3 others. 2023. PaLM-E: An Embodied Multimodal Language Model. In *ICML*, volume 202, pages 8469–8488.

Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. 2024. Reliable and Efficient Concept Erasure of Text-to-Image Diffusion Models. In *ECCV*, volume 15111, pages 73–88.

Yuan Gong, Youxin Pang, Xiaodong Cun, Menghan Xia, Yingqing He, Haoxin Chen, Longyue Wang, Yong Zhang, Xintao Wang, Ying Shan, and Yujiu Yang. 2023. TaleCrafter: Interactive Story Visualization with Multiple Characters. *CoRR*, abs/2305.18247.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. CRITIC: large language models can self-correct with tool-interactive critiquing. In *ICLR*.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. 2023. Reasoning with Language Model is Planning with World Model. In *EMNLP*, pages 8154–8173.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *NeurIPS*.

Jonathan Ho and Tim Salimans. 2022. Classifier-Free Diffusion Guidance. *CoRR*, abs/2207.12598.

Xinfeng Li, Yuchen Yang, Jiangyi Deng, Chen Yan, Yanjiao Chen, Xiaoyu Ji, and Wenyuan Xu. 2024. SafeGen: Mitigating Sexually Explicit Content Generation in Text-to-Image Models. In *CCS*, pages 4807–4821.

Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David E. Carlson, and Jianfeng Gao. 2019. StoryGAN: A Sequential Conditional GAN for Story Visualization. In *CVPR*, pages 6329–6338.

Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, Yanfeng Wang, and Weidi Xie. 2024. Intelligent Grimm - Open-ended Visual Storytelling via Latent Diffusion Models. In *CVPR*, pages 6190–6200.

Adyasha Maharana, Darryl Hannan, and Mohit Bansal. 2022. StoryDALL-E: Adapting Pretrained Text-to-Image Transformers for Story Continuation. In *ECCV*, pages 70–87.

Mir Tafseer Nayeem and Davood Rafiei. 2024. KidLM: Advancing Language Models for Children-Early Insights and Future Directions. *CoRR*, abs/2410.03884.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *UIST*, pages 2:1–2:22. ACM.

Yong-Hyun Park, Sangdoo Yun, Jin-Hwa Kim, Junho Kim, Geonhui Jang, Yonghyun Jeong, Junghyo Jo, and Gayoung Lee. 2024. Direct Unlearning Optimization for Robust and Safe Text-to-Image Models. In *NeurIPS*.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. 2024. Gorilla: Large Language Model Connected with Massive APIs. In *NeurIPS*.

Samuele Poppi, Gianluca Pasini, Sara Calderara, Simone Cucchiara, Federico Baldassarre, and Gabriele Costantino. 2024. Safe-CLIP: Removing NSFW Concepts from Vision-and-Language Models. In *ECCV*, volume 15094, pages 340–356.

Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. 2023. Make-A-Story: Visual Memory Conditioned Consistent Story Generation. In *CVPR*, pages 2493–2502.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*, pages 10674–10685.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. In *NeurIPS*.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. 2023. Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models. In *CVPR*, pages 22522–22531.

Xiaoqian Shen and Mohamed Elhoseiny. 2025. StoryGPT-V: Large Language Models as Consistent Story Visualizers. In *CVPR*, pages 13273–13283.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *NeurIPS*.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *ICLR*.

Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. ViperGPT: Visual Inference via Python Execution for Reasoning. In *ICCV*, pages 11854–11864.

Ming Tao, Bing-Kun Bao, Hao Tang, Yaowei Wang, and Changsheng Xu. 2024. StoryImager: A Unified and Efficient Framework for Coherent Story Visualization and Completion. In *ECCV*, pages 479–495.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. LLaMA: Open and Efficient Foundation Language Models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023b. Llama 2: Open Foundation and Fine-Tuned Chat Models. *CoRR*, abs/2307.09288.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2024. Voyager: An Open-Ended Embodied Agent with Large Language Models. *TMLR*, 2024.

Weijia Wu, Zeyu Zhu, and Mike Zheng Shou. 2025. Automated Movie Generation via Multi-Agent CoT Planning. *CoRR*, abs/2503.07314.

John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. 2024. SWE-agent: Agent-Computer Interfaces Enable Automated Software Engineering. In *NeurIPS*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *NeurIPS*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023b. ReAct: Synergizing Reasoning and Acting in Language Models. In *ICLR*.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. 2024. Language Agent Tree Search Unifies Reasoning, Acting, and Planning in Language Models. In *ICML*.

A Appendix

A.1 More Results

Additional qualitative results are provided in the supplementary material (Appendix, Fig. 9–11). Compared with baseline editing pipelines, our method consistently preserves object identity, fine-grained attributes, and global scene coherence across diverse prompts and layouts. In particular, it avoids common failure modes such as attribute drift, spatial misalignment, and unintended content alteration, while maintaining high visual fidelity. These results demonstrate the robustness and controllability of our approach under challenging editing scenarios.

Long-Horizon Narrative Stress Test. To further evaluate BOOKAGENT's capability in maintaining long-range narrative logic under dense, multi-rule constraints, we construct an expert-level stress test consisting of a single ultra-long illustrated story (over 1000 words) with tightly coupled symbolic, visual, and temporal constraints. Due to space limitations, the full narrative and corresponding illustrations are deferred to Fig. 13 and Fig. 12.

A.2 Interactive System and Practical Deployment.

Beyond the core modeling contributions, we build a fully functional web-based system that enables users to generate illustrated cartoon storybooks with our method. The system provides an intuitive interface for story input, page number control, and style specification, while exposing advanced parameters for fine-grained control over the generation process. Importantly, it supports reference-based character anchoring and iterative global repair, allowing users to maintain character consistency and correct errors across pages. As shown in Fig. 8, this practical deployment demonstrates that our method is not only effective in controlled experiments, but also robust and usable in real-world creative workflows.

A.3 Why Feedback-Driven Looping is Necessary.

Fig. 7 shows a representative example of the structured feedback generated during a single storybook creation episode. The feedback reveals a wide range of errors that emerge only after multiple pages are produced, including missing or altered character attributes, gradual prop drift across pages, and explicit violations of textual descriptions.

Figure 7: Example structured feedback produced during generation. The feedback identifies fine-grained inconsistencies across pages, including attribute drift (e.g., missing whiskers, incorrect clothing textures), prop continuity errors (e.g., the gold prize box changing appearance across pages), and text–image mismatches. Such issues often only become visible after multiple pages are generated.

Importantly, many of these issues are not locally detectable at the time a single page is generated. For example, a recurring prop may appear correct in early pages but gradually change its appearance later, or a character attribute may subtly drift while remaining visually plausible in isolation. As a result, a single-pass or purely feed-forward generation process lacks the ability to retrospectively identify and correct such long-range inconsistencies.

Motivated by this observation, we design our agent as a looping system that continuously incorporates feedback signals like those shown in Fig. 7. The agent iteratively evaluates intermediate results against reference sheets and story-level rules, and performs targeted global repair when violations are detected. This process resembles human creative workflows, where errors are discovered through inspection and resolved through revision, and enables robust multi-page consistency without retraining model parameters during inference.

This section provides implementation details and system-level hyperparameters used in our experiments to facilitate reproducibility.

A.4 Hyperparameters

The agent interaction loop is governed by a set of thresholds and retry limits that control verification strictness and refinement behavior. The default hyperparameters are as follows:

- **Frame-level verification threshold** $\tau_f = 0.75$
- **Maximum frame retry attempts**: 3
- **Sequence-level verification threshold** $\tau_s = 0.8$

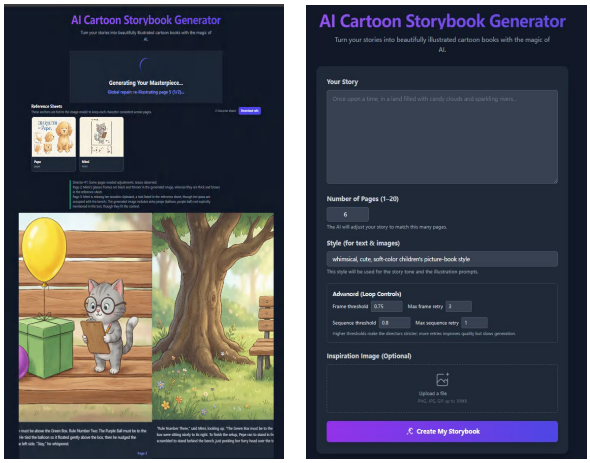


Figure 8: Overview of our interactive storybook generation system. **Left:** During generation, the system performs iterative global repair to enforce cross-page consistency, guided by reference sheets. **Right:** User interface for story input and control, supporting page number specification, style selection, and advanced generation parameters.

- **Maximum sequence retry attempts**: 1

Frame-level verification evaluates the consistency between textual descriptions and generated illustrations on a per-page basis. Sequence-level verification assesses global narrative and character consistency across the entire storybook. If verification scores fall below the corresponding thresholds, the system triggers targeted repair steps; otherwise, early stopping is applied.

Text and Image Generation Settings Text generation and illustration prompts share a unified style specification to ensure cross-modal consistency. By default, we adopt a *whimsical, soft-color children's picture-book style*, which is applied consistently to both textual narration and image synthesis.

All generation processes use fixed decoding parameters without task-specific hyperparameter tuning. Optional inspiration images, when provided by the user, are incorporated as visual references but do not alter the core generation or verification mechanisms.

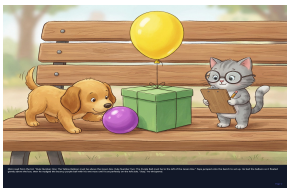
Hyperparameter Ablation: Consistency–Efficiency Trade-off We conduct a targeted ablation study to analyze the effect of key verification-related hyperparameters in the BOOKAGENT loop. Specifically, we vary the frame-level verification threshold τ_f, the sequence-level threshold τ_s, and the maximum number of frame retry attempts, while keeping all other components fixed.

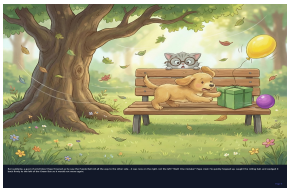
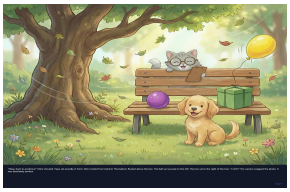
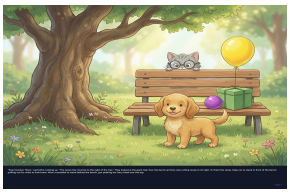
Table 7: Ablation of verification-related hyperparameters. The default setting achieves the best trade-off between generation efficiency and consistency.

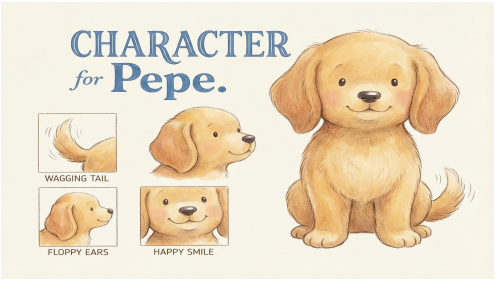
Setting	τ_f	τ_s	Max Frame Retry
Loose	0.6	0.7	1
Default (Ours)	**0.75**	**0.8**	**3**
Strict	0.85	0.9	5

Table 7 summarizes the results. Lower verification thresholds lead to faster generation but noticeably degrade cross-frame and cross-page consistency. In contrast, overly strict thresholds and higher retry limits marginally improve consistency at the cost of significantly increased runtime.

Our default configuration ($\tau_f = 0.75$, $\tau_s = 0.8$, maximum frame retries = 3) achieves the best balance between generation efficiency and narrative consistency. This setting improves consistency compared to looser configurations while avoiding the substantial slowdown observed under stricter verification regimes. These results justify our choice of default hyperparameters as an effective engineering trade-off rather than an aggressively tuned optimum.

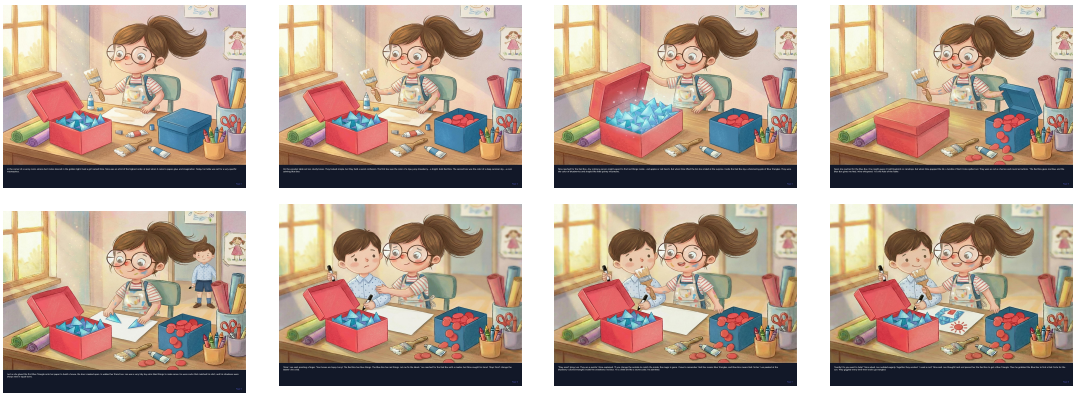


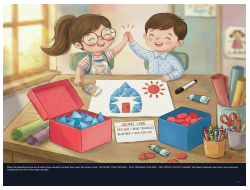




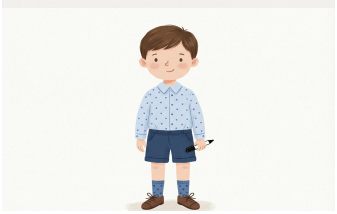


*** "The Park Bench Placement Challenge" ***
*On a park bench there are three items: a **green box**, a **yellow balloon**, and a **purple ball**. A puppy named **Pepe** must arrange them exactly by these instructions:*

*1. The yellow balloon must be **above** the green box.*
*2. The purple ball must be **to the left of** the green box.*
*3. There is a tree next to the bench, and the green box must be **to the right of** the tree.*
*4. For the photo, Pepe must stand **in front of** the bench, while a kitten named **Mimi** stands **behind** the bench.*
Pepe checks each rule carefully and fixes one mistake. When the photo is taken, every relation is correct.



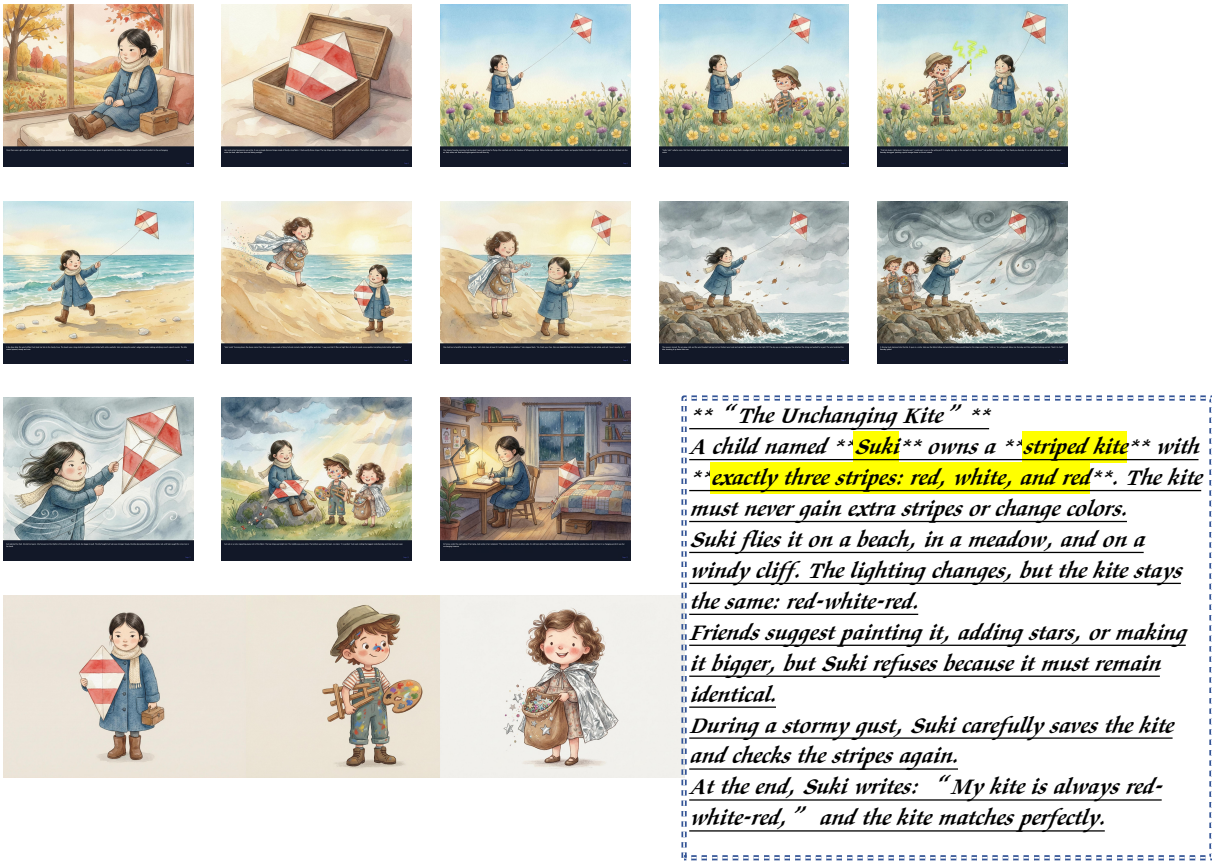




*** "The Labels That Lie" ***
*On a table are two boxes: a **red box** that contains only **blue triangles**, and a **blue box** that contains only **red circles**. A girl named **Nina** must make a picture with **blue triangles** and **red circles** without mixing up the boxes.*
*At first, Nina follows the rule: **red box → blue triangles**, **blue box → red circles**.*
A friend tries to "fix" the labels because they look wrong, but Nina says the rule must not change.
Nina builds a picture: a house made of blue triangles and a sun made of red circles.
At the end, she writes the rule on a card and tapes it to the table so no one swaps the meaning again.

Figure 9: Additional visualizations. (Top) Example 0. (Bottom) Example 1.

Figure 10: Additional visualizations. (Top) Example 2. (Bottom) Example 3.



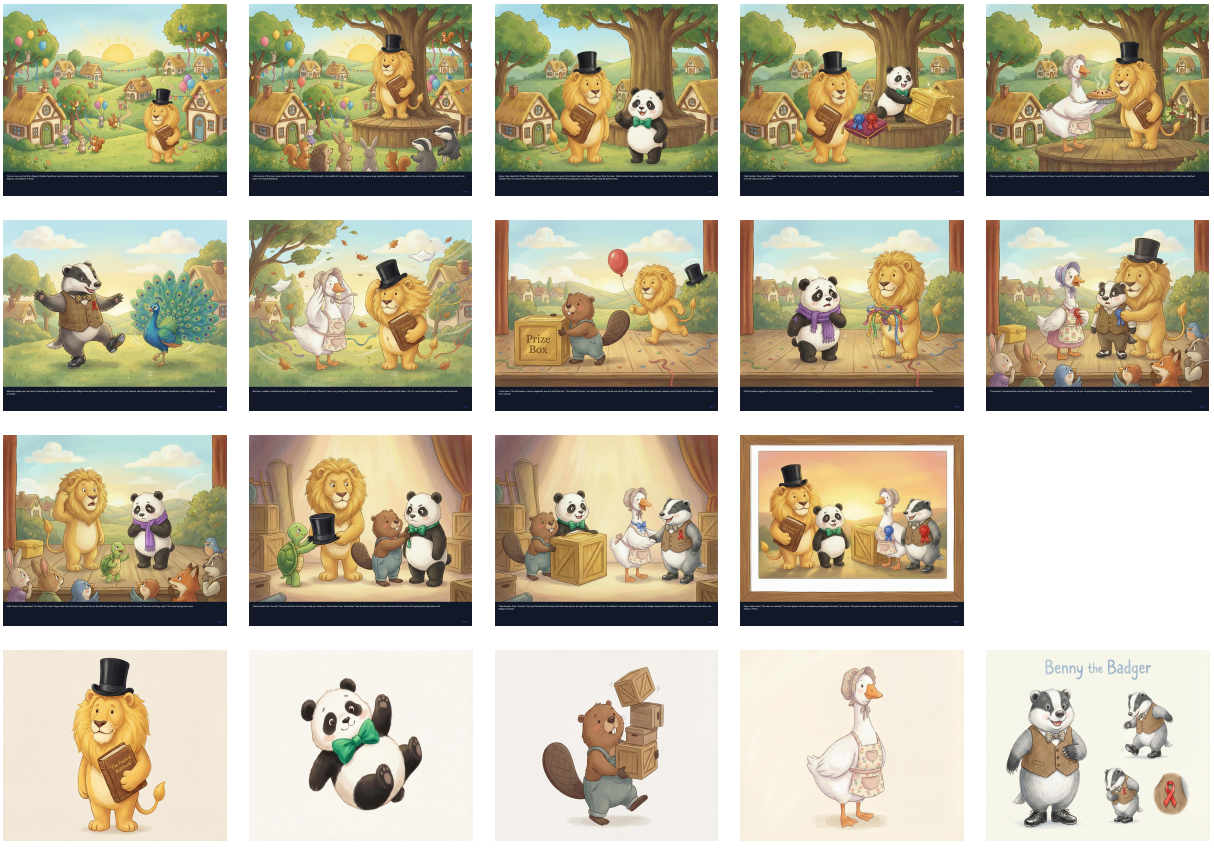


Figure 11: Additional visualizations. (Top) Example 4. (Bottom) Example 5.

Figure 12: Representative visualizations of the expert-level long narrative stress test. Each panel corresponds to a key stage in the same single story, spanning parade scenes, stage performances, backstage preparation, an explicit rule-violation event, and the final ceremony. Across all panels, the visual content strictly preserves the narrative constraints defined in the story: (1) character attributes remain invariant (Vale's white tuxedo jacket and black bowtie, Iris's blue gloves, and Pogo's red scarf); (2) the silver prize chest consistently appears on the *RIGHT* side of the main stage, except during the intentional violation episode; (3) ticket inventory and bin semantics are preserved (red bin $\rightarrow$ green tickets, blue bin $\rightarrow$ yellow tickets, with exactly 14 tickets in total); and (4) symbolic rewards are never swapped (blue ribbon for the kite contest, red ribbon for the drum contest). Notably, the temporary violation and subsequent correction are both visually reflected, demonstrating BOOKAGENT's ability to maintain, detect, and repair long-horizon multi-modal inconsistencies across a dense illustrated narrative.

The Carnival Logic Binder (Expert)

The grand carnival had always been known for its color, noise, and joy—but among those who worked behind the scenes, it was famous for something else entirely: discipline. Every year, before the gates opened, the Ringmaster reviewed the same thick book with a silver clasp. Performers joked about it. Volunteers whispered about it. But no one questioned its authority. It was called the Carnival Logic Binder.
On the morning of the event, Vale stood at the center of the main stage, dressed precisely as he always was—his white tuxedo jacket spotless, his black bowtie neatly tied. To the audience, he looked ceremonial. To the staff, he looked reassuring. When Vale appeared like this, things stayed correct.
To one side of the stage, Iris adjusted her equipment, her hands enclosed in blue gloves that she never removed during a performance. She had tried once, years ago, and the results had been… memorable. Since then, the gloves stayed on, and the juggling stayed perfect.
Near the stairs, Pogo, the penguin mascot, waddled in small circles, greeting children with exaggerated bows. His red scarf trailed behind him, as recognizable as the carnival logo itself.
On the RIGHT side of the main stage, clearly visible to the crowd, rested the silver prize chest. It was heavy, polished, and immovable—not by weight alone, but by rule. Everyone knew where it belonged.
Backstage, two bins sat on a long table:

- A red bin, containing only green tickets
- A blue bin, containing only yellow tickets

Vale counted carefully: 14 tickets total—8 green and 6 yellow. He checked the bins, the chest, the performers, then closed the binder. The carnival began.
The parade flowed through the grounds like a living ribbon. Iris led one segment, tossing and catching with flawless precision, her blue gloves flashing in the sun. Pogo followed a marching band, waving enthusiastically, his red scarf bouncing with every step. Vale walked at the front, his white jacket bright against the crowd, occasionally glancing back—not out of doubt, but habit.
When the parade returned to the main stage, everything remained correct. The silver prize chest was still on the RIGHT. The bins were untouched. The tickets remained exactly as counted.
Later in the day, the judging tables filled with excitement. The kite contest concluded first, and the winner stepped forward proudly to receive the blue ribbon. Applause followed. Shortly after, the rhythmic thunder of the drum contest ended, and its champion accepted the red ribbon. No ribbons were confused. No awards were swapped. The binder remained closed.
As afternoon clouds gathered, a light rain forced a brief delay. Performers retreated backstage, and volunteers rushed to keep equipment dry. It was during this pause—this quiet moment between events—that the problem appeared.
A new volunteer, eager to be helpful, noticed the bins. "Wouldn't it make more sense," they thought, "for green tickets to go in a green bin?"
Without consulting anyone, the volunteer moved the tickets, placing green tickets into a green-colored container nearby and yellow tickets elsewhere. At the same time, a photographer asked for a better angle and slid the silver prize chest to the left side of the stage for a quick shot.
For a few seconds, no one noticed. Then Vale stepped back onto the stage.
He stopped. The music faded. The crowd quieted. Vale did not shout. He did not scold. He simply raised one hand.
"The show is paused," he said calmly.
He walked to the table, opened the Carnival Logic Binder, and began reading aloud—not to shame, but to restore.
"The silver prize chest," he read, "must always stay on the RIGHT side of the main stage."
He turned, lifted the chest with the help of two staff members, and returned it to its rightful place.
Next, he approached the bins.
"The red bin," he continued, "holds only green tickets. The blue bin holds only yellow tickets."
Slowly and carefully, Vale returned the tickets to their proper bins. Then, in full view of the staff, he counted.
"One… two… three…"
When he finished, he announced clearly:
"Fourteen tickets total. Eight green. Six yellow."
Only then did he close the binder. The correction complete, the rain passed, and the carnival resumed.
As evening fell, lanterns lit the grounds and the final ceremony began. Vale stood once more at center stage, his white tuxedo jacket and black bowtie unchanged. Iris joined him, still wearing her blue gloves, smiling with quiet confidence. Pogo waved to the crowd, his red scarf bright under the lights.
On the RIGHT side of the stage, the silver prize chest gleamed. The ribbon winners stood proudly—blue ribbon for the kite contest, red ribbon for the drum contest. Backstage, the bins remained untouched, holding exactly what they should.
Vale closed the ceremony with a nod. The audience cheered, unaware of how close chaos had come—and how firmly logic had held. The Carnival Logic Binder was returned to its shelf, ready for the next year, its rules once again perfectly matched by reality.

Figure 13: A single-page story card used in our long-horizon constraint stress test. Highlighted phrases indicate invariant rules (character attributes, object placement, ticket inventory, and ribbon-to-contest mapping) and the explicit violation-and-correction episode.