

EVOSPARK: Endogenous Interactive Agent Societies for Unified Long-Horizon Narrative Evolution

Shiyu He^{1*}, Minchi Kuang^{2*†}, Mengxian Wang^{1*}, Bin Hu¹, Tingxiang Gu¹,

¹School of Computer Science and Technology, Xinjiang University, Urumqi 830046, China

²Department of Precision Instrument, Tsinghua University, Beijing 100084, China

{heshiyu, wangmengxian, hubin, gutingxiang}@stu.xju.edu.cn

kuangmc@mail.tsinghua.edu.cn

Abstract

Realizing endogenous narrative evolution in LLM-based multi-agent systems is hindered by the inherent stochasticity of generative emergence. In particular, long-horizon simulations suffer from *social memory stacking*, where conflicting relational states accumulate without resolution, and *narrative-spatial dissonance*, where spatial logic detaches from the evolving plot. To bridge this gap, we propose EVOSPARK, a framework specifically designed to sustain logically coherent long-horizon narratives within endogenous interactive agent societies. To ensure consistency, the stratified narrative memory employs a role socio-evolutionary base as living cognition, dynamically metabolizing experiences to resolve historical conflicts. Complementarily, a Generative Mise-en-Scène mechanism enforces role-location-plot alignment, synchronizing character presence with the narrative flow. Underpinning these is the unified narrative operation engine, which integrates an Emergent Character Grounding Protocol to transform stochastic *sparkling* into persistent characters. This engine establishes a substrate that expands a minimal premise into an open-ended, evolving story world. Experiments demonstrate that EVOSPARK significantly outperforms baselines across diverse paradigms, enabling the sustained generation of expressive and coherent narrative experiences.

1 Introduction

The integration of Large Language Models (LLMs) into Multi-Agent Systems (MAS) has fundamentally reshaped the landscape of generative storytelling, enabling agents to simulate complex social interactions with unprecedented fluency (Park et al., 2023; Piao et al., 2025). However, while current systems excel at generating short-term vignettes, achieving long-horizon story evolution—where a

simulation evolves autonomously from a minimal seed into an unbounded, self-sustaining, and logically coherent narrative ecosystem—remains an elusive goal (Xia et al., 2025).

As narratives expand in complexity, two critical systemic deficits emerge in current architectures, preventing the sustainability of long-term logic. First, systems suffer from social memory stacking, where the conventional append-only memory architecture leads to the accumulation of conflicting relational states (e.g., distinct memories of being both a friend and a foe), causing behavioral incoherence (Platnick et al., 2025; Zhong et al., 2023). Second, text-based agents face narrative-spatial dissonance. Lacking a mechanism to synchronize narrative progression with spatial states, agents often generate ungrounded interactions that violate the essential role-location-plot logic—such as characters appearing in disjointed locations during plot-critical transitions—thereby severing the logical link between the story and its setting (Ran et al., 2025; Chen et al., 2025).

Beyond these functional deficits, the field is structurally constrained by a paradigm schism. Traditional interactive narratives rely on rigid script adherence, ensuring logic but sacrificing autonomy (Sun et al., 2025). Conversely, recent LLM-based simulations often prioritize open-ended emergence (Park et al., 2023; Yang et al., 2025), leading to uncontrollable chaotic emergence. Although recent works imply control mechanisms (Han et al., 2024; Wang et al., 2024), existing fragmented architectures fail to support the full spectrum of narrative control necessary for diverse simulation needs—ranging from strict hierarchical planning to open-ended evolution (Sun et al., 2025).

To address these challenges, we propose EVOSPARK, a unified framework that integrates narrative control, cognitive evolution, and spatial grounding to foster interactive agent societies driven by endogenous character emergence. Our

* Equal contribution.

† Corresponding author.

core contributions are:

- **Unified Narrative Operation Engine (NOE):** We operationalize *sparkling*—stochastic LLM hallucinations—not as errors, but as drivers for creativity via the **Emergent Character Grounding Protocol (ECGP)**. Through **Ontological Promotion**, the system validates and transforms fleeting narrative-induced hallucinations into persistent, legitimate main characters, effectively turning stochastic noise into structural assets for infinite world expansion.
- **Role Socio-Evolutionary Base (RSB):** To resolve *social memory stacking*, we introduce the RSB as a mutable, *living* cognition base. Unlike static or append-only RAG approaches, the RSB employs event-driven reflection to continuously metabolize experiences—updating personality, social graphs, and goals via in-place modifications. This ensures the agent’s internal state evolves in real-time, maintaining consistency with the shifting socio-dynamic landscape.
- **Generative Mise-en-Scène (GMS):** We mitigate narrative-spatial dissonance through GMS mechanism, which acts as a virtual stage manager to enforce strict Role-Location-Plot (RLP) alignment. It dynamically synchronizes character presence and transitions with the evolving narrative flow, ensuring spatial contexts remain logically congruent with story progression.

2 Related Work

Endogenous Multi-Agent Systems Research on MAS has transitioned from static orchestration to dynamic, endogenous evolution. Early frameworks like MetaGPT and Camel (Hong et al., 2024; Li et al., 2023) relied on fixed Standard Operating Procedures, limiting adaptability. To address this, recent architectures enable self-improvement: CoMAS and AFlow (Xue et al., 2025; Zhang et al., 2025d) optimize policies via interaction rewards, while the Darwin Gödel Machine (Zhang et al., 2025c) allows agents to recursively modify their own code. In the realm of simulation, AgentSociety, OASIS and Generative Agents (Piao et al., 2025; Yang et al., 2025; Park et al., 2023) scale interactions to observe emergent norms, while BookWorld (Ran et al., 2025) constructs societies directly from fictional texts. However, current endogenous systems exhibit a teleological mismatch

for open-ended storytelling. First, evolutionary frameworks optimize for metric convergence (e.g., pass rates) (Xue et al., 2025; Zhang et al., 2025d), treating the stochasticity vital for narrative expansion as noise. Second, social simulators like S3 and WarAgent (Gao et al., 2023; Hua et al., 2023) suffer from social memory stacking—accumulating interaction logs without a metabolic mechanism to transform transient experiences into persistent structural updates, inevitably degrading long-horizon coherence.

Socio-Evolutionary Dynamics and Memory To support long-term narratives, research primarily targets static consistency to mitigate identity drift. Frameworks like ID-RAG (Platnick et al., 2025) employ identity knowledge graphs to ground agent personas, while Open-Theatre (Xu et al., 2025) and MemoryBank (Zhong et al., 2023) introduce hierarchical stores to ensure retrieval accuracy. Similarly, S3 and AgentSociety (Gao et al., 2023; Piao et al., 2025) utilize memory modules to maintain coherent behavioral patterns across social simulations. However, these architectures typically treat memory as an accumulative log. This rigidity leads to social memory stacking, where obsolete relational states persist and conflict with new developments. While Generative Agents (Park et al., 2023) uses reflection to synthesize observations, it focuses on preserving character states. Recent works attempt to address this: G-Memory (Zhang et al., 2025a) evolves hierarchies based on interaction trajectories, MemEvolve (Zhang et al., 2025b) meta-optimizes memory architectures, and DOME (Wang et al., 2025) tracks temporal state changes. Despite these advances, a gap remains in enabling a socio-evolutionary metabolism that fundamentally transforms agent personality and relations over time, rather than merely accumulating context.

Generative Mise-en-Scène Current text-based agents often suffer from Narrative-Spatial Dissonance, where generated narratives detach from coherent environmental contexts. While Generative Agents (Park et al., 2023) grounds behavior in sandboxes and AgentSociety (Piao et al., 2025) models urban mobility via AOIs, these environments often function as passive containers. BookWorld (Ran et al., 2025) advances this by introducing discrete geospatial tracking and travel constraints. Similarly, NarrativeGenie and HAMLET (Kumaran et al., 2024; Chen et al., 2025) dynamically po-

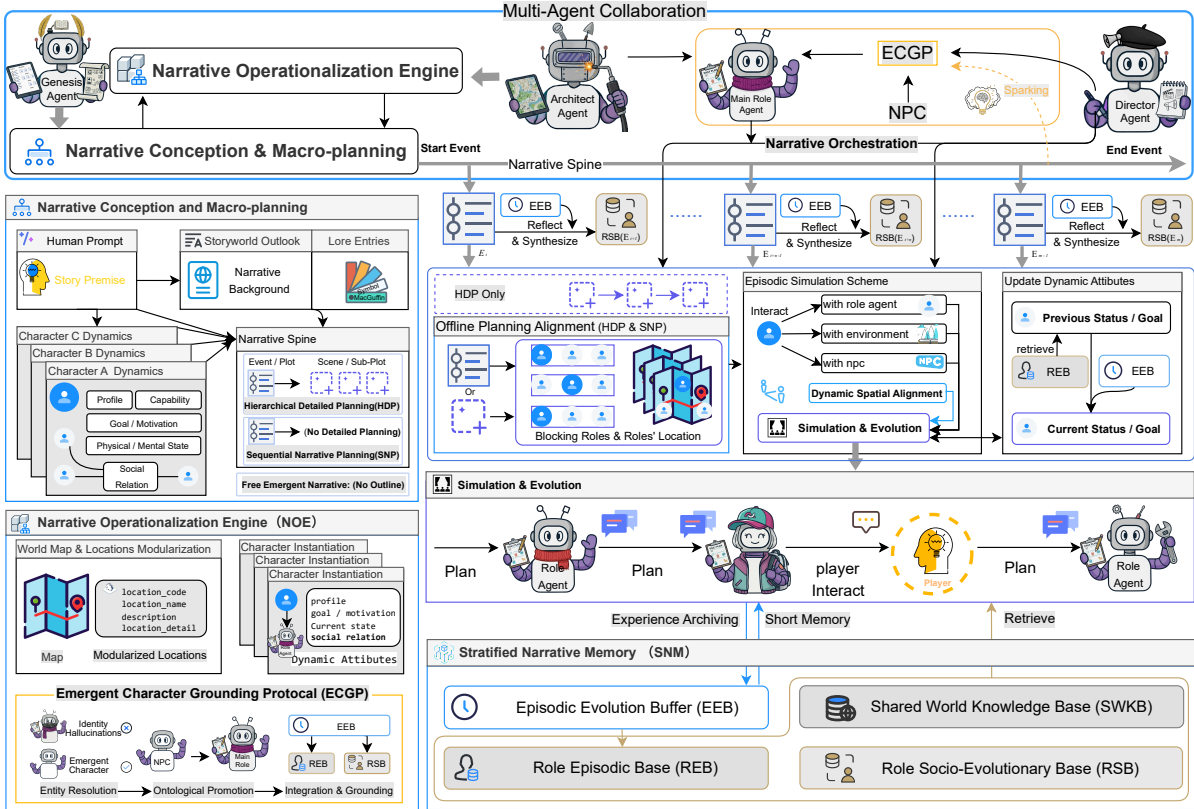


Figure 1: The Architecture of EVOSPAK. The framework initiates with Narrative Conception & Macro-planning, utilizing the Unified Narrative Operation Engine for modularized storyworld and character instantiation. Finally, the Simulation & Evolution module drives the narrative loop, managing continuous interactions via the *Episodic Simulation Scheme* and social memory updates based on the Stratified Narrative Memory.

sition props and adjudicate physical interactions to match narrative beats, while Open-Theatre and HoLLMwood (Xu et al., 2025; Chen et al., 2024) explicitly define spectacle or scene boundaries to constrain agent positioning. However, these frameworks often lack the granularity to maintain essential alignment between the plot, characters, and their specific locations. Standard semantic metrics also remain blind to logical misalignments, leading to interactions that violate spatial consistency (Kumar et al., 2024).

3 EVOSPAK

In this section, we elaborate on the design of EVOSPAK. The primary motive of EVOSPAK is to bridge the gap between static narrative planning and dynamic, open-ended agent interaction. Unlike traditional systems that isolate script control from agent emergence, EVOSPAK functions as a holistic framework tailored for long-horizon consistency. It integrates Narrative Conception and Macro Planning, Simulation and Evolution, and crucially, acts through the Unified Narrative Operation Engine

(NOE), which transforms static blueprints into long-horizon evolutionary story worlds via emergent character grounding. The overall architecture of the EVOSPAK framework is illustrated in Figure 1.

Narrative Conception and Macro Planning Orchestrated by the Genesis Agent, the lifecycle initiates with the synthesis of foundational assets from a user-provided story premise. The agent generates the Storyworld Outlook, lore entries, and character dynamics. Conditioned on the selected control paradigm, these components are structured into a polymorphic Narrative Spine. This spine dictates the simulation’s macro-flow, manifesting as a rigid event hierarchy (HDP), a linear plot sequence (SNP), or an open-ended null state for free emergence (Free EN).

Narrative Operationalization Engine Once the spine is established, the Narrative Operationalization Engine (NOE) structures the simulation environment. As detailed in Figure 1, this layer performs World Map & Locations Modularization

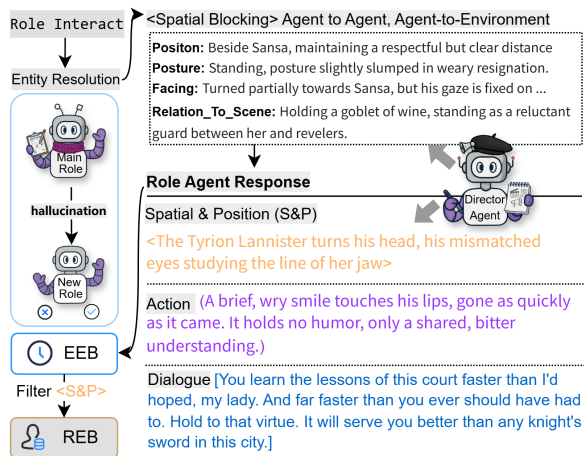


Figure 2: Dynamic Spatial Alignment. The Director Agent orchestrates narrative interactions driven by spatial context, integrating Entity Resolution and precise grounding to ensure logical consistency.

and Character Instantiation, defining static location codes and dynamic agent attributes. Furthermore, it integrates the ECGP, which filters narrative hallucinations and executes Ontological Promotion, transforming valid *sparkling* roles into persistent entities within the system’s cognitive base.

Iterative Simulation & Evolution The execution phase operates under the Episodic Simulation Scheme. As shown in Figure 1, the online simulation executes a continuous loop where Role Agents interact with the environment, Non-Player Characters (NPCs), and players. To ensure **spatial coherence**, this phase integrates the **Generative Mise-en-Scène (GMS)** mechanism. Acting as a “Virtual Stage Manager,” GMS bridges abstract narrative intent with concrete scene execution via a collaborative Plan-Correct protocol between the Genesis and Director agents. It enforces strict Role-Location-Plot (RLP) alignment through a dual-phase process:

- **Offline Planning Alignment:** As detailed in Figure 1, the Genesis Agent establishes foundational logic by aligning constraints across Role, Location, and Plot dimensions, ensuring initial assignments remain logically congruent with authorial intent.
- **Dynamic Spatial Alignment:** The Director Agent leverages spatial blocking to synchronize narrative intentions with real-time contexts. As detailed in Figure 2, this explicitly incorporates an entity resolution step to rectify LLM-induced identity hallucinations (e.g., malformed

role codes), ensuring precise character grounding.

The cycle concludes with *update dynamic attributes*, where the system utilizes the RSB to consolidate the previous status and new experiences into a refreshed current status. This allows agents to assimilate narrative events into long-term cognitive and social evolution.

3.1 Narrative Spectrum Configuration

EVOSPARK serves as a unified, paradigm-agnostic substrate supporting the full spectrum of control granularities. At initialization (T_0), the NOE is configured to one of three paradigms, dynamically aligning the narrative spine and the Director Agent’s intervention policy with user intents:

Hierarchical Detailed Planning (HDP) The system operates on a rigorous event tree. The Director Agent enforces hierarchical constraints to align agent behavior with authorial intent via granular guidance, ensuring high plot fidelity without rigid hard-coding.

Sequential Narrative Planning (SNP) The system generates linear Key Nodes. Agents are motivation-driven to reach these milestones but retain improvisational freedom in path planning and interaction details between nodes.

Free Emergent Narrative (Free EN) Initialized with only foundational settings (Storyworld, Roles), the Director Agent removes plot constraints while retaining interaction guidance. Consequently, narrative trajectories are driven entirely by endogenous agent decisions and emergent conflicts.

3.2 Multi-Agent Collaboration

Our approach empowers the narrative to evolve not through isolated modules, but through the dynamic collaboration of four distinct types of specialized agents.

Genesis Agent As the executor of Narrative Conception, the Genesis Agent processes the human premise to generate the Narrative Spine. Crucially, it initiates the GMS Plan-Correct protocol by collaborating with the Director Agent. During the Offline Planning Alignment phase, it establishes the foundational Role-Location-Plot logic, ensuring that the initial blueprint handed off to the Architect and Director is structurally sound and logically congruent with authorial intent.

Architect Agent This agent acts as the operational core of the NOE. While it executes World Map & Locations Modularization to instantiate the environment, its critical collaborative role lies in the ECGP. It synergizes with the Director Agent to monitor the simulation for *sparkling*. When the Director identifies a valid narrative hallucination, the Architect executes Ontological Promotion, transforming these fleeting mentions into legitimized entities and seamlessly integrating them into the storyworld outlook originally defined by the Genesis Agent.

Director Agent Serving as the conductor of Iterative Simulation, the Director Agent bridges the gap between static plans and dynamic execution. It orchestrates a continuous feedback loop with Role Agents to provide real-time interaction guidance based on the configured paradigm (HDP/SNP/Free EN). Simultaneously, it engages in Dynamic spatial Alignment (part of GMS), correcting potential narrative-spatial misalignments derived from the Genesis Agent’s blueprint. It also acts as the primary filter for the Architect, validating whether a *sparkling* aligns with the current narrative flow before requesting promotion.

Role Agents These agents execute the Episodic Simulation Scheme. Driven by the Role Socio-Evolutionary Base (RSB)—their living cognition—they engage in decentralized interactions with each other and the environment. Rather than acting in isolation, they form the endogenous social graph. Their behaviors are continuously modulated by the Director’s guidance, while their interaction outcomes are metabolized back into the RSB, ensuring that the collective narrative evolution is grounded in consistent, long-term character growth.

3.3 Emergent Character Grounding Protocol (ECGP)

The ECGP operationalizes endogenous character emergence by harnessing stochastic *sparkling*—narrative hallucinations of uninitialized entities. As shown in Figure 1, the protocol captures and integrates these entities via the following pipeline:

Sparkling via Constraint Violation The protocol is triggered by a generative anomaly we term *sparkling*. Despite strict constraints limiting selection to the existing main character list, the LLM may hallucinate an uninitialized name to bridge a narrative gap. ECGP identifies this constraint vio-

lation not as an error, but as a *sparkling*—a signal of latent narrative necessity.

Entity Resolution Upon detection, the Director Agent immediately intercepts the Spark to perform rigorous verification. This step acts as a filter to distinguish genuine new entities from mere aliases (e.g., nicknames or variants of existing IDs). Only when a *sparkling* is confirmed as structurally novel yet contextually coherent is it validated as a legitimate candidate for emergence.

Ontological Promotion This phase executes the core logic of endogenous evolution, formalizing the transition of a *sparkling* from a stochastic byproduct into a simulation entity. Driven by the entity’s plot-criticality, this process elevates its hierarchical status, actively repurposing hallucinations from transient noise into drivers of character evolution and ecosystem expansion.

Integration & Grounding To finalize the emergence, the Architect Agent executes structural integration. It leverages context awareness to instantiate a new profile for the promoted character within the Storyworld Outlook and initializes their cognitive state in the RSB, which serves as the vital bridge for integrating the new main character’s narrative following Ontological Promotion within the ECGP. This process effectively anchors the generated information, transforming what was once a fleeting narrative mention into a permanent, consistent component of the world state, ready to participate in future memory formation and narrative derivation.

3.4 Dynamic Spatial Alignment in GMS

We first observe scenarios where agents exhibit spatial dissonance in the absence of the *Dynamic Spatial Alignment* within GMS. As shown in Table 1, while the semantic plot may progress logically, physical contradictions often emerge during complex interactions, illustrating how agents can become “lost in space” without a persistent spatial context.

To address such discrepancies, the Dynamic-Spatial-Alignment within GMS acts as an implicit spatial perceiver to help agents ground their physical reality. Table 2 illustrates interactions where character responses and dialogues are coherently anchored to their perceived physical environment through this mechanism.

Examples of Narrative-Spatial Dissonance	
Turn	Interaction Detail
T1	Cersei Lannister: (Her voice is a cold, precise command as she addresses the guards without turning, standing rigidly by the hearth with her silhouette sharp against the firelight.) “Bring me the signed contract. Then take the Stark girl to her chambers...” ... (Turn 2: Cersei waits with her back to the room, demanding to see the seals and signatures) ...
T3	Cersei Lannister: (A slow, dangerous pivot brings her eyes to bear on the hesitant guard , her hand remaining outstretched and her posture a blade of ice in the firelight.) “My father’s orders do not supersede mine in this room. The parchment. Now...”
T4	Cersei Lannister: (She snatches the heavy parchment from the man’s trembling fingers, though her eyes, chips of green ice, never leave Sansa’s placid mask.) “Good. Now see to your duties, before I find a captain who knows how to obey.”

Table 1: A fragment exhibiting a physical contradiction in T4. The character is described as maintaining unbroken eye contact with Sansa Stark, which is logically inconsistent with her previous pivot to confront the guard in T3.

Examples with Dynamic Spatial Alignment	
Turn	Interaction Detail
T1	Tyrion Lannister: <i><The Imp stands a half-step behind his wife, Sansa, his mismatched eyes scanning the preening lords and ladies.></i> (His hand drifts toward the small of her back, hesitating before falling away, a silent, aborted gesture of comfort. The Imp’s gaze lingers for a moment on Joffrey, enthroned and sneering, before he leans infinitesimally closer to Sansa’s ear, the rich velvet of his doublet brushing her arm.) “Try not to listen to the crows, little dove. They only caw when there’s carrion. Keep your eyes on me, if you can bear it. I am, at the very least, a smaller monster than most here.”
T2	Sansa Stark: <i><Standing rigidly beside Tyrion, her shoulders drawn tight.></i> (The Little Dove finches almost imperceptibly at the warmth of his breath. Her gaze, which had been fixed on Joffrey with a practiced emptiness, falls to the polished floor, refusing to meet her husband’s eyes.) “As you say, my lord.”
T3	Tyrion Lannister: <i><His gaze shifts from Sansa’s averted face to the celebrating court.></i> (The Imp gives a faint, humorless smile, a barely perceptible twitch of his lips. His attention turns outward, scanning the room with a practiced, weary eye.) “A wise choice, my lady. The floor, at least, is honest. It makes no pretense of friendship while waiting to trip you. A virtue one can’t find in many a lord here.”
T4	Sansa Stark: <i><Standing beside him, a captive bird in a gilded cage.></i> (The Little Dove’s fingers tighten on the silken fabric of her gown, the knuckles of her gloved hands turning white for an instant. Her gaze remains lowered, tracing the veins in the polished marble as if they were a map leading far away from this place.) “It also keeps its counsel, my lord. A virtue rarer than honesty in this court, and far more valuable.”

Table 2: Examples of generated narratives with GMS. The *<orange text>* denotes spatial constraints, while text in parentheses indicates non-verbal actions. The Dynamic-Spatial-Alignment within GMS acts as an implicit spatial perceiver to help agents ground their physical reality.

4 Stratified Narrative Memory (SNM)

To support long-horizon story evolution, we introduce Stratified Narrative Memory (SNM). Unlike flat memory systems that suffer from *social memory stacking*—where contradictory historical states persist—SNM adopts a layered architecture. It systematically segregates global truth, linear provenance, and evolving socio-cognitive states, ensuring agents act based on a *living* cognition.

4.1 Hierarchical Memory Architecture

As shown in Figure 1, SNM bridges immediate perception and persistent storage via four distinct components:

- **Episodic Evolution Buffer (EEB):** A short

episode memory that caches real-time interactions (E_i) and sensory data, serving as a staging area before long-term metabolization.

- **Shared World Knowledge Base (SWKB):** Stores immutable **global truths** (e.g., lore, world-view), providing consistent ground truth for all agents.
- **Role Episodic Base (REB):** An immutable **experience log** for provenance tracking. It is decoupled from immediate decision-making to prevent context pollution.
- **Role Socio-Evolutionary Base (RSB):** The core **mutable storage** for the agent’s current snapshot (personality, social graphs). It evolves continuously via in-place updates.

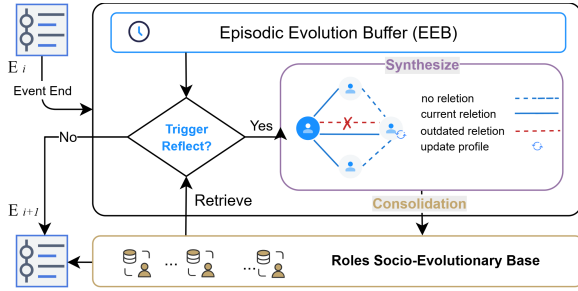


Figure 3: The event-driven Reflect-Synthesize-Consolidation mechanism.

4.2 Reflect-Synthesize-Consolidation

To resolve conflicts between past and present states, we implement an event-driven Reflect-Synthesize-Consolidation mechanism (Figure 3). Instead of simply stacking memories, the system actively *assimilates* data from the EEB into the persistent RSB:

- **Reflection Trigger:** During an event (E_i), raw interactions accumulate in the **EEB**. Upon event conclusion, the system executes a trigger check. If the interaction intensity exceeds a threshold, the system retrieves relevant context from the RSB; otherwise, it proceeds directly to the next event (E_{i+1}).
- **Synthesize:** This phase functions as a cognitive workspace to compute the evolutionary delta. It contrasts the emerging data cached in the EEB against established RSB states. As illustrated in Figure 3, the system explicitly resolves topological shifts by severing outdated relations (visualized as crossed-out edges) and synchronously updating both social connections and character profiles.
- **Consolidation:** The synthesized state is committed to the **RSB**. Crucially, this is an *in-place update* operation. Old personality vectors and social edges are overwritten by the new synthesis, ensuring the RSB remains a consistent, conflict-free snapshot of the agent’s current reality.

5 Experiments

In this section, we validate the effectiveness of EVOSPARK through a rigorous comparative evaluation. We assess the framework’s performance across diverse narrative control paradigms, with a primary focus on the quality of story simulation and evolution.

5.1 Experimental Setup

We utilize curated scenarios covering six distinct genres (e.g., Mystery, Sci-Fi, Epic Fantasy). While we conduct validation across varying narrative lengths (Appendix B), our primary evaluation prioritizes the challenging long-horizon setting within a representative subset of these genres. In this setting, each simulation entails a continuous sequence of 15 significant events (~ 45 scenes), yielding an average corpus of 200k–250k words per run to rigorously test evolutionary consistency.

Baselines We compare EVOSPARK against three representative frameworks representing distinct narrative paradigms:

- **Open-Theatre** (Xu et al., 2025): A script-driven framework utilizing a Director-Global Actor architecture, representing the centralized control paradigm.
- **BookWorld** (Ran et al., 2025): A virtual world simulation framework that models established environments and NPC interactions with high fidelity.
- **HoLLMwood** (Chen et al., 2024): A creative writing agent framework that replicates writer-editor workflows to refine narrative quality via multi-agent collaboration.

5.2 Evaluation Metrics

We define a comprehensive set of metrics tailored to specific narrative paradigms.

Universal Metrics:

- **Role Performance (RP):** Evaluates agent believability, ensuring actions and dialogues align with preset personas and evolving memories.
- **Immersion (Im):** Measures user engagement by assessing how effectively agents interact with the environmental context to provoke emotional resonance.

HDP & SNP Metrics:

- **Narrative Resonance (NR):** Evaluates structural depth. It measures thematic adherence to the blueprint and structural integrity to evoke reader empathy.
- **Long-Horizon Consistency (LC):** Assesses logical stability across extended segments, ensuring smooth transitions and strict adherence to the narrative spine.

Free EN Metrics:

- **Narrative Soundness (NS):** Verifies causal feasibility, ensuring event preconditions are met and agent actions remain rational and goal-oriented.

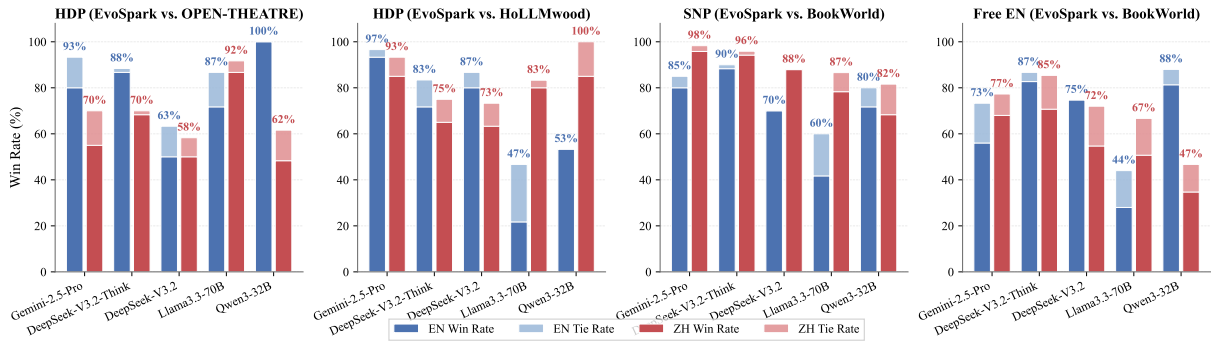


Figure 4: Comparison of win/tie rates between EVOSPARK and baseline frameworks across different narrative modes (HDP, SNP, Free EN), languages, and LLM backbones. Detailed metric breakdowns are in Appendix B.

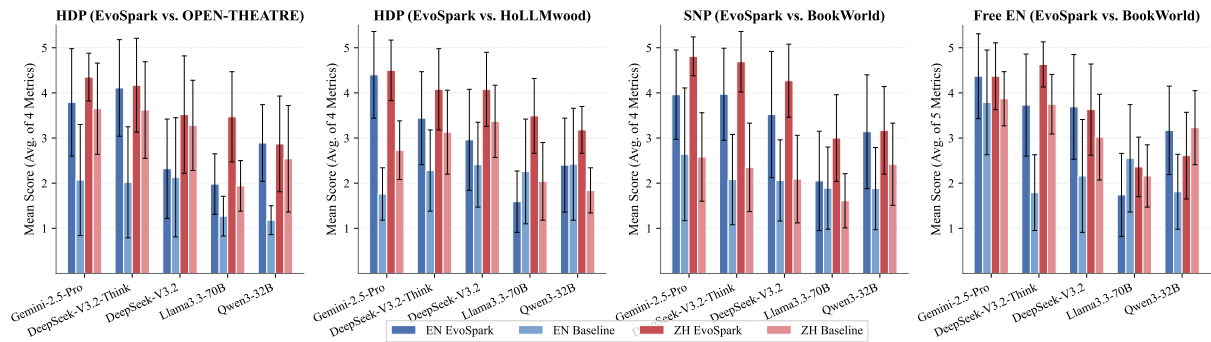


Figure 5: Comparison of overall average scores. The reported values are aggregated mean scores of underlying metrics. Detailed results are provided in Appendix B.

- **Creativity (Cr):** Assesses content novelty. It evaluates the uniqueness of plot twists and character portrayals, penalizing generic stereotypes.
- **Plot Advancement (PAC):** Evaluates simulation momentum, rewarding logical conflict escalation and penalizing stagnation.

5.3 Evaluation Methodology.

Following BookWorld (Ran et al., 2025), we employ a pairwise **LLM-as-a-Judge** protocol with position swapping, using Gemini-2.5-Pro (English) and Deepseek-v3.2-Think (Chinese). We report Win Rates and Average Likert Scores (1–5) (Likert, 1932). The credibility of our evaluation pipeline is substantiated by its consistency with human evaluation, presented in Appendix A.

5.4 Evaluation Results

Figures 4 and 5 illustrate the performance of EVOSPARK across varying paradigms, languages, and backbones.

Overall Superiority on Reasoning Models. EVOSPARK significantly outperforms baselines in most settings, particularly on reasoning-enhanced models (e.g., Gemini-2.5-Pro, DeepSeek-V3.2-

Think). It achieves dominant win rates and margins in *role performance*, *narrative resonance*, and *immersion*. This success is attributed to the synergy between advanced reasoning capabilities and our cognitive modules (ECGP, GMS), which demand complex instruction following to maintain consistency.

Complexity Gap and Stochasticity. Conversely, performance dips with non-reasoning models (e.g., Llama3.3-70B, Qwen3-32B) and in *Free EN* mode. The former reflects an instruction-complexity gap, where EVOSPARK’s high cognitive load (spatial/memory constraints) overwhelms weaker models. In *Free EN*, volatility stems from minimal framework intervention: relaxing planning constraints increases autonomy and stochasticity, naturally causing higher variance than rigidly controlled paradigms.

5.5 Long-Horizon Evolutionary Alignment

Distinct from previous evaluations focused on single events, this analysis investigates system performance across continuous narrative horizons spanning 1, 5, and 10 events. The primary objective is to verify whether agent behaviors can dynamically

		HDP			SNP			Free EN				
Horizon Length	w vs. w/o RSB-Rel	1	46.7 +13.3	40.0 +26.7	60.0 +6.7	46.7 +20.0	46.7 +20.0	60.0 +6.7	53.3 +33.3	40.0 +13.3	73.3 +6.7	Win+Tie (%)
		5	54.5 +27.3	36.3 +27.2	45.4 +54.5	81.8 +0.0	72.7 +0.0	90.9 +9.1	36.3 +54.5	36.4 +9.1	45.5 +36.4	
		10	50.0 +16.7	50.0 +16.7	83.3 +0.0	100.0 +0.0	100.0 +0.0	100.0 +0.0	66.7 +0.0	50.0 +33.3	33.3 +33.3	
			RP	LC	EAA	RP	LC	EAA	RP	PAC	EAA	
	w vs. w/o RSB	1	60.0 +13.3	46.7 +33.3	80.0 +6.7	46.7 +13.3	33.3 +20.0	66.7 +6.7	46.7 +13.3	53.3 +6.7	80.0 +13.3	
		5	45.4 +9.1	81.8 +0.0	63.7 +0.0	90.9 +9.1	81.8 +0.0	100.0 +0.0	63.6 +9.1	54.5 +9.1	90.9 +9.1	
10		66.7 +0.0	83.3 +0.0	83.3 +0.0	100.0 +0.0	100.0 +0.0	100.0 +0.0	66.7 +0.0	50.0 +16.7	83.3 +16.7		
		RP	LC	EAA	RP	LC	EAA	RP	PAC	EAA		

Figure 6: Long-Horizon Evolutionary Alignment Results: Win rates (bold) and tie rates of the full model vs. variants across 1, 5, and 10 events.

		HDP				SNP				Free EN				
w/o RSB-Rel	1	46.7	40.0	46.7	46.7	46.7	46.7	46.7	46.7	53.3	40.0	46.7	46.7	Win+Tie (%)
	5	46.7	40.0	46.7	46.7	46.7	46.7	46.7	46.7	53.3	40.0	46.7	46.7	
	10	46.7	40.0	46.7	46.7	46.7	46.7	46.7	46.7	53.3	40.0	46.7	46.7	
			RP	LC	NR	Im	RP	LC	NR	Im	RP	NSPAC	Cr	
w/o RSB	1	60.0	46.7	60.0	40.0	46.7	33.3	40.0	46.7	46.7	53.3	37.3	35.3	Win+Tie (%)
	5	60.0	46.7	60.0	40.0	46.7	33.3	40.0	46.7	46.7	53.3	37.3	35.3	
	10	60.0	46.7	60.0	40.0	46.7	33.3	40.0	46.7	46.7	53.3	37.3	35.3	
			RP	LC	NR	Im	RP	LC	NR	Im	RP	NSPAC	Cr	
w/o GMS	1	86.7	66.7	77.3	386.7	60.0	53.3	66.7	53.3	60.0	53.3	61.3	61.3	Win+Tie (%)
	5	86.7	66.7	77.3	386.7	60.0	53.3	66.7	53.3	60.0	53.3	61.3	61.3	
	10	86.7	66.7	77.3	386.7	60.0	53.3	66.7	53.3	60.0	53.3	61.3	61.3	
			RP	LC	NR	Im	RP	LC	NR	Im	RP	NSPAC	Cr	
w/o ECGP	1	46.7	53.3	346.7	66.7	53.3	346.7	60.0	60.0	66.7	46.7	60.0	66.7	Win+Tie (%)
	5	46.7	53.3	346.7	66.7	53.3	346.7	60.0	60.0	66.7	46.7	60.0	66.7	
	10	46.7	53.3	346.7	66.7	53.3	346.7	60.0	60.0	66.7	46.7	60.0	66.7	
			RP	LC	NR	Im	RP	LC	NR	Im	RP	NSPAC	Cr	

Figure 7: Ablation Study Results: Pairwise comparison heatmap between full EVOSPARK and ablated variants. Darker red indicates higher win rates for the full model.

shift in accordance with changes in character relationships or social identities, while maintaining long-horizon consistency. Central to this evaluation is **Evolutionary Action Alignment (EAA)**, a metric designed to quantify the synchronization between these evolutionary changes and agent actions.

As illustrated in Figure 6, comparisons across horizons reveal that win rates for **RP**, **LC**, and **EAA** improve significantly as the event count increases, particularly under the **SNP** and **Free EN** paradigms. While **HDP** also demonstrates gains, they are less pronounced than in other modes. We attribute this to the rigid constraints of **HDP**’s deep plot planning, which partially attenuate the dynamic social evolution driven by the **RSB**.

5.6 Ablation Study

To isolate the contributions of specific modules, we evaluated four ablated variants: (1) **No-RSB** (excluding the Role Socio-Evolutionary Base); (2) **No-RSB-rel** (disabling relationship evolution); (3)

No-GMS (omitting the Generative Mise-en-scène mechanism); and (4) **No-ECGP** (removing the Emergent Character Grounding Protocol).

Results using `gemini-2.5-pro` (Figure 7) align with expectations. Removing **GMS** causes the most severe degradation across role performance, resonance, and immersion, confirming that **GMS**’s offline and dynamic alignment is foundational for believability.

Disabling **ECGP** notably impairs immersion and creativity by restricting endogenous character emergence. In contrast, the **No-RSB** variant exhibits a relatively marginal decline in these immediate evaluations. We attribute this to the temporal nature of the module: the **RSB** is designed to mitigate cumulative memory conflicts, the effects of which become pronounced primarily over longer horizons (as detailed in Section 5.5) rather than in short-term comparisons.

6 Conclusion

We present **EVOSPARK**, a framework designed to sustain logically coherent, long-horizon narratives within endogenous agent societies. Unlike static or script-driven approaches, our method resolves *social memory stacking* via *living* cognitive metabolism and mitigates *narrative-spatial dissonance* through generative mise-en-scène. Experiments demonstrate that **EVOSPARK** significantly outperforms baselines in logical consistency and social fidelity across extended horizons. By leveraging stochastic hallucinations as structural narrative assets, our system enables the infinite expansion of open-ended story worlds. We hope this work paves the way for future advancements in autonomous narrative intelligence.

Limitations

Despite **EVOSPARK**’s advancements in long-horizon narrative consistency, certain limitations persist.

Efficiency and Computational Overhead Primarily, although the **GMS** and **RSB** updates are event-driven to conserve computational resources, the progressive accumulation of extensive narrative histories and evolving relationship graphs incurs substantial memory overhead and increased inference latency over extended simulations. This currently constrains the framework’s deployment in resource-limited or strictly real-time environments.

Interactivity Boundaries and HITL Integration

Furthermore, because our current evaluation protocols prioritize autonomous agent-to-agent endogenous evolution to rigorously validate internal systemic coherence, the dynamics of real-time human-player interactivity are not empirically quantified in this work. Nevertheless, EVOSPARK’s architecture already implements functional interfaces for Human-in-the-Loop (HITL) integration. Specifically, the framework provides dual participation channels—unconstrained natural language inputs and discrete selection options—allowing users to be embedded directly as equal, embodied Role Agents. This embodied integration is particularly adaptable to potential applications such as interactive educational drama and virtual gaming. To preserve user agency when facing unpredictable human inputs in these scenarios, future iterations can further focus on amplifying the contextual weight of user interactions. By processing player actions as high-priority variables within GMS constraints, the framework enables users to subtly steer narrative outcomes and trigger coherent society-wide socio-evolutionary reactions. This established synthesis of autonomous emergence and human co-creation lays a solid foundation for future dedicated user-centric evaluations.

Ethics Statement

Data Provenance and Synthetic Generation

Distinct from scraped public corpora, our datasets are **synthetically generated** within the EVOSPARK framework. We cover six narrative scenarios, with long-horizon experiments concentrating on *Epic Fantasy* and *Eastern Fantasy* genres. These scenarios are constructed based on specific domain constraints to simulate fictional social dynamics. Consequently, our data contains no personally identifiable information (PII) or private real-world data, eliminating risks related to privacy infringement or copyright violation of existing literary works.

Human Evaluation To validate our automatic metrics, we conducted human evaluations involving university students proficient in both English and Chinese. We strictly adhered to ethical research practices: all participants were provided with compensation well above the local minimum wage, and informed consent was obtained regarding the usage of their annotations. We maintained strict anonymity to protect annotator privacy.

Societal Impact and Risks EVOSPARK simulates complex social dynamics and narrative evolution. We acknowledge the potential risk that the model could be misused to generate misleading content or simulate harmful social biases inherent in the training data. To mitigate this risk, we implemented safety constraints within our system prompts to filter toxic outputs. However, as with all generative agents, we emphasize that EVOSPARK should be utilized responsibly for educational, creative, and research applications, with careful oversight required for any deployment in open-ended user interactions.

Acknowledgments

We gratefully acknowledge the National Key Laboratory of Time and Space Information Precision Sensing at the Department of Precision Instrument, Tsinghua University, for providing the computational platform and resources that made this research possible.

References

- Jing Chen, Xinyu Zhu, Cheng Yang, Chufan Shi, Yadong Xi, Yuxiang Zhang, Junjie Wang, Jiashu Pu, Tian Feng, Yujiu Yang, and Rongsheng Zhang. 2024. Hollmwood: Unleashing the creativity of large language models in screenwriting via role playing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8075–8121, Miami, Florida, USA. Association for Computational Linguistics.
- Sizhou Chen, Shufan Jiang, Chi Zhang, Xiao-Lei Zhang, and Xuelong Li. 2025. [Hamlet: Hyperadaptive agent-based modeling for live embodied theatrics](#). *Preprint*, arXiv:2507.15518.
- Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. 2023. [S³: Social-network simulation system with large language model-empowered agents](#). *arXiv preprint arXiv:2307.14984*.
- Senyu Han, Lu Chen, Li-Min Lin, Zhengshan Xu, and Kai Yu. 2024. Ibsen: Director-actor agent collaboration for controllable and interactive drama script generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1607–1619, Bangkok, Thailand. Association for Computational Linguistics.
- Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. [MetaGPT: Meta programming for a multi-agent collaborative framework](#).

- In *The Twelfth International Conference on Learning Representations*.
- Wenyue Hua, Lizhou Fan, Lingyao Li, Kai Mei, Jianchao Ji, Yingqiang Ge, Libby Hemphill, and Yongfeng Zhang. 2023. [War and Peace \(WarAgent\): Large language model-based multi-agent simulation of world wars](#). *arXiv preprint arXiv:2311.17227*.
- Vikram Kumaran, Jonathan Rowe, and James Lester. 2024. Narrativegenie: Generating narrative beats and dynamic storytelling with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 20(1):76–86.
- Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. [CAMEL: Communicative agents for “mind” exploration of large language model society](#). In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, volume 36, pages 51991–52008. Curran Associates, Inc.
- Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 22(140):1–55.
- Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simula-
cra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, New York, NY, USA. Association for Computing Machinery.
- Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. 2025. [Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society](#). *Preprint*, arXiv:2502.08691.
- Daniel Platnick, Mohamed E. Bengueddache, Marjan Alirezaie, Dava J. Newman, Alex “Sandy” Pentland, and Hossein Rahnama. 2025. [Id-rag: Identity retrieval-augmented generation for long-horizon persona coherence in generative agents](#). *Preprint*, arXiv:2509.25299.
- Yiting Ran, Xintao Wang, Tian Qiu, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2025. Bookworld: From novels to interactive agent societies for story creation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15898–15912, Vienna, Austria. Association for Computational Linguistics.
- Yuqian Sun, Phoebe J. Wang, John Joon Young Chung, Melissa Roemmele, Taewook Kim, and Max Kreminski. 2025. [Drama llama: An llm-powered storylets framework for authorable responsiveness in interactive narrative](#). *Preprint*, arXiv:2501.09099.
- Qianyue Wang, Jinwu Hu, Zhengping Li, Yufeng Wang, and Daiyuan Li. 2025. Generating long-form story using dynamic hierarchical outlining with memory-enhancement. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Albuquerque, New Mexico. Association for Computational Linguistics.
- Yi Wang, Qian Zhou, and David Ledo. 2024. Storyverse: Towards co-authoring dynamic plot with llm-based character simulation via narrative planning. <https://arxiv.org/abs/2405.13042v2>.
- Haotian Xia, Hao Peng, Yunjia Qi, Xiaozhi Wang, Bin Xu, Lei Hou, and Juanzi Li. 2025. [Storywriter: A multi-agent framework for long story generation](#). *Preprint*, arXiv:2506.16445.
- Tianyang Xu, Hongqiu Wu, Weiqi Wu, and Hai Zhao. 2025. Open-theatre: An open-source toolkit for llm-based interactive drama. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 453–460, Suzhou, China. Association for Computational Linguistics.
- Xiangyuan Xue, Yifan Zhou, Guibin Zhang, Zaibin Zhang, Yijiang Li, Chen Zhang, Zhenfei Yin, Philip Torr, Wanli Ouyang, and Lei Bai. 2025. [Comas: Co-evolving multi-agent systems via interaction rewards](#). *Preprint*, arXiv:2510.08529.
- Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, Prateek Gupta, Shuyue Hu, Zhenfei Yin, Guohao Li, Xu Jia, Lijun Wang, Bernard Ghanem, Huchuan Lu, Chaochao Lu, and 4 others. 2025. [Oasis: Open agent social interaction simulations with one million agents](#). *Preprint*, arXiv:2411.11581.
- Guibin Zhang, Muxin Fu, Guancheng Wan, Miao Yu, Kun Wang, and Shuicheng Yan. 2025a. [G-memory: Tracing hierarchical memory for multi-agent systems](#). *Preprint*, arXiv:2506.07398.
- Guibin Zhang, Haotian Ren, Chong Zhan, Zhenhong Zhou, and Junhao Wang. 2025b. [Memevolve: Meta-evolution of agent memory systems](#).
- Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune. 2025c. [Darwin Gödel machine: Open-ended evolution of self-improving agents](#). *arXiv preprint arXiv:2505.22954*.
- Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xionghui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, Bingnan Zheng, Bang Liu, Yuyu Luo, and Chenglin Wu. 2025d. [AFlow: Automating agentic workflow generation](#). In *The Thirteenth International Conference on Learning Representations*.
- Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2023. [Memorybank: Enhancing large language models with long-term memory](#). *Preprint*, arXiv:2305.10250.

A Consistency with Human Evaluation

To validate the reliability of our model-based evaluation approach, we conducted a comprehensive agreement analysis between model assessments and human assessments.

Setup and Methodology We recruited 8 human annotators to evaluate outputs from our proposed method and baselines. To ensure a robust evaluation, we randomly sampled partial data across diverse narrative genres. The evaluation was conducted using a 5-point Likert scale, where annotators and the model independently scored the generated narratives. To quantify human-model agreement, we computed Cohen’s Kappa (κ) between aggregated human judgments (majority vote) and model assessments. As shown in Table 3, results indicate substantial agreement ($\kappa = 0.62$ – 0.76) across all paradigms, confirming that our automated metrics align closely with human perception.

As shown in Table 3, the results indicate substantial agreement across all paradigms, confirming that our automated evaluation metrics align closely with human perception.

Paradigm	Metric	Kappa (κ)
HDP	Role Performance (RP)	0.71
	Logical Consistency (LC)	0.69
	Narrative Resonance (NR)	0.65
	Immersion (Im)	0.69
SNP	Role Performance (RP)	0.73
	Logical Consistency (LC)	0.76
	Narrative Resonance (NR)	0.68
	Immersion (Im)	0.71
Free EN	Role Performance (RP)	0.70
	Narrative Soundness (NS)	0.68
	Creativity (Cr)	0.62
	Immersion (Im)	0.72
	Plot Adv. & Conflict (PAC)	0.73

Table 3: Agreement between Human and Model Judges across different control paradigms.

B Implementation Details and Supplementary Results

This appendix supplements the main experimental analysis with three key components: (1) validation performance on standard narrative benchmarks across diverse genres, establishing the framework’s versatility; and (2) a granular breakdown of the specific metric scores for the long-horizon experiments reported in Section 5.4; and (3) the specific hyperparameter configurations and generation settings

utilized during inference.

Specifically for the experimental setup, all language model API requests were executed using non-streaming outputs. We set the sampling temperature to 0.8 to strike an optimal balance between narrative creativity and logical coherence. All other decoding parameters, including `top_p` and `top_k`, were retained at their default values.

B.1 Performance on Standard Benchmarks

To verify the framework’s versatility and stability across distinct narrative styles, we conducted evaluations on a standard benchmark suite covering six genres: *mystery*, *classical drama*, *science fiction*, *modern drama*, *epic fantasy*, and *eastern fantasy*.

Unlike the long-horizon stress tests focused on evolutionary consistency in the main text, these simulations follow a standard episodic structure (~ 3 significant events, $\sim 6k$ words) to ensure broad genre coverage. Figure 8 presents the comparative results. EVOSPARK demonstrates robust performance and stylistic adaptability across all tested domains, confirming its effectiveness in handling diverse genre-specific constraints alongside complex long-term evolution.

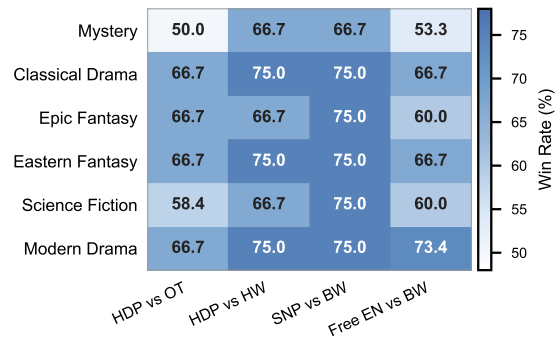


Figure 8: Cross-Domain Performance Comparison. Average win rates (%) of EVOSPARK against baselines across three paradigms (HDP, SNP, and Free EN) in six narrative domains. OT: OpenTheatre, HW: HoLLM-wood, BW: BookWorld.

B.2 Granular Metric Analysis for Long-Horizon Experiments

In the main paper, we reported aggregated performance metrics to provide an overall assessment. Here, we present the comprehensive breakdown across all individual evaluation metrics.

Figure 9 illustrates the detailed pairwise Win Rates for every specific metric (RP, Im, NR, LC,

NS, Cr, PAC) across different languages and backbones. Correspondingly, Figure 10 provides the detailed Average Likert Scores (1–5).

These granular results further substantiate that EVOSPARK’s improvements are not limited to a single dimension but are distributed across character fidelity, spatial immersion, and narrative logic.

B.3 Granular Ablation for RLP Alignment in GMS

To explicitly isolate the functional contribution of the Generative Mise-en-Scène (GMS) module, we conduct a fine-grained ablation study focusing specifically on the Role-Location-Plot (RLP) Alignment mechanism. While Section 5.6 presents the aggregated ablation results, Table 4 provides a detailed breakdown of pairwise win/tie rates and absolute mean scores across different narrative control paradigms (HDP, SNP, and Free EN).

Paradigm	Metric	Win / Tie (%)
HDP	Role Performance (RP)	86.7 / 6.7
	Logical Consistency (LC)	66.7 / 6.7
	Narrative Resonance (NR)	73.3 / 13.3
SNP	Role Performance (RP)	60.0 / 33.3
	Logical Consistency (LC)	53.3 / 13.3
	Narrative Resonance (NR)	66.7 / 33.3
Free EN	Role Performance (RP)	60.0 / 13.3
	Narrative Soundness (NS)	53.3 / 20.0
	Plot Adv. & Conflict (PAC)	60.0 / 20.0
	Creativity (Cr)	73.3 / 13.3

Table 4: Granular ablation results for the GMS mechanism (Full Model vs. w/o RLP Alignment). Values denote the pairwise **Win / Tie** rates (%). The high win rates demonstrate the indispensability of spatial alignment across all paradigms.

B.4 Efficiency and Latency Analysis

Efficiency and Latency Analysis To investigate the computational overhead required to sustain long-horizon consistency, we evaluated the running efficiency of all systems over 100 aligned narrative steps. Specifically, this benchmark was conducted under the SNP mode utilizing the `deepseek-chat` model as the underlying reasoning engine. We define an *Interaction Turn* as a complete simulation cycle where all agents perceive and act, and an *LLM Call* as a single API request.

Table 5 summarizes the execution overhead across all systems. While EVOSPARK exhibits higher total duration and turn latency compared to

baselines, this is primarily attributed to the multi-agent coordination required by the GMS and RSB modules to ensure narrative logic. Notably, the *Avg/LLM Call* metric reveals that EVOSPARK maintains high inference efficiency at the individual request level, suggesting that the latency trade-off is a direct result of the enhanced cognitive modeling necessary for long-term narrative evolution.

C Evaluation Scenario Details

In this section, we provide the detailed specifications of the six narrative scenarios used in our experiments. Table 6 outlines the metadata for each domain, including genre, title, language, and source type. The specific story premises utilized for simulation initialization are detailed in the subsequent text.

Impact of Prior Knowledge A critical consideration in evaluating long-horizon narrative frameworks is verifying whether a system’s performance relies on memorization acquired during the LLM’s pre-training phase. To objectively assess genuine deductive reasoning capabilities and rule out the interference of prior knowledge, our dataset purposefully includes both *Existing* scenarios (based on famous literary works, thus possessing prior knowledge) and *Synthesized* scenarios (completely original narratives lacking prior knowledge).

Table 7 presents a comprehensive comparative analysis of EVOSPARK’s performance under both scenario conditions. The results detail the win rates against baselines as well as the absolute Average Likert scores achieved by EVOSPARK itself. The varying distribution of winning conditions across different LLM backbones demonstrates that EVOSPARK’s superiority is not strictly dependent on pre-trained memorization.

D Main Prompts Details

To facilitate reproducibility and provide transparency into the EVOSPARK implementation, we present the core prompt templates utilized across key functional modules. Specifically, we provide the system instructions for dynamic relationship evolution (Table 8), long-term memory updates (Table 9), generative spatial blocking (Table 10), and emergent character instantiation (Table 11). In these templates, terms enclosed in curly braces (e.g., `{relation}`) denote dynamic placeholders filled by the EVOSPARK engine during runtime based on the real-time simulation context.

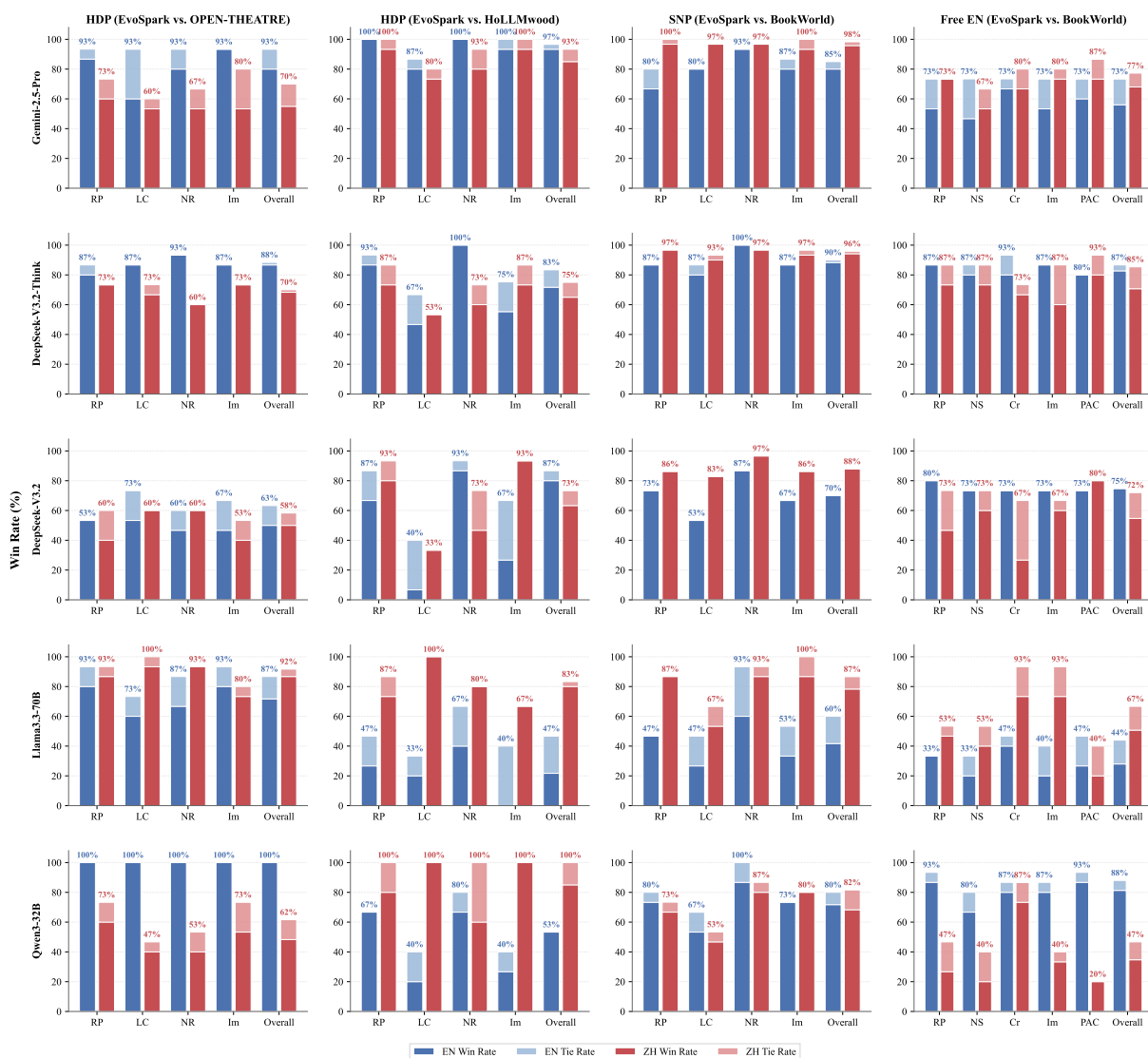


Figure 9: Detailed Win Rates of EvoSpark vs. Baselines across all individual evaluation metrics. This breakdown covers Role Performance (RP), Immersion (Im), Narrative Resonance (NR), Long-Horizon Logical Consistency (LC), Narrative Soundness (NS), Creativity (Cr), and Plot Advancement and Conflict (PAC).

System	Total Duration	Avg / Turn	Median / Turn	Min / Max Turn	Avg / LLM Call
EVOSPARK	63.6 min	38.17 s	42.56 s	3.99 s / 80.93 s	3.30 s
OpenTheatre	41.2 min	24.72 s	24.74 s	16.51 s / 33.49 s	7.89 s
BookWorld	25.9 min	15.53 s	12.66 s	2.60 s / 58.01 s	2.15 s
HOLLMwood	9.1 min	5.46 s	5.04 s	2.29 s / 18.82 s	4.77 s

Table 5: Turn Latency Comparison based on a 100-Turn Benchmark. All statistics reflect resource consumption during the evolutionary simulation. Avg / LLM Call isolates pure inference efficiency.

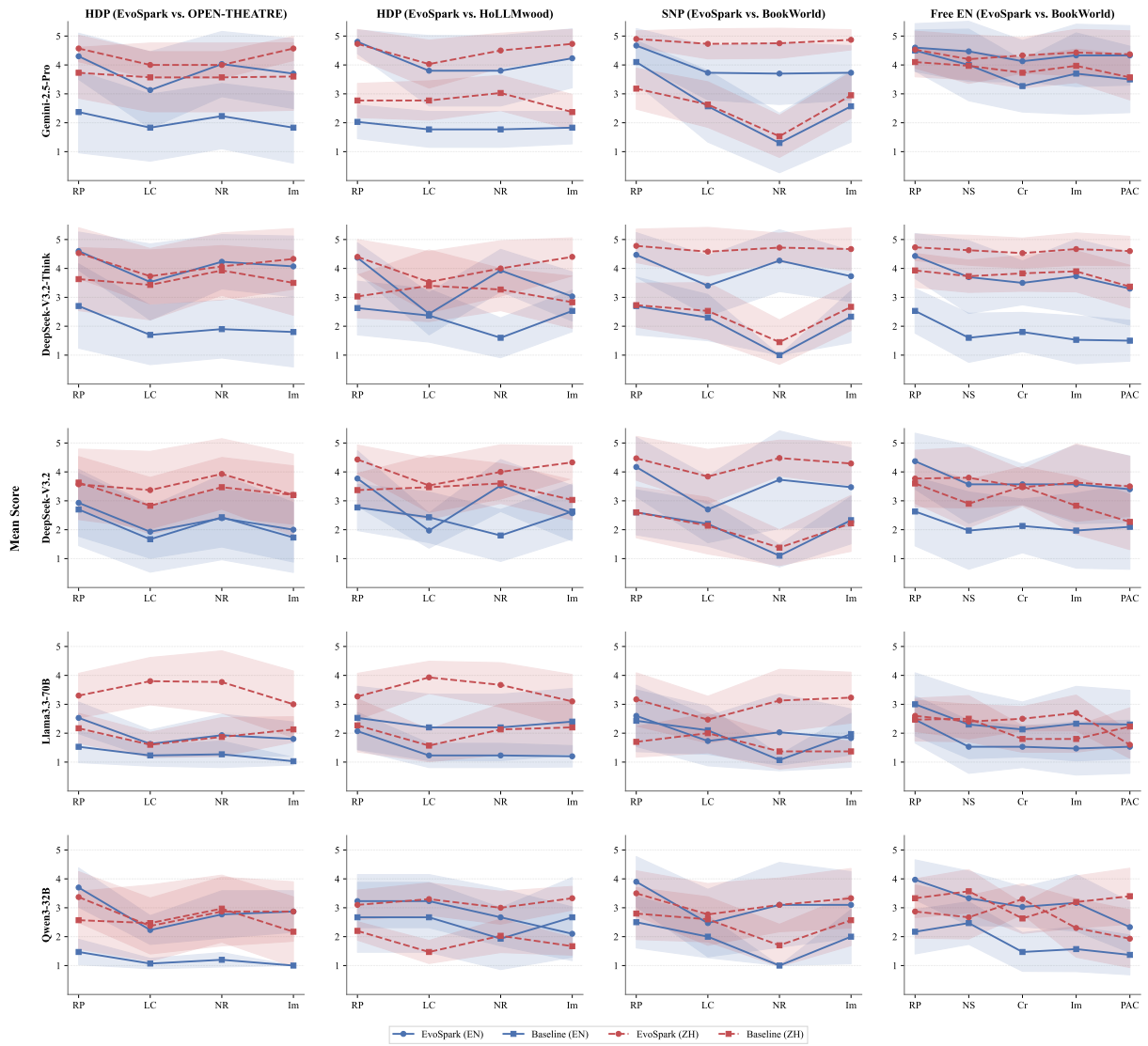


Figure 10: Detailed Average Scores (1–5) of EvoSpark vs. Baselines across all individual evaluation metrics. The results demonstrate consistent superiority in HDP, SNP and Free EN paradigms across diverse LLM backbones.

Genre (Title)	Lang.	Type	Story Premise
Mystery	ZH	Existing	<i>The Longest Day in Chang'an</i> . On the Lantern Festival, undercurrents surge in Chang'an as Turkic "Wolf Guards" infiltrate, plotting a devastating fire attack named "Quelehuoduo." Li Bi, head of the Jing'an Department, recruits death-row inmate Zhang Xiaojing to save the city. Within twelve hours, this duo navigates both street chases and court politics. They uncover that the terror plot intertwines with the Crown Prince's struggle and official conspiracies aimed at the Tang Dynasty's foundation. Ultimately, they risk their lives to prevent the Lantern Tower explosion, resolving the imperial crisis.
Classical Drama	EN	Existing	<i>Romeo and Juliet</i> . Amidst the feud between Montagues and Capulets in Verona, two star-crossed lovers marry in secret. A chain of tragic misunderstandings and fatal duels forces them toward a heartbreaking destiny that ultimately unites their warring families in grief.
Epic Fantasy	EN	Existing	<i>A Song of Ice and Fire</i> . Following the Red Wedding, the scattered remnants of House Stark must forge new identities through trauma. They aim to dismantle the alliance of terror, reclaim Winterfell, and restore honor in a brutal saga of vengeance and rebirth.
Eastern Fantasy	ZH	Synthesized	<i>Ten Thousand Years of Qi Refining</i> . Xu Yang, ostensibly a low-level cultivator, is actually a 100,000-year-old entity stronger than gods. When his sect faces destruction, he emerges to crush enemies with "accidental" displays of overwhelming power. Bound by a "Source Shackle," he must eventually face upper-realm deities to protect his legacy.
Sci-Fi	EN	Synthesized	<i>The War of Lost Will (2145)</i> . Scientist Lin Shen discovers his "Memory Chip" tech is corrupted by a "Cognitive Pollution Program." To prevent humanity from losing free will, he teams up with a rogue AI "Zero" and agent Su Li to crack the code within 72 hours, battling both external enemies and his own fading memory.
Modern Drama	ZH	Synthesized	<i>I am the Sky of Dragon City</i> . Huo Tian, a humiliated delivery man, awakens a "Future Vision" (3-second foresight). Discovering that a wealthy heir plotted his mother's illness to steal a family recipe, he uses his ability to win high-stakes bets and dismantle the heir's family empire in a story of tactical revenge.

Table 6: Detailed specifications of the evaluation scenarios. The dataset covers six distinct genres, comprising three existing literary works and three synthesized open-ended narratives. *Lang.* denotes the simulation language (ZH: Chinese, EN: English).

Model	Baseline	Win Rate (%)		EVOSPARK Avg. Likert		Winner
		Existing	Synthesized	Existing	Synthesized	
Gemini-2.5-Pro	vs. OpenTheatre	80.00	55.00	3.79	4.35	Existing ✓
	vs. HoLLMwood	93.30	85.00	4.40	4.50	Existing ✓
	vs. BookWorld (SNP)	80.00	95.80	3.96	4.81	Synthesized ✓
	vs. BookWorld (FEN)	56.00	68.00	4.37	4.37	Synthesized ✓
DeepSeek-V3.2-Think	vs. OpenTheatre	86.70	68.30	4.11	4.17	Existing ✓
	vs. HoLLMwood	71.70	65.00	3.44	4.08	Existing ✓
	vs. BookWorld (SNP)	88.30	94.20	3.97	4.69	Synthesized ✓
	vs. BookWorld (FEN)	82.70	70.70	3.73	4.63	Existing ✓
DeepSeek-V3.2	vs. OpenTheatre	50.00	50.00	2.32	3.52	Tie
	vs. HoLLMwood	80.00	63.30	2.96	4.08	Existing ✓
	vs. BookWorld (SNP)	70.00	87.90	3.52	4.27	Synthesized ✓
	vs. BookWorld (FEN)	74.70	54.70	3.69	3.63	Existing ✓
Llama3.3-70B	vs. OpenTheatre	71.70	86.70	1.98	3.47	Synthesized ✓
	vs. HoLLMwood	21.70	80.00	1.59	3.49	Synthesized ✓
	vs. BookWorld (SNP)	41.70	78.30	2.05	3.00	Synthesized ✓
	vs. BookWorld (FEN)	28.00	50.70	1.74	2.36	Synthesized ✓
Qwen3-32B	vs. OpenTheatre	100.00	48.30	2.89	2.87	Existing ✓
	vs. HoLLMwood	53.30	85.00	2.40	3.18	Synthesized ✓
	vs. BookWorld (SNP)	71.70	68.30	3.14	3.17	Existing ✓
	vs. BookWorld (FEN)	81.30	34.70	3.17	2.61	Existing ✓

Table 7: Comprehensive breakdown of Win Rates and EVOSPARK’s absolute Average Likert Scores across *Existing* (prior knowledge) and *Synthesized* (no prior knowledge) scenarios. The results confirm that the framework’s performance does not fundamentally rely on LLM pre-training memorization.

UPDATE_RELATION_PROMPT
<p>You need to update your relationships with relevant characters based on the following information.</p> <p>## Character Description {role_profile}</p> <p>## Character Relationships {relation}</p> <p>## Character Current Status {status}</p> <p>## Conversation History {history_text}</p> <p>## Character Relationship Update Requirements Please strictly follow the requirements below and return the updated relationships in JSON format:</p> <ol style="list-style-type: none"> Decision Logic: Combine the “Character Current Status” and “Conversation History” to determine whether the relationships need to be updated. Only update if there are significant interactions or changes in the dynamic. Update Strategy: If changes are needed, please modify or supplement the original “detail” field content to reflect the new state. If the existing description is still accurate and sufficient, do not change it. You can only modify the values of the “relation” and “detail” fields in each sub-object. The value of the “relation” field must be a list of strings (List[str]), for example: [“new relationship1”, “new relationship2”]. The value of the “detail” field must be a string. Keep it concise and summarized (recommended 300-500 words maximum). Focus on core relationship points and recent changes; avoid lengthy historical reviews. Do not change any other keys (e.g., “ZhaoKai-en”, “LinWanYue-en”, etc.) or the overall JSON structure. Your response cannot contain any extra text or explanations besides the updated JSON. You cannot delete characters, even if there is no relationship. <p>Important: Ensure the total JSON length does not exceed the model’s output limit. Prioritize JSON completeness.</p>

Table 8: Prompt template for updating character relationship networks based on recent interactions.

UPDATE_PROFILE_PROMPT

You need to update the character's "profile" field based on the following information.

Original Character Description (JSON format)

```
{role_profile}
```

Character Current Status

```
{status}
```

Conversation History

```
{history_text}
```

Character Description Update Requirements

Please strictly follow the requirements below and **only return the updated "profile" field's string content**:

1. Analyze the "profile" field in the "Original Character Description".
2. Combine the "Character Current Status" and "Conversation History" to determine whether the "profile" field needs to be updated.
3. The "profile" field can only be changed when major changes related to the character occur in the story and have an impact on them.
4. If changes are needed, please modify or add to the original "profile" field content.
5. If no changes are needed, please **return the original "profile" field's string content**.
6. **Your response must be pure text string**, and can only contain the content of the "profile" field after updating (or without updating).
7. **Do not** include any JSON structure
8. **Do not** include any extra text or explanations (such as "Okay, here's the updated...").

For example, if the original "profile" is "a student", after updating it should become "a student who just finished an exam", you **can only return** "a student who just finished an exam".

Table 9: Prompt template for updating character profiles (long-term memory) based on narrative progression.

GENERATE_SPATIAL_POSITIONING_PROMPT
<p># Role Definition You are a professional stage director specializing in spatial blocking, skilled at arranging character positions to create dramatic tension and visual composition.</p> <p># Core Task Design reasonable spatial positioning for all participating characters (including NPCs) based on the current scene/event, character relationships, and dialogue history.</p> <p># Input Information ## Current Scene/Event: {scene_or_event} ## Participating Characters List: {roles_list} ## Current Location: Location Name: {location_name}; Location Description: {location_description} ## Recent Dialogue History: {recent_history} ## Current Dialogue Round: Round {current_round}</p> <p># Spatial Positioning Design Principles ## 1. Relative Position: Describe distances (face-to-face, side-by-side...), power dynamics, emotional relationships. ## 2. Embodied Posture: Standing, Sitting, Other postures. ## 3. Facing Direction: Face-to-face, Back turned, Sideways, Same direction. ## 4. Scene Interaction: Furniture interaction, Environmental interaction, Prop interaction.</p> <p># Output Format Requirements Must output a strict JSON object in the following format: <pre> {{ "spatial_layout": "One-sentence description of overall spatial composition (20-40 characters)", "positions": {{ "Character A Name": {{ "position": "Position in space (e.g., by window)", "posture": "Body posture (e.g., standing)", "facing": "Facing direction (e.g., facing Character B)", "relation_to_scene": "Relationship to scene elements" }}, ... }} }}</pre> </p> <p># Design Considerations 1. Dynamic Adjustment: Fine-tune based on dialogue development. 2. Relationship Hints: Use distance and facing to suggest relationships. 3. Dramatic Tension: Increase/reduce distance based on conflict/reconciliation. 4. Logical Consistency: Matching location characteristics. 5. Include All Characters: Ensure every participating character has clear position description.</p> <hr style="width: 10%; margin-left: 0;"/> <p>Now, design the spatial positioning for this round of dialogue based on the above information. Output the JSON object directly without any other explanations or markdown code block markers.</p>

Table 10: Prompt template for the Generative Mise-en-Scène module, instructing the LLM to act as a stage director.

FIND_NEW_ROLE_INFO_PROMPT
<p>You are a skilled screenwriter. Based on the following information, generate character information for {role_code}.</p> <p>### Records of Previous Scenes {history_scene_json}</p> <p>### Current Event {event}</p> <p>### Information of All Other Characters {role_agents}</p> <p>### Requirements</p> <ol style="list-style-type: none"> 1. Based on the records of previous scenes, generate character information. 2. The character information should include character profile, gender, identity, and relation. 3. Return in JSON format, formatted as follows: <pre> {{ "profile": "character profile", "gender": "character gender", "identity": "character identity", "relation": "character relationships", "name": "character name", "nickname": "character nickname" }}</pre> 4. Forbidden to output any explanations, comments, or Markdown markers (e.g., ““json, ““python).

Table 11: Prompt template for the Emergent Character Grounding Protocol (ECGP), used to instantiate new characters from narrative context.