DIRECTING THE UNSCRIPTED: ALT-MIRAGE FOR EMERGENT SOCIAL DECEPTION DRAMA

Yan Liu, Yang Zhao & Lidong Zhai
Institute of Information Engineering, CAS, China
{liuyan245, zhaoyang, zhailidong}@iie.ac.cn

Baoyang Chen
CAFA, Beijing, China
Chenbaoyang@cafa.edu.cn

Huamin Qu
HKUST, Hong Kong, China
huamin@cse.ust.hk

ABSTRACT

We introduce **Alt-Mirage**, a multi-agent AI Director framework for *unscripted* interactive storytelling in social-deduction games. Unlike prior director systems that primarily enforce predefined plot structures, Alt-Mirage targets the reality-show-like appeal of improvisation, deception, and emergent conflict by balancing autonomy with controllability. Our Director decomposes narrative orchestration into four specialized sub-agents: a *Planner* that converts high-level outlines into narrative goals, a *Director* that schedules interventions online, a *Verifier* that audits belief updates with a **log-grounded DEL-ToM** epistemic model to prevent hallucinated world-state changes while allowing strategic lying, and a *Critic* that quantifies narrative tension via **information-theoretic signals** (audience uncertainty, conflict entropy, and event surprisal) and casts intervention as an optimization objective. To steer dynamics without overriding agent reasoning, Alt-Mirage introduces **soft control** through a bounded *SuperEgo* modulation layer that reranks Ego-proposed action candidates under persona-consistency gating, preserving character autonomy while enabling interpretable behavioral steering; **hard control** complements this by reshaping environment conditions (e.g., blackout, lockdown, rule variants). We further provide a **lead-director dashboard** enabling mixed-initiative control by visualizing target vs. realized tension, agent belief states, and inconsistency alerts, and allowing high-level intervention commands. Experiments in an *Among Us*-style simulator with nine LLM-driven agents show that soft-control directives reliably steer outcomes and pacing (e.g., increasing faster-win rates and accelerating escalation) while maintaining emergent variability, demonstrating a practical route toward controllable, belief-consistent, tension-aware interactive drama.

1 INTRODUCTION

Traditional AI Directors in games and interactive narratives typically rely on predefined scripts executed by single-agent or multi-agent systems. In such frameworks, character agents behave like actors with limited deviation. While this supports narrative coherence and structural control, it often suppresses emergent dynamics, spontaneity, and the rich interactivity that make human experiences compelling.

By contrast, reality shows and unscripted social scenarios derive their appeal from improvisation, tension, and unexpected interactions. These spontaneous elements, rather than strict adherence to a script, form much of their emotional and dramatic core. Enabling more immersive and co-creative narrative experiences therefore requires rethinking the AI Director not as a script executor, but as a system that balances structural intent with emergent behavior.

To address this challenge, we propose Alt-Mirage, a multi-agent AI Director framework for social deduction-style games built on the mechanics of Among Us. In this environment, autonomous agents complete tasks, make inferences, and participate in voting rounds that simulate deception and

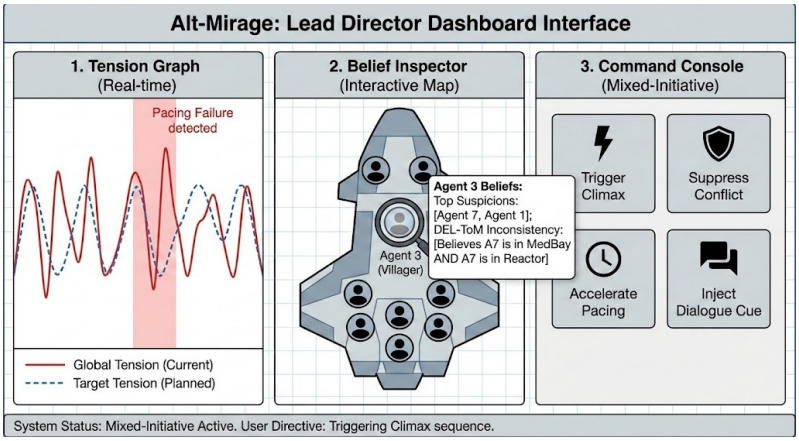

Figure 1: **Alt-Mirage lead-director dashboard** for mixed-initiative control, showing tension tracking, belief inspection, and high-level intervention commands.

collaboration. Guided by user-provided or system-generated story outlines, the Director monitors game dynamics in real time and intervenes when the unfolding trajectory deviates from the intended narrative arc.

Alt-Mirage supports two modes of intervention:

Hard interventions that directly manipulate the environment (e.g., injecting events, altering resource conditions);

Soft interventions that influence the internal states of agents through indirect adjustments, particularly via subconscious modulation.

Specifically, agent cognition is abstracted into two components: Ego, which encompasses perception, memory, and reasoning chains; and SuperEgo, which defines personality traits and behavioral tendencies. Rather than altering the agent's immediate decisions, the Director modifies the SuperEgo to steer future behaviors without overriding autonomy.

At the system level, the Alt-Mirage Director adopts a multi-agent architectural paradigm, decomposing the responsibilities of narrative orchestration into a set of specialized sub-agents:

A Planner, which translates abstract user intent into long-term narrative goals;

A Director, which adapts event pacing and coordination in response to environmental feedback;

A Verifier, which ensures the logical consistency of agent reasoning using a DEL-TOM-based model grounded in dynamic epistemic logic;

A Critic, which evaluates the dramatic tension of agent behaviors through a computational model of narrative tension.

Alt-Mirage represents, to our knowledge, the first systematic framework that integrates goal-driven story progression, logical validation of agent reasoning, and tension-aware behavior modulation within a social-deduction game setting. To make these capabilities actionable during live play, we provide a lead-director dashboard that exposes key narrative control signals—tension trajectories, agent belief states, and DEL-ToM inconsistency alerts—and enables mixed-initiative, high-level interventions such as pacing adjustments or dialogue cue injection. Figure 1 illustrates the interface and its core functions.

2 Related Work

Alt-Mirage combines three research threads: AI directors for interactive narrative, epistemic belief tracking for social reasoning, and tension-aware mixed-initiative control. Early systems such as *Façade* (Mateas & Stern, 2003) and *Versu* (Evans & Short, 2014) demonstrated drama management,

while recent LLM-based directors such as **CoDi** (Kim et al., 2025) and **IBSEN** (Han et al., 2024) steer autonomous character agents toward narrative goals. Separately, DEL and ToM-based approaches model belief updates, deception, and coordination in social interaction (Chetcuti-Sperandio & Bolander, 2020; Bolander & Dissing, 2020; Zhang et al., 2025), while tension-aware and mixed-initiative systems support engagement shaping and human-in-the-loop control (Hwang et al., 2022; Song et al., 2024; Kreminski & Wardrip-Fruin, 2022). Alt-Mirage integrates these directions for social-deduction gameplay with log-grounded belief verification and tension-aware intervention.

3 METHODOLOGY

3.1 ENVIRONMENT SETUP: SOCIAL DEDUCTION GAME SIMULATOR

To evaluate Alt-Mirage, we build an *Among Us*-style multi-agent social-deduction simulator in Unreal Engine 5 with nine autonomous agents and asymmetric information between **Villagers** and **Heretics**. Heretics know each other's identities, while Villagers infer them from partial observations and dialogue. Agents assume predefined roles (e.g., Prophet, Hunter) on a modular 3D map and act in discrete timesteps under partial observability.

Each episode cycles through **Task**, **Reporting**, **Communication**, and **Voting** phases: Villagers complete tasks to charge a shared Totem while Heretics sabotage or eliminate; discovered bodies trigger discussion and a subsequent vote. Villagers win by charging the Totem or eliminating all Heretics; Heretics win by reaching parity or surviving a time limit.

The environment supports global events (e.g., **Darkness**, **Door Lockdown**, **Task Suspension**) that alter perception, movement, or task progress. These events can be triggered by the AI Director or a human audience, enabling mixed-initiative modulation by changing the shared world state without prescribing individual actions.

3.2 ROLE AGENT COGNITIVE ARCHITECTURE

In the Alt-Mirage framework, each role-playing agent is modeled as a cognitively structured social actor capable of perceiving the environment, storing and recalling information, engaging in reasoning, and participating in interactive dialog. To support rich social behaviors and narrative-aligned decision-making, we adopt a layered cognitive architecture inspired by structured agent modeling (Park et al., 2023). each agent is internally composed of five key components: a *Perception Module*, a *Memory Module*, an *Ego* (reasoning engine), a *SuperEgo* (behavioral modulation layer) and a dynamic *Belief State*(Magee et al., 2024).

Perception Module Agents receive structured observations from the environment, bounded by spatial visibility constraints (e.g., occlusion, light level) and temporal state (e.g., time of round, global events). Observations include the presence and status of nearby players, interactive objects, environmental conditions (e.g., doors, tasks, bodies), and public role information. These inputs are encoded into semantic textual prompts for subsequent decision-making processes.

Memory Module Agents maintain a persistent, structured memory store segmented into interpretable subfields (e.g., known facts, received statements, suspicions, behavioral logs). This memory supports both short-term state tracking and long-term belief consolidation. It is selectively read and written during interaction and serves as a grounding context for reflection, reasoning, and language generation.

Ego: Deliberative Reasoning The agent's reasoning engine (*Ego*) integrates current observations, private memory, and role behavior templates to produce (i) updated beliefs and intentions and (ii) a small set of actionable candidates (e.g., move, accuse, probe, defend, ally, stay silent). Ego is invoked cyclically and is responsible for inferring strategies under uncertainty and generating coherent plans grounded strictly in accessible knowledge.

SuperEgo: Trait Vector and Structured Monologue State SuperEgo parameterizes *how* the agent acts on Ego's beliefs rather than *what* it believes. It modulates (i) **candidate selection** among Ego-proposed actions and (ii) **surface realization** of an action into dialogue style.

Crucially, the SuperEgo does not alter the agent's factual reasoning chain or access to information. Agents continue to reason strictly based on their visible observations and private memory. Instead, the SuperEgo modulates the *expression style* and *interaction strategy*—that is, how an agent chooses to communicate or act upon a given belief. For the same inference, different agents may select distinct rhetorical forms (e.g., direct accusation, indirect hinting, evidence-driven pressure, alliance-building, deflection, or strategic silence), thereby enabling interpretable behavioral diversity and narrative richness without additional fine-tuning.

During runtime, the AI Director may perform lightweight adjustments to the SuperEgo to achieve *soft control*. Trait vectors are subject to low-frequency, bounded modifications to avoid character drift, while the monologue state allows more frequent updates to adapt to current narrative tempo and tension. This layered architecture—combining stable personality with dynamic narrative alignment—makes the SuperEgo both a carrier of long-term social identity and an actionable interface for controllable, explainable style modulation.

3.3 Multi-Agent Narrative Controller

To support collaborative narrative control in the Alt-Mirage framework, we decompose the AI Director into four specialized sub-agents: the **Planner**, **Director**, **Verifier**, and **Critic**. Together, they form a modular architecture that enables structured story progression while preserving agent autonomy and emergent interaction(Kim et al., 2025). Each sub-agent operates with a distinct responsibility but maintains shared access to the story state.

3.3.1 Planner: Goal-Oriented Narrative Structuring

The Planner ingests user-provided or system-generated story outlines and transforms them into formalized narrative goals. These include event checkpoints, causal dependencies, and conflict phases. The output is a sequenced set of narrative targets that serve as an abstract control backbone against which real-time gameplay can be aligned. This goal structure enables the Director to evaluate deviations and coordinate interventions accordingly.

3.3.2 Director: Real-Time Monitoring and Intervention Scheduling

The Director acts as the temporal coordinator, constantly observing game state evolution and aligning it with the Planner's intended trajectory. By comparing actual progress with planned milestones, the Director identifies narrative drift, pacing anomalies, or unbalanced agent behaviors. Upon detecting such deviations, it schedules either soft or hard interventions (detailed in Section 3.4), such as adjusting environmental conditions or nudging agent personalities. The Director can also issue clarification prompts to enrich narrative salience without overt control.

3.3.3 Verifier: Log-Grounded Epistemic Consistency Tracking

The *Verifier* audits agents' belief updates to prevent hallucinated world-state changes while allowing strategic deception. It maintains an explicit DEL-ToM epistemic model (Chetcuti-Sperandio & Bolander, 2020; Bolander & Dissing, 2020) grounded in a global event log $\mathcal{L}_{\leq t}$ (available to the Director) that records both environment events and dialogue acts.

Epistemic state and DEL transition operator. At time t, the global epistemic model is a Kripke structure $\mathcal{M}_t = (W_t, \{R_i^t\}_{i\in\mathcal{A}}, V_t)$. We summarize agent i's monitored belief state as a set of (possibly higher-order) belief literals:

$$\beta_i^t = \{\, B_i p,\ B_i B_j p,\ B_i B_j(\neg p),\ \ldots \}. \tag{1}$$

A discrete interaction is represented by an event (action) model $\mathcal{E}_t = (E_t, \{R_i^{E_t}\}_{i\in\mathcal{A}}, \mathrm{pre}_t, \mathrm{post}_t)$, whose accessibility relations encode per-agent observability (e.g., dialogue heard by all vs. private utterances). The DEL product update $\mathcal{M}_{t+1} = \mathcal{M}_t \otimes \mathcal{E}_t$ is defined by:

$$W_{t+1} = \{(w,e) \in W_t \times E_t \mid \mathcal{M}_t, w \models \mathrm{pre}_t(e)\}, \tag{2}$$

$$((w,e),(w',e')) \in R_i^{t+1} \Leftrightarrow (w,w') \in R_i^t \wedge (e,e') \in R_i^{E_t}, \tag{3}$$

$$V_{t+1}((w,e)) = \mathrm{post}_t(e)(V_t(w)). \tag{4}$$

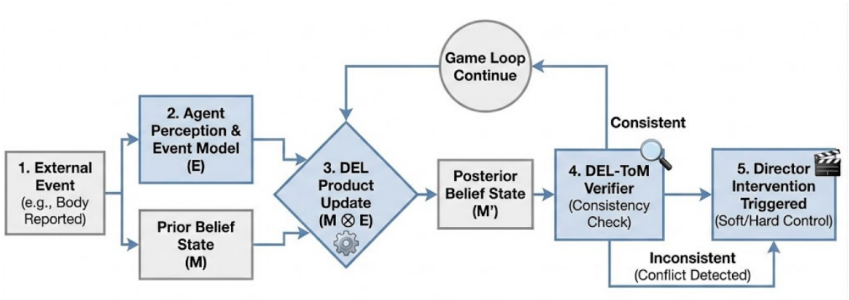

Figure 2: **DEL-ToM update cycle.** Logged environment and dialogue events induce a (constrained) DEL product update from $\mathcal{M}$ to $\mathcal{M}'$; detected anomalies can trigger Director interventions.

Deception as an epistemic action (belief shift without factual change). Since agents may lie, we model utterances as epistemic actions that typically do not modify environment facts. For an utterance $a = \mathsf{say}_i(p)$, we use two events $E = \{e_T, e_L\}$ (truthful vs. deceptive) with $\mathrm{pre}(e_T) = p$, $\mathrm{pre}(e_L) = \neg p$, and identity postconditions on environment propositions, $\mathrm{post}(e_T) = \mathrm{post}(e_L) = \mathrm{id}$. The speaker distinguishes whether they lied ($e_T \not\sim_i^E e_L$), while a listener may not ($e_T \sim_j^E e_L$ for $j \neq i$), capturing deception-induced ToM dynamics through the updated accessibility relations.

Log-grounded constraint for hallucination avoidance. To rule out hallucinated world updates, we ground the valuation of environment propositions P_{env} in the global log. Let $\Gamma_{t+1}(w) = 1$ denote that world w is consistent with $\mathcal{L}_{\leq t+1}$ (including recorded environment outcomes and dialogue occurrences/visibility). We apply a constrained transition:

$$\mathcal{M}_{t+1} = (\mathcal{M}_t \otimes \mathcal{E}_t)_{\Gamma_{t+1}}, \tag{5}$$

where $_\Gamma$ restricts the updated model to log-consistent world–event pairs. Intuitively, deception may change agents' beliefs about P_{env}, but it cannot directly change P_{env} without logged support.

Anomaly detection and intervention triggers. After each constrained update, the Verifier flags:

- **Epistemic contradictions**: $B_i p \wedge B_i(\neg p)$;
- **Inter-agent inconsistencies**: incompatible beliefs under a shared perceptual context (as implied by $R_i^{E_t}$ and the log);
- **Grounding violations (hallucinations)**: updates that require unlogged events or yield no log-consistent worlds (i.e., the restriction eliminates all plausible world–event pairs).

All anomalies are logged and forwarded to the *Critic* and *Director*, enabling soft/hard interventions (e.g., clarification cues, evidence reframing, pacing control). Figure 2 summarizes this log-grounded DEL-ToM update-and-intervention loop.

3.3.4 Critic: Information-Theoretic Tension Evaluation

The *Critic* monitors narrative tension as an online, quantitative signal to guide discrete Director interventions. Rather than treating tension as an abstract notion, we operationalize it via information-theoretic measures over (1) the audience's uncertainty about hidden game truth, (2) the degree of social conflict/disagreement, and (3) event-level surprise, consistent with drama management settings where engagement is controlled through monitoring and intervention Roberts & Isbell (2008); Mateas & Stern (2005).

Audience belief and uncertainty. Let z_t denote the hidden game truth (e.g., impostor identity, latent states, and evidence validity), which is known to the Director. The audience only observes the public log $\mathcal{L}_{\leq t}^{\text{pub}}$ (dialogue acts and public environment events). We define a spectator belief distribution

$$b_t(z) = P(z \mid \mathcal{L}_{\leq t}^{\text{pub}}), \tag{6}$$

and quantify suspense as posterior uncertainty

$$U_t = H(b_t) = -\sum_z b_t(z) \log b_t(z) \quad \text{(Shannon entropy Shannon (1948)).} \tag{7}$$

Conflict entropy. To capture social tension from disagreement, we compute an entropy over publicly expressed accusations/votes. Let $q_t(x)$ be the normalized distribution of targets x mentioned in accusation acts (or votes) within a rolling window.

$$H_{\text{conf}}(t) = -\sum_x q_t(x) \log q_t(x). \tag{8}$$

High H_{conf} indicates fragmented consensus and active conflict, a hallmark of social-deduction drama.

Event surprise (prediction error). Let ℓ_t be the next logged public event (including dialogue acts). Using a lightweight predictive model over the public log, we define surprisal as

$$S_t = -\log P(\ell_t \mid \mathcal{L}^{\text{pub}}_{<t}), \tag{9}$$

so that rare, unexpected turns contribute more to perceived dramatic intensity than frequent but predictable events.

Global tension score. We combine these terms (optionally augmented with affective arousal A_t from NRC-VAD Mohammad (2018)) into:

$$T_t^{\text{global}} = w_U U_t + w_C H_{\text{conf}}(t) + w_S \bar{S}_t + w_A A_t, \tag{10}$$

where $\bar{S}_t$ is a windowed average of S_t to reduce noise.

Tension as an optimization objective for interventions. Director interventions are discrete actions $a_t \in \mathcal{A}$ (e.g., modifying agent inclinations, injecting clues, adjusting environment conditions). We interpret Critic alerts as signals that the current trajectory under-serves information progression or tension pacing. Concretely, we use information gain

$$IG_t = H(b_{t-1}) - H(b_t) \tag{11}$$

as a proxy for audience information acquisition, and choose interventions to maximize expected information gain and surprise while tracking a desired tension arc $T_t^\star$:

$$a_t^\star = \arg\max_{a_t \in A} \mathbb{E}\left[\eta\, IG_{t+1} + \mu\, S_{t+1} - \lambda\big(T_{t+1} - T_{t+1}^\star\big)^2 - \text{cost}(a_t)\right]. \tag{12}$$

In practice, A_t is a small discrete candidate set of director actions (e.g., cue injection, pacing shift, environmental event trigger, or no-op), and we optimize Eq. (12) by enumeration with one-step heuristic scoring. Since the candidate set is small, the per-step selection cost is linear in $|A_t|$, which is tractable in our online setting. This formulation makes the Critic's notion of "tension" explicit, measurable, and directly tied to an intervention objective.

This formulation makes the Critic's notion of "tension" explicit, measurable, and directly tied to an intervention objective. In practice, A_t is a small discrete candidate set of director actions (e.g., cue injection, pacing shift, environmental event trigger, or no-op), and we optimize Eq. (12) by enumeration with one-step heuristic scoring. Since the candidate set is small, the per-step selection cost is linear in $|A_t|$, which is tractable in our online setting.

3.4 Control Mechanisms: Intervention via Agents

In the Alt-Mirage framework, narrative modulation is achieved through two primary intervention strategies: **Hard Control** and **Soft Control**. These interventions are orchestrated by the central **Director**, which synthesizes input from the **Critic** (tension evaluator), **Verifier** (belief consistency checker), and **Planner** (narrative structure manager) to determine when and how intervention is warranted. In addition to system-triggered modulation, Alt-Mirage also supports *human-initiated intervention*, enabling a mixed-initiative interactive storytelling framework.

3.4.1 HARD CONTROL: STRUCTURAL NARRATIVE RESHAPING

Hard control manipulates the environment directly to influence pacing, perception, and agent strategy without violating autonomy. The Director triggers hard control in response to the following contextual cues:

- The current narrative trajectory significantly deviates from the Planner-defined storyline (e.g., sustained peace during a conflict-intended phase);
- The Critic reports sustained low global tension and lack of critical events;
- Agent behavior is unbalanced or converging (e.g., key agents remain silent, or one faction dominates information flow).

Corresponding hard control actions include:

- **Environmental Event Injection**: trigger blackout, door lockdown, or task suspension to disrupt current strategies;
- **Task/Resource Modification**: alter Totem energy accumulation, pause regional tasks, or disable key zones;
- **Narrative Node Advancement**: forcibly initiate an emergency meeting or corpse discovery to shift phase;
- **Voting Rule Adjustment**: introduce double-elimination, anonymous voting, or alternate outcomes to enhance unpredictability.

These operations reshape the agents' observable world state and constraints, prompting adaptive behavior without interfering with their internal logic or deliberation.

Soft control as bounded SuperEgo modulation. Soft control operates exclusively on the agent's SuperEgo, leaving Ego's belief update and factual reasoning unchanged. Concretely, each agent i maintains a stable base trait vector $p_i \in \mathbb{R}^d$ and a Director-issued modulation $u_{i,t} \in \mathbb{R}^d$. The effective preference is

$$\theta_{i,t} = \mathrm{clip}(p_i + u_{i,t}, \theta_{\min}, \theta_{\max}), \qquad u_{i,t+1} \leftarrow (1-\kappa)u_{i,t}, \tag{13}$$

where clipping and exponential decay ensure interventions are bounded and reversible, preventing character drift.

At each timestep, Ego produces a small candidate set $C_{i,t} = \{c_k\}_{k=1}^{K}$ of feasible actions/utterances. SuperEgo then selects among candidates via transparent re-ranking:

$$c^\star = \arg\max_{c \in C_{i,t}} \Big[\mathrm{Consis}(c; p_i) + \lambda\, \theta_{i,t}^\top \phi(c)\Big], \tag{14}$$

where $\phi(c)$ encodes behavior features (e.g., aggressiveness, dominance, probing vs. defending) extracted by a lightweight rubric-based scorer, and $\mathrm{Consis}(c; p_i)$ gates persona violations. Candidates below a consistency threshold are rejected, and the agent falls back to base behavior when needed. Finally, the chosen action is realized into surface dialogue using the structured monologue state $m_{i,t}$ (stance, affect, intended persona), which the Director may update more frequently than traits.

This design makes soft interventions interpretable and controllable (via $\theta_{i,t}$ and $m_{i,t}$) while preserving character autonomy (stable p_i and consistency gating), and avoids intrusive token-level manipulation of the LLM.

3.4.2 HUMAN-INITIATED COMMANDS AND MIXED-INITIATIVE MODULATION

Beyond automated control, Alt-Mirage provides a lightweight human intervention interface that enables real-time director-style input during gameplay. Users may issue parameterized commands targeting either the environment or specific agents(Kreminski & Wardrip-Fruin, 2022), such as: Increase Agent X's aggressiveness or Trigger blackout in Place_02

The Director processes both system-driven and human-authored interventions through a unified control queue, weighing current tension state, storyline adherence, and cooldown constraints before execution. This ensures smooth coordination between autonomous narrative orchestration and external human agency, realizing Alt-Mirage's vision of collaborative, mixed-initiative storytelling.

Faction	Win (%)	Correct Acc.	Survive (s)	First Death (s)
No-Director				
Villager	32	0.366	592.91	135.12
Heretic	68	/	613.74	241.21
Soft: Heretic-fast-win				
Villager	25	0.24355	478.30	85.57
Heretic	75	/	595.39	210.84
Soft: Villager-win				
Villager	38	0.27422	547.26	182.34
Heretic	62	/	534.91	196.62

Table 1: **Controllability and pacing under soft-control directives.**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP AND METRICS

We evaluate Alt-Mirage in the *Among Us*-style simulator under two complementary lenses: **(i) controllability and pacing** (can the Director steer outcomes and temporal dynamics?) and **(ii) behavioral diversity** (does control collapse multi-agent variation?).

Implementation details. All role agents and director modules use GLM-5. Unless otherwise stated, decoding uses temperature 1.0, the API default top-p, and a maximum generation length of 10,000 tokens. At each timestep, Ego proposes $K{=}3$ candidate actions or utterances. SuperEgo re-ranks candidates under Eq. (14) using a consistency threshold $\tau{=}0.5$, bounded modulation $[\theta_{\min}, \theta_{\max}] = [-1, 1]$, and decay $\kappa{=}0.1$. For tension evaluation, we use fixed weights $(w_U, w_C, w_S, w_A) = (0.35, 0.25, 0.3, 0.1)$, with the optional affective arousal term disabled.

Conditions. For **controllability**, we compare **No-Director** against two **soft-control-only** directives: **Soft-Control (Heretic-fast-win)** and **Soft-Control (Villager-win)**. For **diversity**, we compare **No-Director**, a **Script Director** baseline with fixed non-adaptive interventions, and **Alt-Mirage** with online Planner–Director–Verifier–Critic orchestration and discrete hard/soft interventions. All conditions are run in the same environment with matched initialization distributions (map, role assignment policy, and randomization protocol).

Metrics. For controllability and pacing, we report faction **win rate**, Villager **correct-accusation rate**, **average survival duration**, and **first-death time**. To quantify the Verifier's epistemic grounding effect, we additionally report the per-step rate of three logged anomaly types defined in Section 3.3.3: epistemic contradictions, inter-agent inconsistencies, and grounding violations, where the latter denote unsupported belief updates. For diversity, we extract each agent's within-episode *structured action sequence* from the global log and discretize it into action tokens. We then compute **Self-BLEU-n** as a proxy for cross-agent homogenization, together with **action-type entropy** and **Distinct-n** as within-agent variety measures. Finally, using the Critic score in Eq. (10), we compare the realized tension trajectory T_t against the target arc $T_t^\star$ via per-episode mean absolute error (MAE) and peak-time deviation.

4.2 CONTROLLABILITY AND PACING VIA SOFT CONTROL

Soft control, grounding, and pacing. Table 1 shows that soft-control directives shift both outcomes and tempo relative to No-Director. *Heretic-fast-win* increases Heretic win rate (68% $\rightarrow$ 75%) while accelerating escalation (first death: 135.12s $\rightarrow$ 85.57s; Villager survival: 592.91s $\rightarrow$ 478.30s), consistent with stronger deception pressure and earlier lethal opportunities. Although *Villager-win* lowers correct-accusation accuracy relative to baseline, it still improves Villager win rate (32% $\rightarrow$ 38%), suggesting that soft control can favor task completion and coordination over high-variance eliminations. More broadly, soft control shifts behavioral tendencies distributionally rather than prescribing exact actions, preserving emergent variability and failure modes.

Setting	Contradictions ↓	Inconsistencies ↓	Grounding Violations ↓
Alt-Mirage w/o Verifier	0.074	0.129	0.082
Alt-Mirage	0.056	0.117	0.029

Table 2: Verifier ablation. We report per-step rates of the three anomaly types defined in Section 3.3.3. Grounding violations serve as our primary hallucination metric. Lower is better.

Setting	MAE ↓	Peak-Time Dev. ↓
No-Director	0.184	0.217
Script Director	0.121	0.143
Alt-Mirage	0.073	0.081

Table 3: Illustrative tension-trajectory deviation metrics relative to the target arc. Lower is better.

Setting	Self-BLEU2 ↓	Self-BLEU3 ↓	Self-BLEU4 ↓	Entropy ↑	Distinct-1 ↑	Distinct-2 ↑
No-Director	0.5897	0.4523	0.3614	1.1815	0.3271	0.6547
Script Director	0.6352	0.5501	0.4887	1.1898	0.4361	0.7743
Alt-Mirage	0.6109	0.5159	0.4521	1.1582	0.4047	0.7063

Table 4: **Behavior diversity and homogenization.** Self-BLEU-n measures cross-agent n-gram similarity over action-token sequences (i.e., less homogenization). Entropy and Distinct-n measure within-agent variety.

Verifier ablation and tension alignment. Table 2 compares Alt-Mirage with a *w/o Verifier* variant and shows the largest reduction in grounding violations, supporting the claim that the log-grounded DEL-ToM Verifier suppresses hallucinated world-state changes rather than merely shifting downstream outcomes. Table 3 further shows that Alt-Mirage more closely follows the target tension arc than No-Director and Script Director, yielding lower tracking error and smaller peak-time deviation, which supports the role of the Critic–Director loop in improving pacing alignment.

4.3 Behavior Diversity and Homogenization

Alt-Mirage preserves diversity under control. Table 4 shows that **Script Director** yields the highest Self-BLEU across n, indicating the strongest cross-agent homogenization, while **No-Director** is lowest. **Alt-Mirage** reduces long-horizon homogenization relative to scripted control, suggesting that online, signal-driven interventions can steer dynamics without collapsing agents into a single scripted pattern. Although scripted control can increase Distinct-n and entropy through broader phase and event coverage within each agent sequence, this does not contradict the Self-BLEU result, which measures cross-agent overlap. Overall, Alt-Mirage improves controllability over outcomes and pacing while better preserving multi-agent behavioral diversity than scripted control.

References

Thomas Bolander and Lasse Dissing. Implementing theory of mind on a robot using dynamic epistemic logic. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2020.

Matthew Chetcuti-Sperandio and Thomas Bolander. Del-based intention recognition in epistemic games. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, 2020.

Richard Evans and Emily Short. Versu: A simulationist storytelling system. In *IEEE Transactions on Computational Intelligence and AI in Games*, volume 6, pp. 113–130. IEEE, 2014.

Jiwoon Han et al. Ibsen: Director-actor agent collaboration for interactive drama generation. In *Proceedings of the ACL Conference*, 2024.

Jiho Hwang et al. Modeling tension in stories via commonsense reasoning and emotional word embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE)*, 2022.

Seungwoo Kim et al. Codi: A director-actor framework for goal-driven interactive story. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE)*, 2025.

Margaret Kreminski and Noah Wardrip-Fruin. Loose ends: A mixed-initiative story planning assistant for helping authors connect narrative threads. In *Foundations of Digital Games (FDG)*, 2022.

Liam Magee, Vanicka Arora, Gus Gollings, and Norma Lam-Saw. The drama machine: Simulating character development with LLM agents. *arXiv preprint arXiv:2408.01725*, 2024.

Michael Mateas and Andrew Stern. Façade: An experiment in building a fully-realized interactive drama. In *Game Developers Conference*, 2003.

Michael Mateas and Andrew Stern. Structuring content in the façade interactive drama architecture. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE)*, pp. 93–98, 2005. doi: 10.1609/aiide.v1i1.18722. URL https://cs.uky.edu/~sgware/reading/papers/mateas2005structuring.pdf.

Saif Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 174–184. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1017. URL https://aclanthology.org/P18-1017/.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proc. of the 36th ACM Symposium on User Interface Software and Technology (UIST)*, 2023.

David L. Roberts and Charles L. Isbell. A survey and qualitative analysis of recent advances in drama management. 2008. URL https://faculty.cc.gatech.edu/~isbell/papers/itssa08-survey.pdf.

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948. URL https://www.mpi.nl/publications/item2383162/mathematical-theory-communication.

Youngrok Song et al. A conflict-embedded narrative generation using commonsense reasoning. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

Xinyu Zhang et al. Multimind: Multimodal theory of mind modeling for social deduction games. *arXiv preprint arXiv:2502.12345*, 2025.