From Words to Worlds: Transforming One-line Prompts into Multi-modal Digital Stories with LLM Agents

Danrui Li*
Computer Science, Rutgers, the State University of New Jersey
USA
danrui.li@rutgers.edu

Samuel S. Sohn*
Computer Science, Rutgers, the State University of New Jersey
USA
samuel.sohn@rutgers.edu

Sen Zhang*
Computer Science, Rutgers, the State University of New Jersey
USA
sen.z@rutgers.edu

Che-Jui Chang
Computer Science, Rutgers, the State University of New Jersey
USA
chejui.chang@rutgers.edu

Mubbasir Kapadia
Roblox
USA
mkapadia@roblox.com

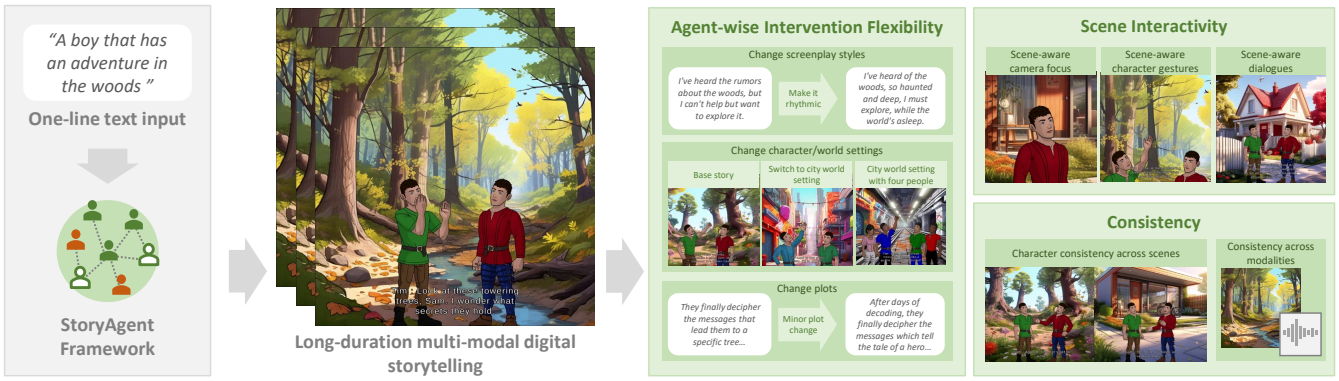

Figure 1: StoryAgent is a digital storytelling generation framework that integrates communicative Large Language Model agents with state-of-the-art generative models and tools. Taking one-line text instruction as input, it produces digital storytelling content with scene interactivity, long-duration consistency, and intervention flexibility.

Abstract

Digital storytelling, essential in entertainment, education, and marketing, faces challenges in generation efficiency. The StoryAgent framework, introduced in this paper, utilizes Large Language Models and generative tools to automate and refine digital storytelling. Employing a top-down story drafting and bottom-up asset generation approach, StoryAgent tackles key issues such as manual intervention, interactive scene orchestration, and narrative consistency. This framework enables efficient production of interactive and consistent digital storytellings across multiple modalities, democratizing content creation and enhancing engagement.

*These authors contributed equally to this research.

MIG '24, November 21–23, 2024, Arlington, VA, USA

ACM ISBN 979-8-4007-1090-2/24/11
https://doi.org/10.1145/3677388.3696321

CCS Concepts

• **Computing methodologies** → **Procedural animation**; *Intelligent agents*; Multi-agent planning.

Keywords

Digital storytelling, Large Language Models, Communicative Agents

ACM Reference Format:
Danrui Li, Samuel S. Sohn, Sen Zhang, Che-Jui Chang, and Mubbasir Kapadia. 2024. From Words to Worlds: Transforming One-line Prompts into Multi-modal Digital Stories with LLM Agents. In *The 17th ACM SIGGRAPH Conference on Motion, Interaction, and Games (MIG '24), November 21–23, 2024, Arlington, VA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3677388.3696321

1 Introduction

Digital storytelling has emerged as a powerful medium across various domains, including entertainment, education, and marketing [De Jager et al. 2017; Lambert 2013; Wu and Chen 2020] due to its ability to combine multimedia elements such as text, images, audio, and video to create immersive and interactive digital narratives. Its versatile applications, ranging from interactive narratives

in video games to immersive experiences in virtual reality, make it an indispensable tool for conveying information and eliciting emotions in today's digital age. Traditionally the production processes of digital storytelling narratives are often time-consuming and resource-intensive with requirements of professional skills, limiting the speed and scale of content output even for early-stage prototyping.

Hence, there arises a critical need to streamline and automate the production pipeline to meet the growing demands for engaging and dynamic storytelling content. Recent advancements in text-based generative models[Betker et al. 2023; Brooks et al. 2024; Esser et al. 2024; Ho et al. 2022; Kreuk et al. 2023; Li et al. 2024; Liu et al. 2023b], which synthesize text, image, sound, and motion from user inputs, have facilitated a multimodal hands-free digital storytelling asset production. Meanwhile, several previous works [Cavazza et al. 2007; Kapadia et al. 2016a; Louarn et al. 2018; Marti et al. 2018] developed authoring tools for part of the production pipeline such as storyboard making, staging, and script writing. While it seems promising to combine the works above by integrating generative models into authoring tools, significant challenges persist as below.

One notable challenge is the need for flexible intervention, as human creators often require the ability to modify initial generation results according to their preferences (**Challenge 1**). Text-based generative models[Brooks et al. 2024; Ho et al. 2022] are proficient at creating high-quality short clips but offer limited fine-grained control over outcomes, such as modifying characters while maintaining the same storyline. Conversely, procedural methods [Kapadia et al. 2016a,b] enable fine-grained control but typically require a specialized interaction interface with the framework, which often lacks a universal and convenient approach for human intervention. Moreover, orchestrating the interactions between characters and scene objects remains a difficult task (**Challenge 2**), yet indispensable to improving the visual fidelity and elevating the digital storytelling experience [Yi et al. 2024]. Finally, long-term consistency is needed for audience engagement (**Challenge 3**). For instance, character appearances and voice tones should remain consistent with the narrative context throughout the story. In addition, consistency should also include the synchronization of the textual plots and downstream modalities, including audio, speech, and visuals [Chang et al. 2023, 2022a,b; Cummings and Bailenson 2016; Kybartas and Bidarra 2017; Salselas and Penha 2019]. Despite advancements in diffusion-based animation generation [Feng et al. 2023; Guo et al. 2023; Liew et al. 2023], existing methods struggle to ensure long-term consistency or require additional inputs like reference videos or skeletons.

We propose a novel StoryAgent framework (see Fig. 1). It integrates Large Language Model (LLM) agents [Li et al. 2023; Wu et al. 2023] with generative models and tools, primarily served as a rapid prototyping and ideation tool for digital storytelling designers. Our framework operates by initially drafting the story by a top-down approach, using communicative LLM agents to decompose text instructions into a hierarchical textual representation of the digital storytelling content, where the leaf nodes are descriptions of a single modality for a snippet of the timeline. Subsequently, it employs generative models and tools in a bottom-up fashion to create and assemble the corresponding assets from text descriptions.

The framework effectively addresses several key challenges. Initially, its textual representation and generation pipeline provide a means for human developers to exert fine-grained control through simple natural language instructions (**Challenge 1**). By leveraging the reasoning capabilities of Large Language Models, the framework is adept at interpreting instructions to selectively modify leaf nodes in its hierarchical structure without affecting other components, thus preserving the overall content integrity. Furthermore, the framework captures semantic and spatial information from generated images and integrates this data into its textual hierarchy, which facilitates enhanced scene interactivity for downstream components (**Challenge 2**). Additionally, the top-down hierarchical textual representations maintain long-term consistency across various outputs (**Challenge 3**). A notable implementation is the consistent use of costume asset IDs to maintain a character's appearance across different time frames within the video. By reusing the same asset, the framework guarantees visual uniformity and continuity throughout all scenes, ensuring a cohesive viewing experience. This capability allows for the creation of contextually appropriate content at scale.

Finally, by leveraging text as the intermediate product to decouple story drafting and asset generation, our framework facilitates a plug-and-play structure. It not only allows for unprecedented coverage of modalities (see Tab. 1), but also easy integration of the latest generative models, which ensures that our framework performance can continuously benefit from ongoing research developments.

2 Background

2.1 Digital storytelling as authoring tools

Digital storytelling has traditionally focused on providing authoring tools for human developers [Cavazza et al. 2007; Kapadia et al. 2016a; Louarn et al. 2018; Marti et al. 2018], typically adopting a procedural generation approach rather than end-to-end solutions like [Brooks et al. 2024]. Usually, they are made for a specific stage in the production pipeline, where domain knowledge and specific software skills are required for users. Our StoryAgent framework covers wide range of stages and enables flexible human intervention through natural language instructions, thus creating integrated and accessible authoring experience.

2.2 Enable scene interactivity

Recent studies have advanced the dynamics of character interactions within 3D environments, significantly aided by the explicit spatial representations of 3D objects [Chang et al. 2024a,b, 2023; Hassan et al. 2023; Starke et al. 2019; Xu et al. 2023; Yi et al. 2024; Zhang et al. 2021]. In contrast, for 2D art styles, despite diffusion-based models delivering superior visual quality, the lack of inherent spatial scene information within 2D images poses challenges for implementing interactivity.

Nevertheless, existing research in image understanding offers many tools for deducing spatial data from images, including techniques like segmentation [Xie et al. 2021] and depth estimation [Bhat et al. 2023a]. These methodologies can underpin a procedural pipeline designed to facilitate interactivity in 2D contexts. This paper represents an initial effort to enable such scene interactivity for

Table 1: Digital storytelling components covered in prior works and ours. ✓= generative models. △ = retrieval methods.

		World		Character		Audio			Others
	Plot	Semantic	Visual	Appearance	Animation	Music	Speech	SFX	Cinematography
[Ammanabrolu et al. 2020]	✓								
[Balint and Bidarra 2023]		✓							
[Hartsook et al. 2011]		✓	△						
[Louarn et al. 2018]									✓
[Zhang et al. 2021]			△	△	✓+ △				
[Kumaran et al. 2023]	✓		△	△	△				
[Liu et al. 2023b]						✓	✓+ △	✓	
Ours	✓	✓	✓	✓+ △	△	✓	✓	✓+ △	✓

2D art styles, allowing characters to engage with the background environment.

2.3 Consistency in digital storytelling

Following the definition of [Kybartas and Bidarra 2017], the components in digital storytelling can be categorized into plots (textual contents such as story arc and events) and space (world settings, characters, scene props, etc.). Prior works have explored various approaches to ensure coherence and consistency within the plot generation, such as event-driven [Kapadia et al. 2016b], persona-driven [Xu et al. 2020; Zhang et al. 2022], state-space planning[Miller et al. 2019], and top-down decomposition [Kim et al. 2023]. However, for the consistent joint generation between plot and space components, previous studies have been limited, generally focusing on one or two areas such as crowd motion [Chen et al. 2020], cinematography [Louarn et al. 2018], world settings [Hartsook et al. 2011; Merino et al. 2023] etc. (see Tab. 1). While some approaches have aimed to address the alignment of all visual components simultaneously, typically using latent diffusion text-to-image models [Liu et al. 2024; Maharana et al. 2022; Shen and Elhoseiny 2023], their effectiveness is still constrained by the time length, rendering them unsuitable for long-term digital storytelling scenarios. In this work, we achieve a wide range of consistency between various components from a procedural approach.

2.4 Generative digital storytelling with LLM agents

Prior works have explored LLM-assisted generation frameworks covering several components such as text[Ammanabrolu et al. 2020; Balint and Bidarra 2023], audio[Liu et al. 2023b], and visual[Kumaran et al. 2023], where LLMs have shown their capabilities in assisting single-modal asset generation in the following aspects:

Firstly, they can extract information from natural language descriptions and convert it into formatted parameters [Kumaran et al. 2023; Qing et al. 2023]. Leveraging their world knowledge, LLMs can also break down complicated concepts into several simpler components [Liu et al. 2023b], which lowers the difficulties for downstream generative models. In addition, the organization of components can be stored and updated in explicit formats [Torre et al. 2024]. Finally, with reasoning capabilities [Wei et al. 2023; Yao et al. 2023b], LLMs can plan for the generation tasks with predefined toolsets and real-time feedback. But prompt-based narrative scene generation tools like [Kumaran et al. 2023] are limited in the ability to generate multiple coherent and compelling narrative scenes in sequence.

LLM agent systems are widely used in complex generation tasks such as game world narratives [Park et al. 2023] and computer programs [Qian et al. 2023]. In these cases, the consistencies are achieved by predefined hierarchical memories and reflection procedures. In this paper, we apply an LLM agent system to ensure consistency in both temporal and modal dimensions.

3 Method

The entire pipeline (see Fig. 2) begins with a text instruction, then develops an intricate storyline by an LLM-agent-based story cluster (see 3.1) with scene understanding capabilities (see 3.2). Previous research ([Wei et al. 2024], [Yao et al. 2023a], [Liu et al. 2023a]) showed that task decomposition can boost the agent system's problem-solving, creative writing, and task planning. Our story cluster deconstructs the digital storytelling task into multiple subtasks, each targeting a specific modality. For each modality, the story cluster specifies all output asset files of the story through the textual descriptions. Based on these descriptions, generative models and tools are organized into asset generation teams (see 3.3) to create tangible assets of the story. Finally, approaches for human interactions are introduced in 3.4.

3.1 Story cluster

The story cluster drafts the storyline through a network of LLM agent teams. Inspired by the pipeline workflow in the film industry, the framework starts with a story arc generation team. Then, the story arc is distributed to numerous downstream teams to handle different narrative components: characters, settings, story beats, setting affordances, story scenes, and the screenplay.

The intermediate products that are circulated between teams are always in text. Although the cluster is built upon AutoGen[Wu et al. 2023], the circulation follows a predefined procedure (see Fig. 2) to ensure generation stability.

3.1.1 Story Arc. The network begins with a story arc team, which composes a blueprint of the narrative in text, instructing the entire

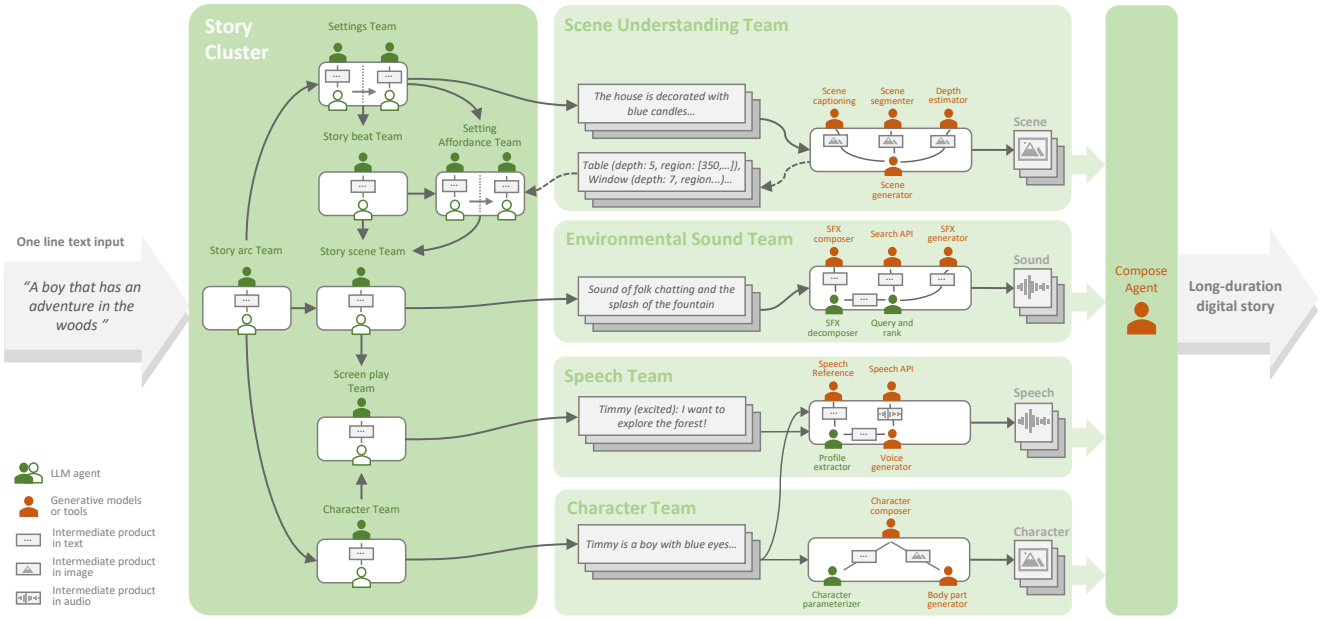

Figure 2: The framework of StoryAgent. Beginning with a text instruction, the framework builds the story with task decomposition, specifying all asset files for each modality in the textual description. Then generative models and tools are organized to create and compose tangible assets of the story.

production at the highest level. The story arc is produced in a JSON format as below:

```
"story_arc": {
    "exposition": "A curious boy named Tim and his
    adventurous friend, Sam, decide to explore the woods
    near their homes, despite the rumors of it being haunted
    .",
    "rising_action": "As they venture deeper into the woods,
     they find a series of cryptic messages carved into the
    trees. They decide to decipher these messages, believing
     they might lead to a hidden treasure.",
    "climax": "They finally decipher the messages which lead
     them to a specific tree. They find an old, weathered
    box buried at the base of the tree.",
    "falling_action": "They open the box to find a map of
    their town from decades ago, with a path marked leading
    back to their homes.",
    "resolution": "They follow the path on the map and
    realize that the 'treasure' was the journey and the
    memories they made. They return home, their friendship
    strengthened by the adventure."
}
```

3.1.2 Downstream Components. Receiving the story arc and upstream contents as input, each team incrementally optimizes their results through single-stage or multi-stage dialogues (see Fig.3). All multi-stage configurations are manually designed to further decompose the generation process, as some agent teams require complex reflections that the current LLM backbone cannot handle. In each stage, the dialogue takes place between two LLM agents. Inspired by the role-playing method in [Li et al. 2023], each agent is initialized with a system prompt, where dialogue history and upstream content are attached to an agent-specific template.

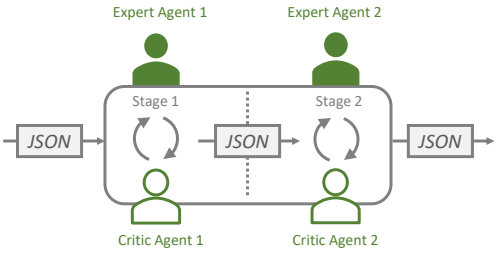

Figure 3: Structure of a two-stage LLM agent team. It takes upstream JSON as inputs and uses two expert-critic LLM agent pairs to process. Finally, another JSON string will be generated for downstream teams.

Among the two, one "expert agent" is responsible for creating component specifications based on predefined requirements and upstream inputs. The other "critic agent", evaluates and scores them against predefined criteria. The roles of the agents, their constraints, and the output formats are shared between the expert and the critic within individual team stages but differ across different agent teams

and stages. For comprehensive prompts for all agents, please refer to the appendix.

Starting from the expert agent, the two respond to each others' outputs in a round-robin way, where the circulation ends when the critic is satisfied with the expert's outcome. Utilizing GPT-4 as the LLM agent backbone, this expert-critic design facilitates a robust capacity for self-correction, which ensures consistency between the upstream commands and downstream outcomes. To illustrate, consider the following example where the critic agent helps the expert agent refine a character profile:

Admin

(provide a story arc in JSON format)

Expert in Character

{"name": "Village elders"...}

Critic in Character

...The character name "Village elders" is plural and should be singular...The expert should break down "Village elders" into individual characters with singular names. For example, they could create characters for a few prominent elders and villagers.

Expert in Character

{"name": "Elder John"...},{"name": "Elder Mary"...}...

Critic in Character

(pass the check)

An example of generated character profile is shown below:

```
"story_characters": [
    {
        "name": "Tim",
        "gender": "Male",
        "age": 12,
        "personality": "Tim is a curious boy with a cautious
            nature. He is always eager to learn new things
            but tends to think twice before diving into
            unknown situations.",
        "beliefs": "Tim has a fascination with the
            supernatural, often indulging in stories about
            ghosts and haunted places.",
        "motivations": "His curiosity and the desire for
            adventure are the main driving forces behind his
             actions. He is always looking for something new
             and exciting to explore.",
        "development": "Throughout the story, Tim starts off
            as a cautious boy but his adventurous side
            slowly comes to light as he delves deeper into
            the woods.",
        "physical_description": {
            "tunic": "He wears a green cotton tunic, soft to
                the touch and slightly faded from numerous
                washes.",
            "pants": "His brown corduroy pants are sturdy and
                 well-worn, perfect for an adventure in the
                woods.",
            "boots": "His black leather boots are scuffed
                from use, but they are waterproof and
                comfortable for long walks."
        }
    },
    ...
]
```

3.2 Image-based scene understanding and interaction

Simply putting the characters in the foreground and the story setting images in the background could result in unreasonable arrangements like a character standing on the water or up in the air. To enhance the fidelity of each scene and give audiences a better immersion in the story, the pipeline conducts several measures: first fusing image semantic segmentation and depth estimation explicitly and implicitly as the visuospatial information and then letting setting affordance agents to enable foreground characters to interact with background images, which would otherwise be disjoint. The visuospatial information could also benefit cinematography by providing the object's estimated position and distance to the camera.

3.2.1 Scene understanding. The scene understanding process is structured in three key stages, designed to equip setting agents with comprehensive knowledge of the narrative environment, enabling them to provide relevant affordances for dynamic story interaction.

Initially, story setting agents construct a hierarchical graph of all narrative settings. Each node within this graph represents a distinct setting and includes essential details like the setting's name, visual prompts, and its relationship with other settings (parent and children). From these visual prompts, background images are generated using [Li et al. 2024], forming the visual foundation of the story's environments.

The generated images undergo semantic segmentation to identify and categorize objects within each setting. This segmentation ([Xie et al. 2021]), paired with depth estimation ([Bhat et al. 2023b]), provides a three-dimensional understanding of each scene. Objects are encapsulated within bounding boxes, which highlight their spatial position and median depth, aiding in the precise placement within the narrative space. This crucial spatial data informs how characters and cameras will interact with these objects, ensuring accurate and realistic scene compositions(see 3.2.2).

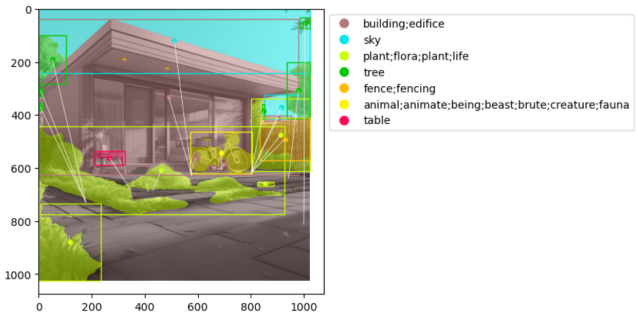

Figure 4: Semantic scene understanding example

In the final stage, setting affordance agents utilize the detailed object and spatial data to create a rich layer of interaction possibilities through affordances. These affordances are meticulously documented with object relations, narrative relevance, and evidence of the object's existence in the image. This structural information is then passed to story scene agents, who use it to script interactions and narrative events, ensuring that characters can interact naturally with their environment.

3.2.2 Character and camera interaction. Our pipeline processes and utilizes object location information to enhance storytelling through precise cinematography. Once this data is captured, screenplay agents integrate it to orchestrate scene dynamics effectively. This involves using the object centers as focal points in animation and cinematography—characters interact with key objects via targeted movements, and cameras adjust focus and framing based on the object's position and estimated distance. This approach allows for the strategic selection of shot types (close, medium, or wide) to best capture the narrative moment.

3.3 Asset generation teams

Each agent generation team represents a hybrid of LLM agents and generative models. Generally, the LLM agents here serve as the bridge between the upstream textual descriptions and the downstream generative models and tools. Specifically, LLM agents aim to convert specifications articulated in natural language into model parameters and reflect on feedback from generative models and tools.

In our study, we employ the rigging framework, CC2D[Simpleton 2021], to construct and animate human avatars. To generate image assets for different parts of the character, such as pants and clothing, we input appearance descriptions into a text-to-image generative model [Li et al. 2024]. These assets are produced using a supplemental prompt—such as "detailed, cartoon, 8k"—to enhance the initial description. Since diffusion-based generative models are hard to generate meaningful 2D textures directly, the generative procedure goes through a composite and decompose process. First We composite the 2D texture of the tunic and pants into a meaningful shape of the asset (see Fig. 5), then generate assets using its assembled mask. After the generation is done, we decompose the whole assets back with their texture mask and then integrate them within the CC2D framework (see Fig. 6). Textures are loaded during runtime (see Fig. 7).

Figure 5: Character pants asset examples

For background image creation, we utilize the same generative model to ensure style consistency. We use [Stewart [n. d.]] to lift the token limitation and generate backgrounds by visual prompts

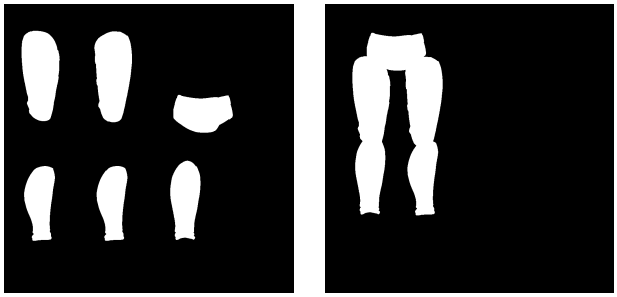

Figure 6: Character parts mask. Left is the original texture format and right is the assembled mask for better generation quality

Figure 7: Pants asset in the runtime

provided by the setting configuration agents. Furthermore, these images undergo further processing in a scene understanding cluster (see 3.2) to enable interactivity. This additional step allows for dynamic interaction within the generated environments (detailed in 3.2).

In speech generation, a character description simplified by LLMs is firstly fed to AudioGen[Kreuk et al. 2023] to create a voice fingerprint, containing the voice feature of that character. Then it is smoothed by XTTS [CoquiAI 2021], producing a reference speech audio. Next, the reference audio, speech in text, and the emotion state in text are fed to ElevenLabs API [ElevenLabs 2024] to create the final speech audio.

Inspired from [Liu et al. 2023b], the sound description is firstly decomposed to components by LLMs, where different decomposition methods are tailored for sound effect types (looping or one-time). Then, for each component, an LLM agent is applied to search corresponding assets on FreeSound [of Universitat Pompeu Fabra [n. d.]], where the assets are evaluated by feeding their descriptions to the agent. The search is repeatedly triggered with keywords that are optimized by the agent using previous search results. If there are no suitable assets, AudioGen[Kreuk et al. 2023] is called to create one. All music clips are created by MusicGen [Kreuk et al. 2023], a text-to-music generative model.

Consistency is optimized by communication between LLM agents and other generative models and tools. For instance, in sound retrieval, after the initial description is broken down into simpler components, each component's description is transformed into a query string, which is iteratively refined based on the results of previous searches. This method underscores the dynamic use of LLMs to enhance the consistency between the text description and the retrieved assets.

Admin
The sound of townsfolk chatting, children playing, and the gentle splash of the fountain water.

Decomposition Agent
{ "name": "townsfolk chatting", "desc": "the sound of townsfolk chatting in a town square", "volume": -25 }, { "name": "children playing", "desc": "the sound of children playing in a town square", "volume": -20 }, { "name": "fountain water", "desc": "the gentle splash of fountain water in a town square", "volume": -30 }

Decomposition Agent
@Query Agent the gentle splash of fountain water in a town square

Query Agent
@Search API water +fountain +splash +gentle -ocean -river

Search API
(a list of sound with descriptions)

Query Agent
water +fountain +gentle +town -ocean -river -greenway

Search API
(a list of sound with descriptions)

Query Agent
fountain +gentle +town

Search API
(a list of sound with descriptions)

Query Agent
(The index of the best sound asset)

3.4 Intervention

While our framework facilitates a hands-free generation process, it also offers flexible intervention approaches for human creators. The framework allows for the following intervention approach:

- Full regeneration from the beginning: by feeding prior generation results and extra instructions to the framework inputs, the framework updates all contents with maximum consistency ensured. We apply this approach in Section 4.1. Please refer to the instruction prompt in the appendix.
- Intermediate the process: by fixing upstream results and altering agent system prompt, the framework updates all downstream contents with finer control over selected modalities.
- Replacement: by replacing generated results with human-crafted ones, human creators get their maximum control over the outcome. However, replacing downstream assets may cause inconsistency issues since the framework cannot reflect on them.

4 Results

In this section, we present the qualitative results of our pipeline, highlighting the benefits of our text-based pipeline. Based on an example story (see Fig. 8), we first demonstrate how the reasoning capabilities of StoryAgent facilitate consistency between plots and downstream components. Then we illustrate how the coordination of various agents and generative tools ensures scene interactivity in 2D art styles. Finally, we show its capability to adapt human intervention during the generation process, thus providing story alternatives for human creators.

4.1 Agent-wise human intervention

Our framework's agentic approach to generating digital narratives allows for flexible human intervention. Below, we present various intervention outcomes based on the same story, demonstrating how flexibility is achieved by instructing different agents at distinct stages within the pipeline.

4.1.1 Low-level intervention. In the following example, we adjust the screenplay writing style of the story, which represents a fine-grained low-level intervention case. By directly instructing the downstream agent team (Screenplay team) to compose the dialogue in rhyming couplets, we ensure that the upstream content of the base story (story's outline) remains unchanged.

Base Storyscript:

NARRATOR: In a quiet neighborhood, we meet Tim, a curious boy with a thirst for adventure.
TIM: I've heard the rumors about the woods, but I can't help but want to explore it.
NARRATOR: Despite the rumors, Tim's curiosity is not deterred.
TIM: No matter how far I go, I know this house will always be my safe haven.

Screenplay-Intervened Story Script:

NARRATOR: In a cozy home, lived a boy named Tim, his spirit was adventurous, his curiosity never dim.
TIM: I've heard of the woods, so haunted and deep, I must explore, while the world's asleep.
NARRATOR: With a heart full of courage, and a mind full of wonder, into the woods, Tim decided to wander.
TIM: No matter how far, into the woods I delve, this house will always be my safe haven, my safe shelf.

4.1.2 High-level intervention. We demonstrate two interventions with higher-level instructions, which impose larger-scale impacts on the narrative generation process (i.e., downstream agents): one where the characters and plot are fixed, but the setting is changed (**S**-intervention), and another where the plot is fixed, but the characters and setting are changed (**CS**-intervention).

(a) A curious boy named Tim and his adventurous friend, Sam, decide to explore the woods near their homes, despite the rumors of it being haunted.

(b) As they venture deeper into the woods, they find cryptic messages carved into the trees. They decide to decipher these, believing they might lead to a treasure.

(c) They finally decipher the messages that lead them to a specific tree. They find an old, weathered box buried at the base of the tree.

(d) They open the box to find a map of their town from decades ago, with a path marked leading back to their homes.

(e) They follow the path on the map and realize that the treasure was the journey and the memories they made. They return to Sam's home with friendship strengthened.

Figure 8: Screenshots of a story generation, with one frame selected from each stage of a five-stage storytelling arc.

The interventions were achieved by using the story arc of the base story as input in the framework with an additional instruction to preserve some elements of the story while changing others. Following the pipeline, we firstly observe story arc changes. The expositions (starting part) of the story arcs are shown in Fig. 9. While the characters and settings are modified as the response to user input, the story motivation remains unchanged: characters are about to explore some places with rumours. Besides, there exists diversities when responding to the same "city" settings: while the **S**-intervention depicts the "rumour" with "confusing streets and alleys", the **CS**-intervention presents "subway system".

Base Story:
A curious boy named Tim and his adventurous friend, Sam, decide to explore the woods near their homes, despite the rumors of it being haunted.

S-Intervention:
Tim, a curious boy, and his adventurous friend, Sam, decide to explore the vast cityscape around their apartment complex, despite the rumors of it being a maze of confusing streets and alleys.

SC-Intervention:
In a bustling city, a curious boy named Tim, his adventurous friend Sam, and their classmates, the shy Lily and the skeptic Mike, decide to explore the labyrinthine subway system, despite the rumors of it being haunted.

Figure 9: The expositions of story arcs in the original base story (top), a variant on world settings (middle), and a variant where both world settings and character settings are intervened (bottom). In the latter two, the responses to the intervention instructions are underscored.

Then in Figure 10 we can see the changes in the story arc are propagated to downstream agent teams, resulting in the variations of speech and graphics of the digital storytelling. Between both interventions, we observe that the cryptic carvings in the forest of the base story are mirrored in the cryptic graffiti in the **S**-intervention and the cryptic etchings in the **CS**-intervention (Event 1). This demonstrates that our framework can make precise interventions, varying settings while preserving key story concepts.

In Events 2 and 3, we see that each story's plot leads the characters on a circuitous adventure back to where they started and they share the same moral of appreciating the adventure. In the **CS**-intervention, the dialogue that was originally distributed between two characters has been distributed among four characters, meaning that the characters have not only been added but are being fully utilized.

4.2 Scene interactivity

StoryAgent can integrate detailed scene information into digital storytelling, creating scene interactivity. We focus on two common practices in digital storytelling for children, effectively engaging young audiences by emphasizing narrative elements.

Firstly, the interactions take place at the semantic level. When the story unfolds, characters will refer to the scene objects in their dialogue, maintaining the story's coherence at the same time. In Fig.11, the characters are talking about their childhood memories, while raising a fence in the background as a reference.

The interactivity is also enhanced by the synchronous interplay between the dialogue, character gestures, and cinematography. As characters discuss an object within the scene, not only does the camera shift focus to and zoom in on the object for detailed visualization, but the characters also direct their gazes and gestures toward it. This coordination makes digital storytelling more accessible and engaging for children, as detailed in Fig. 12.

4.3 Story consistency

Our framework keeps a hierarchical organization of all asset files, which enables asset re-usages. Consequently, this structure ensures appearance consistency across all visual elements, both within neighboring frames and across different scenes (see Fig.13).

Additionally, our framework ensures consistency across modalities, as visual and audio elements together create an emotive storytelling experience. In this manuscript, the generated auditory

Figure 10: Story outcomes (screenshots and speech) with agent-wise intervention. The original base story is shown in the first column, while a variant on world settings is shown in the second. The last column is the variant where both world settings and character settings are intervened. The results show a strong complaint to the original plot, as all outcomes show the same key events (shown in rows). However, the diversity of the intervention modality is ensured at the same time.

descriptions of scenes are paired with corresponding visualizations in Fig.14, illustrating how our framework maintains alignment between visual and audio styles.

In addition, although a naive LLM function call with examples can achieve a similar level of consistency, its generated length is largely constrained. Our experiments show that this naive approach produces stories with an average length of 33 screenplay items (approximately two minutes in videos). We attribute this to the length limitation of single LLM function call. On the contrary, with the hierarchical decomposition approach, the example story generated by our framework is two times longer. And the length can be further extended with instructions.

5 Discussions

In this paper, we introduce the "StoryAgent" framework, which utilizes text as a central medium to organize existing generative

Figure 11: Character dialogue with scene understanding. (left) "Look at our house Sam. It is not the treasure we expected, but it's a symbol of our adventure today" (right) "You are right Tim. And remember the times we used to sit on that fence, dreaming about finding treasure?"

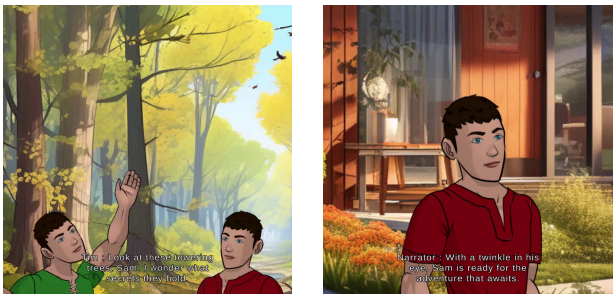

Figure 12: Camera shoot and character gesture with scene understanding. (left) The camera shifts to focus on the trees while simultaneously, the character points at them as he discusses them. (right) The camera zooms in on the character's face while the narrator describes his eyes.

Figure 13: Character consistency across scenes. Based on the textual descriptions of the generative narrative, the character assets are reused across scenes, ensuring visual consistency.

models to produce long-duration digital storytelling. By addressing three principal hurdles in storytelling (long-duration consistency between modalities, scene interactivity, and accessible human intervention), the outputs of this framework can either be of personal use by non-professionals or serve as prototypes in professional digital storytelling production.

In addition, we argue such a framework offers more flexibility in real-world digital storytelling production: First, the rapid emergence of new generative models and tools for various modalities necessitates their swift applications into a larger pipeline at minimal

[Visual] A small room with a single bed, a wooden desk, a bookshelf filled with books, a window overlooking the yard, and a blue rug. [Audio] Background sounds of a quiet suburban neighborhood, with distant sounds of children playing and birds chirping.

[Visual] A dense forest with towering trees, a carpet of fallen leaves, a narrow trail, bird nests in branches, and a quiet stream. [Audio] Background sounds of rustling leaves, chirping birds, and distant animal sounds create an atmosphere of being deep in the woods.

Figure 14: Consistency across modalities. Our framework ensure the coherence between visual and audio elements by aligning their text descriptions.

cost. Our model-independent framework facilitates a plug-and-play approach, allowing the latest research advancements to be easily incorporated. This ensures that the performance of our framework continuously benefits from ongoing research developments. Second, by replacing generative assets with human-crafted ones, our framework addresses the specific limitations faced by indie developers in certain modalities. For example, graphic artists can use our pipeline to integrate music into their digital stories, whereas musicians can employ it to generate visual representations of their audio works. Finally, the text-based nature of our framework permits human artists to readily adjust intermediate products, thereby enhancing the controllability of the production process.

However, we identify several limitations in our study as follows: (1) Although the performance of the pipeline is shown by the examples in the paper, a quantitative user study or objective evaluation can illustrate its capability further. (2) While our approach has achieved broad coverage of story components, the narrative presentation—encompassing duration, order, and style of storytelling (collectively referred to as 'discourse' in [Kybartas and Bidarra 2017])—is only minimally explored. (3) Generation results are constrained by the performance and capabilities of state-of-the-art singular-modality models. This limitation affects the alignment between the plots and the downstream components, potentially impacting the overall coherence and immersion of the generated results. (4) Our experimentation with scene interactivity is limited to a small number of scenarios. Future research should explore more complex interaction modes, such as enabling characters to pick up items or alter the status of objects within the environment, to enhance the interactive experience in 2D art styles. (5) Visual storytelling can be further improved by introducing more complex and diversified cinematography.

References

Prithviraj Ammanabrolu, Ethan Tien, Wesley Cheung, Zhaochen Luo, William Ma, Lara J. Martin, and Mark O. Riedl. 2020. Story Realization: Expanding Plot Events into Sentences. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 05 (April 2020), 7375–7382. https://doi.org/10.1609/aaai.v34i05.6232

J. Timothy Balint and Rafael Bidarra. 2023. Procedural Generation of Narrative Worlds. *IEEE Transactions on Games* 15, 2 (2023), 262–272. https://doi.org/10.1109/TG.2022.

3216582 Green Open Access added to TU Delft Institutional Repository 'You share, we take care!' – Taverne project https://www.openaccess.nl/en/you-share-we-take-care Otherwise as indicated in the copyright section: the publisher is the copyright holder of this work and the author uses the Dutch legislation to make this work public..

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. 2023. Improving Image Generation with Better Captions. https://cdn.openai.com/papers/dall-e-3.pdf

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. 2023a. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288* (2023).

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. 2023b. ZoeDepth: Zero-shot Transfer by Combining Relative and Metric Depth. https://doi.org/10.48550/ARXIV.2302.12288

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. 2024. Video generation models as world simulators. (2024). https://openai.com/research/video-generation-models-as-world-simulators

Marc Cavazza, Jean-Luc Lugrin, David Pizzi, and Fred Charles. 2007. Madame bovary on the holodeck: immersive interactive storytelling. In *Proceedings of the 15th ACM International Conference on Multimedia* (Augsburg, Germany) *(MM '07)*. Association for Computing Machinery, New York, NY, USA, 651–660. https://doi.org/10.1145/1291233.1291387

Che-Jui Chang, Danrui Li, Seonghyeon Moon, and Mubbasir Kapadia. 2024a. On the Equivalency, Substitutability, and Flexibility of Synthetic Data. *arXiv preprint arXiv:2403.16244* (2024).

Che-Jui Chang, Danrui Li, Deep Patel, Parth Goel, Honglu Zhou, Seonghyeon Moon, Samuel S Sohn, Sejong Yoon, Vladimir Pavlovic, and Mubbasir Kapadia. 2024b. Learning from Synthetic Human Group Activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21922–21932.

Che-Jui Chang, Samuel S Sohn, Sen Zhang, Rajath Jayashankar, Muhammad Usman, and Mubbasir Kapadia. 2023. The Importance of Multimodal Emotion Conditioning and Affect Consistency for Embodied Conversational Agents. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 790–801.

Che-Jui Chang, Sen Zhang, and Mubbasir Kapadia. 2022a. The IVI Lab entry to the GENEA Challenge 2022–A Tacotron2 based method for co-speech gesture generation with locality-constraint attention mechanism. In *Proceedings of the 2022 International Conference on Multimodal Interaction*. 784–789.

Che-Jui Chang, Long Zhao, Sen Zhang, and Mubbasir Kapadia. 2022b. Disentangling audio content and emotion with adaptive instance normalization for expressive facial animation synthesis. *Computer Animation and Virtual Worlds* 33, 3-4 (2022), e2076.

Chien-Yuan Chen, Sai-Keung Wong, and Wen-Yun Liu. 2020. Generation of small groups with rich behaviors from natural language interface. *Computer Animation and Virtual Worlds* 31, 4-5 (2020), e1960. https://doi.org/10.1002/cav.1960

CoquiAI. 2021. coqui-ai/TTS: a deep learning toolkit for Text-to-Speech, battle-tested in research and production. https://github.com/coqui-ai/TTS?tab=readme-ov-file

James J. Cummings and Jeremy N. Bailenson. 2016. How Immersive Is Enough? A Meta-Analysis of the Effect of Immersive Technology on User Presence. *Media Psychology* 19, 2 (2016), 272–309. https://doi.org/10.1080/15213269.2015.1015740 arXiv:https://doi.org/10.1080/15213269.2015.1015740

Adele De Jager, Andrea Fogarty, Anna Tewson, Caroline Lenette, and Katherine M Boydell. 2017. Digital storytelling in research: A systematic review. *The Qualitative Report* 22, 10 (2017), 2548–2582.

ElevenLabs. 2024. elevenlabs/elevenlabs-python. https://github.com/elevenlabs/elevenlabs-python original-date: 2023-03-26T11:59:52Z.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. 2024. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. arXiv:2403.03206 [cs.CV]

Mengyang Feng, Jinlin Liu, Kai Yu, Yuan Yao, Zheng Hui, Xiefan Guo, Xianhui Lin, Haolan Xue, Chen Shi, Xiaowen Li, Aojie Li, Xiaoyang Kang, Biwen Lei, Miaomiao Cui, Peiran Ren, and Xuansong Xie. 2023. DreaMoving: A Human Video Generation Framework based on Diffusion Models. arXiv:2312.05107 [cs.CV]

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. 2023. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning.

Ken Hartsook, Alexander Zook, Sauvik Das, and Mark O. Riedl. 2011. Toward supporting stories with procedurally generated game worlds. In *2011 IEEE Conference on Computational Intelligence and Games (CIG'11)*. 297–304. https://doi.org/10.1109/CIG.2011.6032020

Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. 2023. Synthesizing Physical Character-Scene Interactions. In *ACM SIGGRAPH 2023 Conference Proceedings* (<conf-loc>, <city>Los Angeles</city>, <state>CA</state>, <country>USA</country>, </conf-loc>) *(SIGGRAPH '23)*. Association for Computing Machinery, New York, NY, USA, Article 63, 9 pages. https://doi.org/10.1145/3588432.3591525

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. 2022. Imagen Video: High Definition Video Generation with Diffusion Models. arXiv:2210.02303 [cs.CV]

Mubbasir Kapadia, Seth Frey, Alexander Shoulson, Robert W. Sumner, and Markus Gross. 2016a. CANVAS: computer-assisted narrative animation synthesis. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (Zurich, Switzerland) *(SCA '16)*. Eurographics Association, Goslar, DEU, 199–209.

Mubbasir Kapadia, Alexander Shoulson, Cyril Steimer, Samuel Oberholzer, Robert W. Sumner, and Markus Gross. 2016b. An event-centric approach to authoring stories in crowds. In *Proceedings of the 9th International Conference on Motion in Games* (Burlingame, California) *(MIG '16)*. Association for Computing Machinery, New York, NY, USA, 15–24. https://doi.org/10.1145/2994258.2994265

Juntae Kim, Yoonseok Heo, Hogeon Yu, and Jongho Nang. 2023. A Multi-Modal Story Generation Framework with AI-Driven Storyline Guidance. *Electronics* 12, 6 (2023). https://doi.org/10.3390/electronics12061289

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. 2023. AudioGen: Textually Guided Audio Generation. arXiv:2209.15352 [cs.SD]

Vikram Kumaran, Jonathan Rowe, Bradford Mott, and James Lester. 2023. SceneCraft: Automating Interactive Narrative Scene Generation in Digital Games with Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* 19, 1 (Oct. 2023), 86–96. https://doi.org/10.1609/aiide.v19i1.27504

Ben Kybartas and Rafael Bidarra. 2017. A Survey on Story Generation Techniques for Authoring Computational Narratives. *IEEE Transactions on Computational Intelligence and AI in Games* 9, 3 (2017), 239–253. https://doi.org/10.1109/TCIAIG.2016.2546063

Joe Lambert. 2013. *Digital storytelling: Capturing lives, creating community*. Routledge.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. 2024. Playground v2.5: Three Insights towards Enhancing Aesthetic Quality in Text-to-Image Generation. arXiv:2402.17245 [cs.CV]

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 *Thirty-seventh Conference on Neural Information Processing Systems*.

Jun Hao Liew, Hanshu Yan, Jianfeng Zhang, Zhongcong Xu, and Jiashi Feng. 2023. MagicEdit: High-Fidelity and Temporally Coherent Video Editing. arXiv:2308.14749 [cs.CV]

Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qiang Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone. 2023a. LLM+P: Empowering Large Language Models with Optimal Planning Proficiency. arXiv:2304.11477 [cs.AI] https://arxiv.org/abs/2304.11477

Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, Yanfeng Wang, and Weidi Xie. 2024. Intelligent Grimm – Open-ended Visual Storytelling via Latent Diffusion Models. arXiv:2306.00973 [cs.CV]

Xubo Liu, Zhongkai Zhu, Haohe Liu, Yi Yuan, Meng Cui, Qiushi Huang, Jinhua Liang, Yin Cao, Qiuqiang Kong, Mark D. Plumbley, and Wenwu Wang. 2023b. WavJourney: Compositional Audio Creation with Large Language Models. arXiv:2307.14335 [cs.SD]

Amaury Louarn, Marc Christie, and Fabrice Lamarche. 2018. Automated staging for virtual cinematography. In *Proceedings of the 11th ACM SIGGRAPH Conference on Motion, Interaction and Games* (Limassol, Cyprus) *(MIG '18)*. Association for Computing Machinery, New York, NY, USA, Article 4, 10 pages. https://doi.org/10.1145/3274247.3274500

Adyasha Maharana, Darryl Hannan, and Mohit Bansal. 2022. StoryDALL-E: Adapting Pretrained Text-to-Image Transformers for Story Continuation. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII* (Tel Aviv, Israel). Springer-Verlag, Berlin, Heidelberg, 70–87. https://doi.org/10.1007/978-3-031-19836-6_5

Marcel Marti, Jodok Vieli, Wojciech Witoń, Rushit Sanghrajka, Daniel Inversini, Diana Wotruba, Isabel Simo, Sasha Schriber, Mubbasir Kapadia, and Markus Gross. 2018. CARDINAL: Computer Assisted Authoring of Movie Scripts. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces* (<conf-loc>, <city>Tokyo</city>, <country>Japan</country>, </conf-loc>) *(IUI '18)*. Association for Computing Machinery, New York, NY, USA, 509–519. https://doi.org/10.1145/3172944.3172972

Timothy Merino, Roman Negri, Dipika Rajesh, M Charity, and Julian Togelius. 2023. The five-dollar model: generating game maps and sprites from sentence embeddings. In *Proceedings of the Nineteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* (Salt Lake City) *(AIIDE '23)*. AAAI Press, Article 11, 9 pages. https://doi.org/10.1609/aiide.v19i1.27506

Chris Miller, Mayank Dighe, Chris Martens, and Arnav Jhala. 2019. Stories of the town: balancing character autonomy and coherent narrative in procedurally generated worlds. In *Proceedings of the 14th International Conference on the Foundations of Digital Games* (San Luis Obispo, California, USA) *(FDG '19)*. Association for Computing Machinery, New York, NY, USA, Article 82, 9 pages. https://doi.org/10.1145/3337722.3341850

Music Technology Group of Universitat Pompeu Fabra. [n. d.]. Freesound. https://freesound.org/

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. arXiv:2304.03442 [cs.HC]

Chen Qian, Xin Cong, Wei Liu, Cheng Yang, Weize Chen, Yusheng Su, Yufan Dang, Jiahao Li, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Communicative Agents for Software Development. arXiv:2307.07924 [cs.SE]

Zhongfei Qing, Zhongang Cai, Zhitao Yang, and Lei Yang. 2023. Story-to-Motion: Synthesizing Infinite and Controllable Character Animation from Long Text. In *SIGGRAPH Asia 2023 Technical Communications* (<conf-loc>, <city>Sydney</city>, <state>NSW</state>, <country>Australia</country>, </conf-loc>) *(SA '23)*. Association for Computing Machinery, New York, NY, USA, Article 28, 4 pages. https://doi.org/10.1145/3610543.3626176

Inês Salselas and Rui Penha. 2019. The role of sound in inducing storytelling in immersive environments. In *Proceedings of the 14th International Audio Mostly Conference: A Journey in Sound* (Nottingham, United Kingdom) *(AM '19)*. Association for Computing Machinery, New York, NY, USA, 191–198. https://doi.org/10.1145/3356590.3356619

Xiaoqian Shen and Mohamed Elhoseiny. 2023. StoryGPT-V: Large Language Models as Consistent Story Visualizers. arXiv:2312.02252 [cs.CV]

Simpleton. 2021. CC2D Essential Bundle | 2D Characters | Unity Asset Store. https://assetstore.unity.com/packages/2d/characters/cc2d-essential-bundle-187410

Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. 2019. Neural state machine for character-scene interactions. *ACM Trans. Graph.* 38, 6, Article 209 (nov 2019), 14 pages. https://doi.org/10.1145/3355089.3356505

Damian Stewart. [n. d.]. Compel: A prompting enhancement library for transformers-type text embedding systems. https://github.com/damian0815/compel

Fernanda De La Torre, Cathy Mengying Fang, Han Huang, Andrzej Banburski-Fahey, Judith Amores Fernandez, and Jaron Lanier. 2024. LLMR: Real-time Prompting of Interactive Worlds using Large Language Models. arXiv:2309.12276 [cs.HC]

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903 [cs.CL]

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 1800, 14 pages.

Jing Wu and Der-Thanq Victor Chen. 2020. A systematic review of educational digital storytelling. *Computers & Education* 147 (2020), 103786. https://doi.org/10.1016/j.compedu.2019.103786

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. arXiv:2308.08155 [cs.AI]

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. 2021. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *Neural Information Processing Systems (NeurIPS)*.

Feifei Xu, Xinpeng Wang, Yunpu Ma, Volker Tresp, Yuyi Wang, Shanlin Zhou, and Haizhou Du. 2020. Controllable Multi-Character Psychology-Oriented Story Generation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 1675–1684. https://doi.org/10.1145/3340531.3411937

Pei Xu, Kaixiang Xie, Sheldon Andrews, Paul G. Kry, Michael Neff, Morgan Mcguire, Ioannis Karamouzas, and Victor Zordan. 2023. AdaptNet: Policy Adaptation for Physics-Based Character Control. *ACM Trans. Graph.* 42, 6, Article 177 (dec 2023), 17 pages. https://doi.org/10.1145/3618375

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *Advances in Neural Information Processing Systems* 36 (2023). Publisher Copyright: © 2023 Neural information processing systems foundation. All rights reserved.; 37th Conference on Neural Information Processing Systems, NeurIPS 2023 ; Conference date: 10-12-2023 Through 16-12-2023.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023b. ReAct: Synergizing Reasoning and Acting in Language Models. arXiv:2210.03629 [cs.CL]

Hongwei Yi, Justus Thies, Michael J. Black, Xue Bin Peng, and Davis Rempe. 2024. Generating Human Interaction Motions in Scenes with Text Control. arXiv:2404.10685 [cs.CV]

Jia-Qi Zhang, Xiang Xu, Zhi-Meng Shen, Ze-Huan Huang, Yang Zhao, Yan-Pei Cao, Pengfei Wan, and Miao Wang. 2021. Write-An-Animation: High-level Text-based Animation Editing with Character-Scene Interaction. *Computer Graphics Forum* 40, 7 (2021), 217–228. https://doi.org/10.1111/cgf.14415

Zhexin Zhang, Jiaxin Wen, Jian Guan, and Minlie Huang. 2022. Persona-Guided Planning for Controlling the Protagonist's Persona in Story Generation. arXiv:2204.10703 [cs.CL]