ACM MM LGM3A’25

Dynamic Storytelling with Multimodal Synchronized Video Generation

Ye Zhiqiu

Introduction

Why it's challenging

- Users want **coherent, engaging videos** from concise input (e.g., one sentence).
- Existing methods often:
  - Requires training (expensive, hard to scale).
  - Lack **structured narrative across scenes**.

Goal:

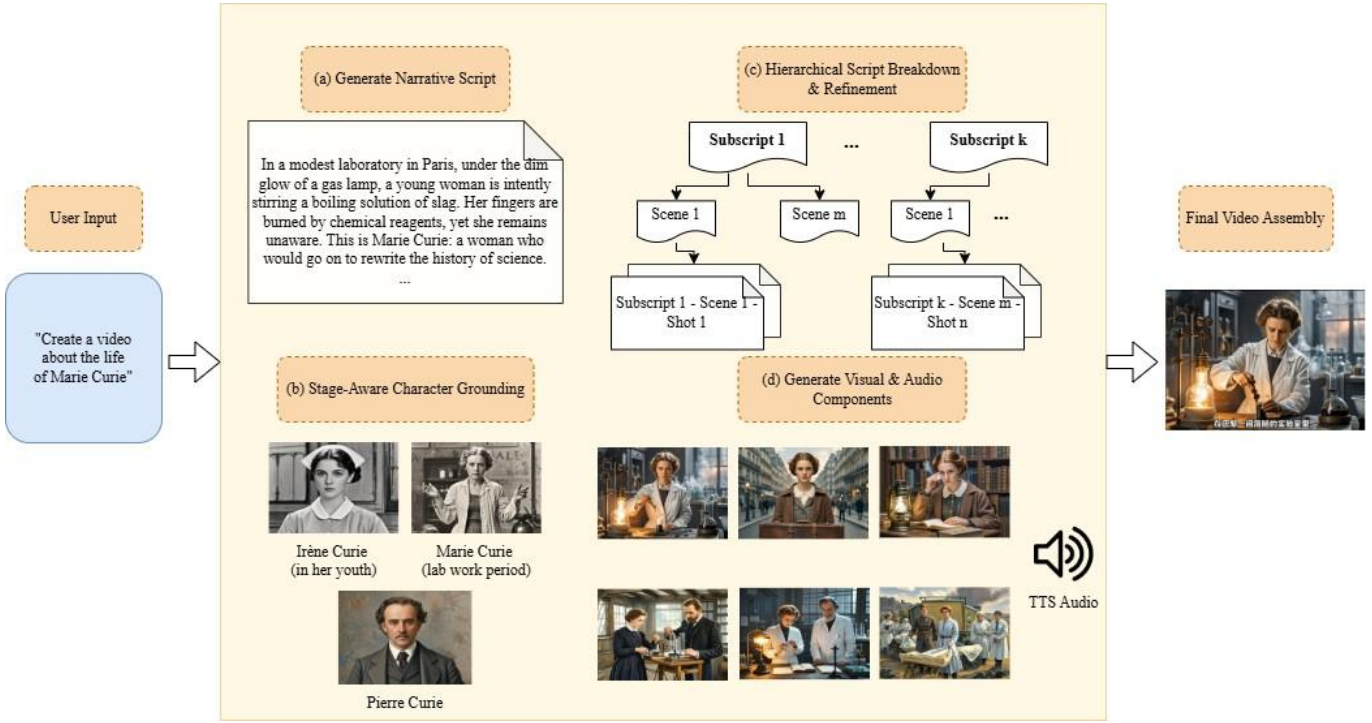
- Turn a **one-sentence input** into a **full, coherent, multimodal video** (text + image + audio).
- Key features:
  a. **Training-free hierarchical multi-agent pipeline** for story & character generation.
  b. **Cross-modal alignment** for text, visual, and audio coherence.
  c. **Stage-aware character grounding** for consistent identity and appearance.

System Overview

Overall Input: Single-sentence user request. → **Overall output:** Video with narrative and audios.

4 modules:

- Narrative Expansion
- Character Grounding
- Hierarchical Script Planning
- Refinement & Multimodal Generation



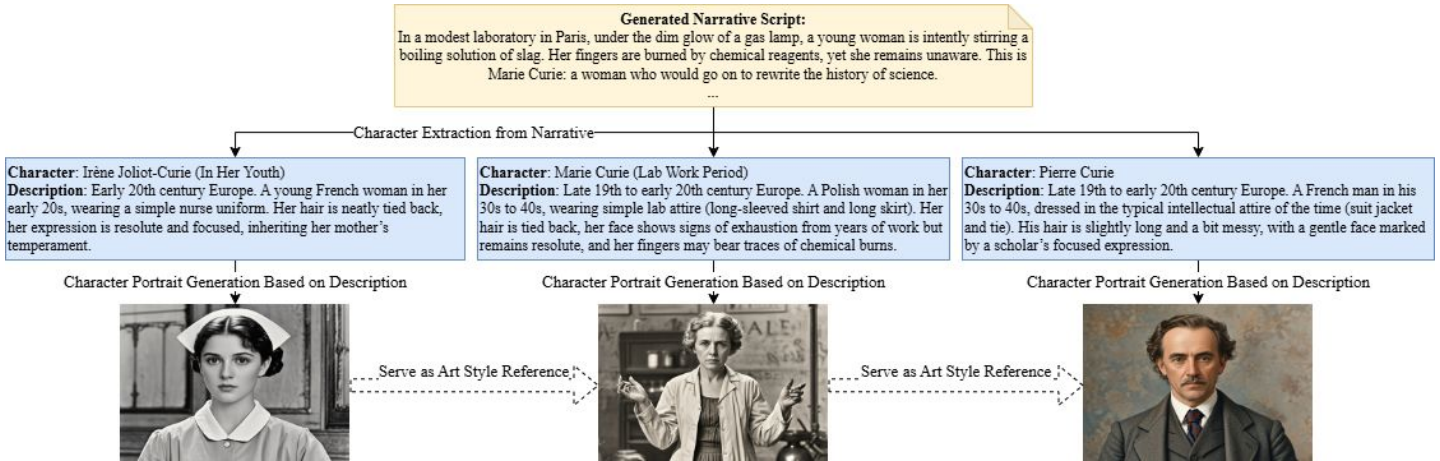
Character Grounding - Maintains **consistent identity** of characters

Key steps:

- Stage-Aware Character Extraction

Identify the key characters and recognizes **life stages or major transitions** (e.g., aging, events in storyline).

- Visual Reference Generation



Script & Scene Planning - Breakdown narrative for smooth generation

Key steps:

- Hierarchical decomposition: **Subscripts → Scenes → Shots**
- Subscript Level
  - Base on **major events or story arcs**.
  - Provides a high-level roadmap for the story.
- Scene Level
  - Base on **emotional tone** and **temporal setting**.
  - Examples: morning vs evening, indoor vs outdoor.
- Shot Level
  - Provide details like: Characters involved; time, location tags; Visual description
  - **Number of shots is adaptively estimated based on narrative length.**

Shot example

Character: Marie Curie (lab work period)

Visual Description: Marie Curie stands at the table, stirring a boiling solution, focused and determined.

Shot Type: Close-up, static

Emotional Highlight: Close-up captures her concentration, emphasizing dedication and persistence.

……

Visual Coherence Refinement - logical consistency

Key steps:

- Logical Consistency Check
  - Detects conflicts in actions, props, or scene elements.
  - Example: character age or attire matches the story context.
- Prompt Refinement for Image/Video Generation
  - Ensures **high-quality, visually coherent results**

Synchronization & Video Assembly

Key steps:

- **Image & Audio Generation**
- **Multimodal Assembly**
  - Calculate audio time after TTS generation → synchronized audio and visual elements

Results & Evaluation

- Automatic evaluation (VBench)

Table 1: Comparison between baseline and our system across four dimensions. Positive percentages indicate improvement.

Dimension	Ours	Baseline	Improvement
Human Anatomy	0.9440	0.8523	9.17%
Human Identity	0.6517	0.4751	17.65%
Imaging Quality	0.7563	0.6735	8.94%
Subject Consistency	0.7529	0.7022	5.08%

- Human evaluation:

Table 2: Comparison between baseline and our system across four dimensions. Positive percentages indicate improvement.

Dimension	Ours	Baseline	Improvement
Narrative coherence	4.71	4.25	0.46
Cross-modal alignment	4.42	3.75	0.67
Character consistency	4.08	3.42	0.67
Temporal synchronization	4.38	4.08	0.29

Conclusion & Takeaways

Contributions:

- Training-free hierarchical multimodal video generation.
- Stage-aware character grounding.
- Cross-modal alignment across text, image, audio.

Future work:

- More complex narratives
- web search integration