

A APPENDIX

A.1 LARGE LANGUAGE MODEL (LLM) USAGE

This paper benefited from the use of LLM (e.g., ChatGPT) for grammar correction and language polishing. All ideas, experimental designs, and analyses are the sole responsibility of the authors.

A.2 DETAILS OF MSB & M²SB

The overview of MSB and M²SB is illustrated in Fig. 8. We first use GPT-4o to generate the story outline and the possible characters/scenes images, then we prompt GPT-4o to generate the story scripts in MSB and M²SB for each keyframe, as shown in Fig. 9.

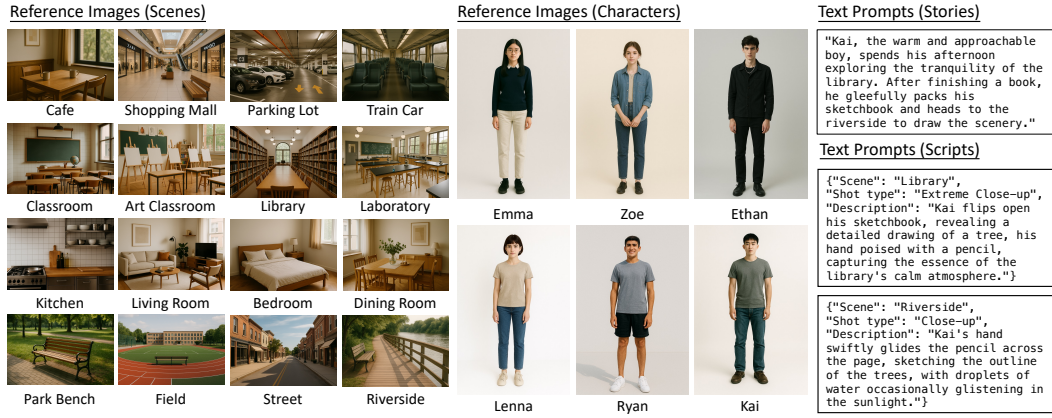


Figure 8: Overview of Multimodal Story Benchmark

A.3 DATASET FOR SHOT-TYPE CONTROL

In this section, we detail the pipeline used to construct the dataset for shot-type control. (1) We collect video data from the Condensed Movie Dataset (Bain et al., 2020). (2) We apply ByteTrack (Zhang et al., 2022) to track character trajectories across frames, enabling retrieval of the same individual across different scenes. (3) We randomly sample two frames to form a pair and use CLIP (Radford et al., 2021) to verify that both frames depict the same character, thereby avoiding trivial duplication or copy-paste artifacts. (4) We apply a shot-type classifier trained by Xie et al. (2025) to categorize the target frame into one of the canonical shot types. (5) Finally, we use Qwen2.5-VL (Bai et al., 2025) to generate a caption for the target frame, which serves as the textual prompt. The resulting dataset contains 715 example pairs, which we use for training. Examples in this dataset are illustrated in Fig. 10. We further provide more example from Story2Screen on MSB in Fig. 11.

A.4 STORY2SCREEN WITH EXISTING TI2V MODEL.

We employ Veo3 to demonstrate that Story2Screen can be integrated with recent TI2V models to form longer, meaningful videos. Specifically, we leverage the text prompts produced in Stage 1 of Story2Screen and the keyframes generated by ConsistFilmer to synthesize short clips, which are then concatenated into longer videos. We also compare against closed-source models (Sora and Veo3). Story2Screen not only enables the generation of longer videos but also improves text-video alignment with the abstract, resulting in more semantically meaningful content. Qualitative comparisons are provided in Appendix A.4. Story2Screen can generate longer videos with global and local consistency in the scene, and more diverse shot types. We provide more Qualitative Comparison with the existing T2V model in Figure 12. We also provide the video in the supplementary.

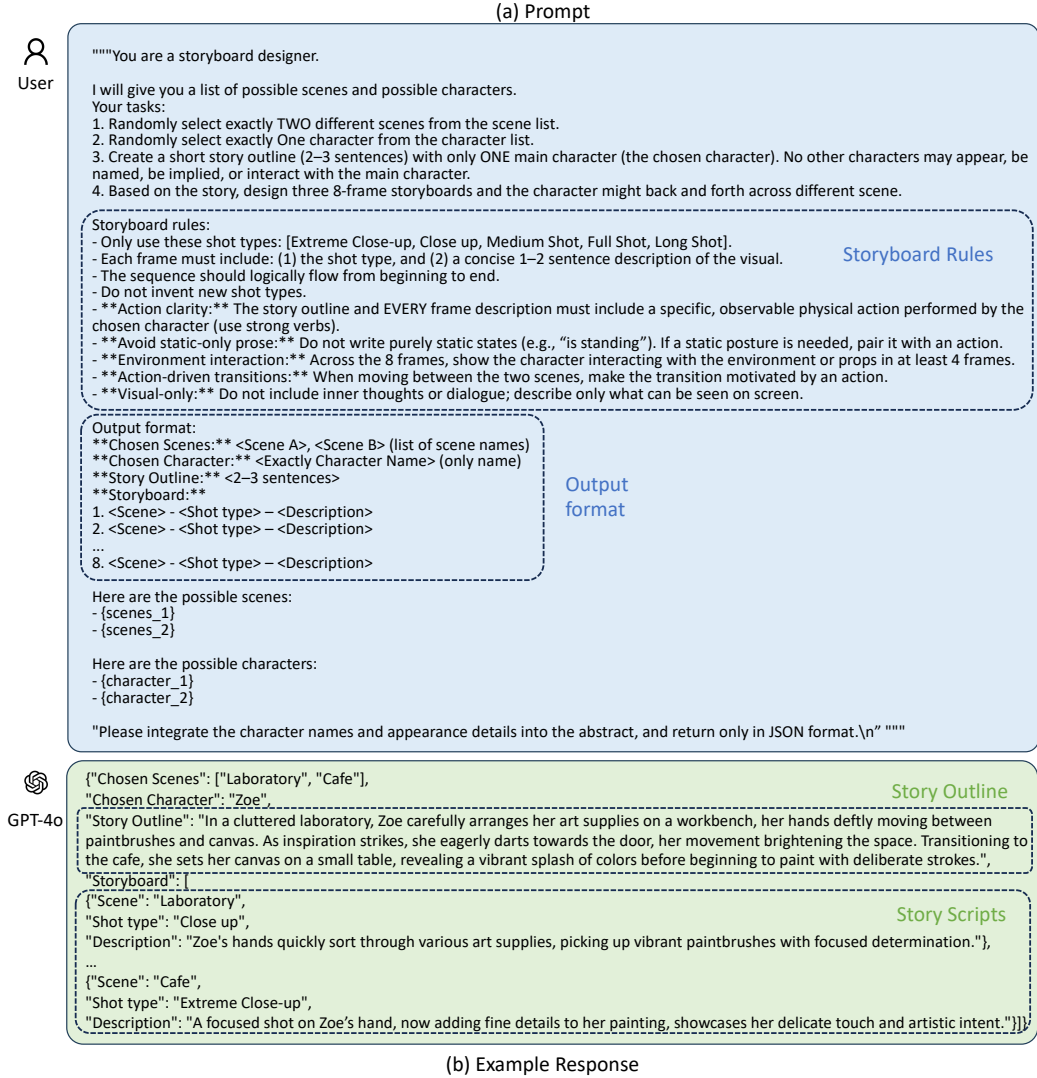


Figure 9: Prompt for GPT to generate scripts prompt in MSB

A.5 MORE QUALITATIVE COMPARISON

We provide more qualitative comparison with existing methods in Fig. 13.

A.6 LIMITATION

ConsistFilmer primarily relies on the previous frame as a reference to maintain temporal consistency. While this is effective for local continuity, it may be insufficient for capturing long-range narrative structures required in real-world story generation. Future directions could involve integrating higher-level semantic representations, such as multimodal knowledge graphs or narrative planning modules, to provide global guidance and enhance the overall coherence of story progression. We regard this as a promising avenue for future work.





Reference Image	Target Images	Text Prompt	Shot Type
		The young man with curly hair appears to be in a moment of deep contemplation or perhaps even distress. his expression is serious, and his gaze is directed off to the side.	<u>Close-up Shot</u>
		A man sitting on a makeshift campsite by a serene pond, surrounded by lush greenery and towering trees.	<u>Full Shot</u>
		the character, dressed in a tweed blazer over a sweater vest and tie, sits in a dimly lit room with a classic, somewhat formal ambiance.	<u>Close-up Shot</u>
		A man stands in a dimly lit kitchen, his body language tense and his expression one of intense concentration or perhaps fear.	<u>Medium Shot</u>

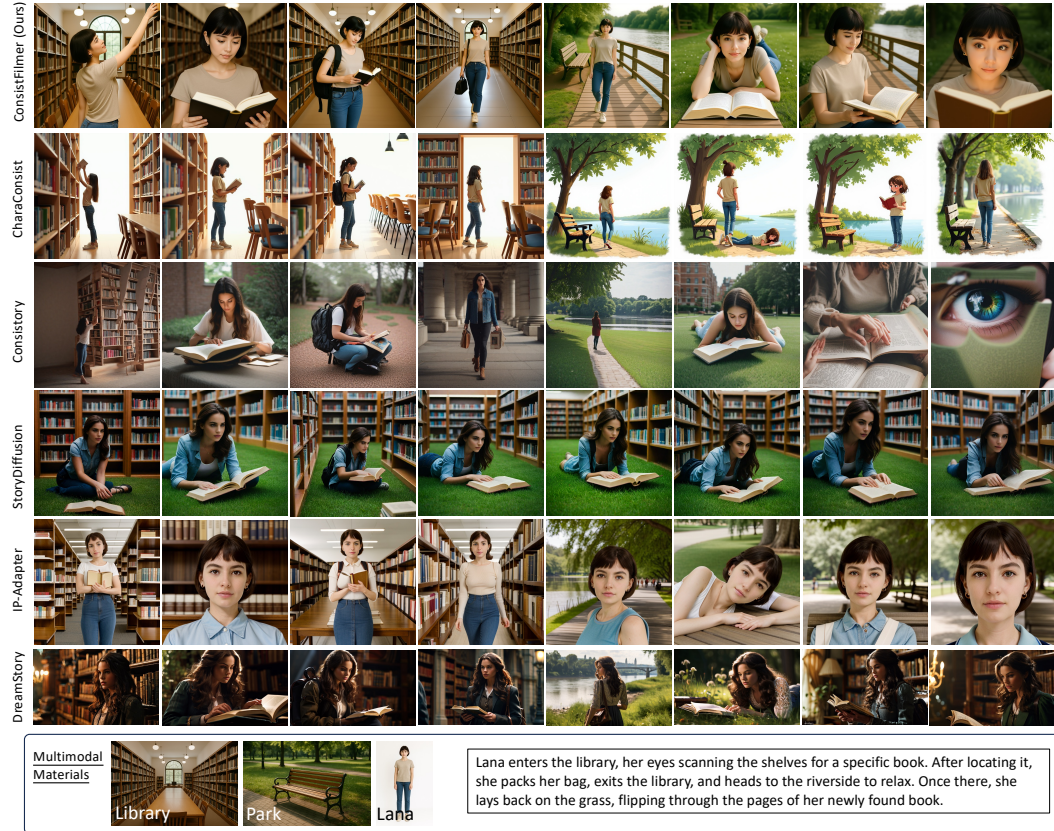
Figure 10: Dataset for shot-type control from CMD

		 Medium Shot	 Close up	 Extreme Close-up
Living Room	Kai	Kai leaning against the wall, arms crossed, still lost in his thoughts while gazing at the street outside.	Kai's face shifts to a thoughtful expression as he gazes out the window, his hands resting on the windowsill as he contemplates something profound.	Kai's eye depicts a spark of reflection, highlighting his deep contemplation and the world just beyond the glass.
		 Full Shot	 Medium Shot	 Close up
Bedroom	Lana	Lana walks towards the door, her bag slung over one shoulder, as she turns to take one last look at her bedroom.	Lana stands and gathers her art supplies from the desk, placing them into a bag while glancing out the window.	Lana leans forward, carefully flipping through the pages of her sketchbook, her focused expression highlighting her determination.
		 Full Shot	 Medium Shot	 Close up
Street	Zoe	Zoe steps out of the art classroom, walking along the vibrant street, her bag slung over her shoulder.	Zoe pauses to look up at a tree, reaching out to touch the leaves as they sway in the breeze.	Zoe takes a deep breath, closing her eyes briefly to savor the fresh air, a look of inspiration on her face.

Figure 11: Additional results from MSB



Figure 12: Qualitative Comparison with T2V models.



- (1) **[Medium Shot]** Lana stands in front of the towering bookshelves, reaching up to pull down a book from the top shelf.
- (2) **[Close Up]** Lana flips open the book, her focused expression reflecting her calm nature as she scans the first few pages.
- (3) **[Full Shot]** Lana carefully places the book inside her backpack, adjusting the straps as she prepares to leave.
- (4) **[Long Shot]** Lana exits the library, stepping onto the street with a determined look as she starts her journey to the riverside.

- (5) **[Long Shot]** Lana approaches the riverside, stepping onto the grass and taking a moment to appreciate the view..
- (6) **[Medium Shot]** Lana lays down on the grass, propping herself on one elbow, and opens the book to the first chapter.
- (7) **[Medium Shot]** Lana's fingers flip through the pages, her calm demeanor evident as she becomes engrossed in the story.
- (8) **[Close-up]** The camera focuses on Lana's eyes as they sparkle with interest, capturing her connection to the book.

Figure 13: Qualitative comparison with prior methods.