Customized Visual Storytelling with Unified Multimodal LLMs

Supplementary Material

6. Details of MSB & M^{2}SB

Since existing storytelling datasets [33] primarily provide story descriptions as text-only inputs, they are insufficient for evaluating story customization, which requires multimodal conditioning such as specified characters, backgrounds, and desired shot types. To address this limitation, we introduce two new datasets, MSB and M^{2}SB, designed specifically for benchmarking multimodal story customization, as shown in Fig. 7.

To construct this dataset, we first employ GPT-4o to generate a diverse set of story outlines along with corresponding candidate character and scene images. We then prompt GPT-4o to produce detailed multimodal story scripts for each keyframe, including the full script prompt, reference character and background mentions, and shot-type annotations. In total, we generate 100 stories, each containing 8 multimodal scripts, resulting in 800 scripts overall. These datasets enable systematic evaluation of customization, consistency, and controllability in story visualization models.

7. Dataset for Shot-type Prompt Tuning

In this section, we describe the pipeline used to construct the dataset for tuning the shot-type prompt. The dataset is built through the following five steps:

- **Video collection** We begin by collecting video data from the Condensed Movie Dataset (CMD) [2].
- **Character tracking** We apply ByteTrack [51] to track character trajectories across frames, enabling retrieval of the same individual across different shots and scenes.
- **Frame pairing and identity verification** We randomly sample two frames whose temporal distance is larger than a predefined threshold, and use CLIP [28] to verify that both frames depict the same character, thereby avoiding trivial duplicates or copy–paste artifacts.
- **Shot-type annotation** We apply the shot-type classifier from [44] to categorize the target frame into one of the canonical cinematic shot types.
- **Caption generation** Finally, we use Qwen2.5-VL [1] 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. Example pairs from this dataset are shown in Fig. 8.

Table 6. Comparison of VstoryGen and Wan2.2 on MSB

Method	Clip-T
Wan2.2-TI2V-5B (Baseline)	0.240
VstoryGen (w/ Wan2.2)	**0.285**

Table 7. Ablation on Visual Reference Retrieval with different numbers of reference images from previous keyframes.

μ	Clip-T	Avg-Consistency
0	0.284	0.855
1	**0.285**	**0.858**
2	0.282	0.856

8. More Experiments and Ablation Study

8.1. Comparison between Wan2.2 and VstoryGen.

We compare VstoryGen with the original Wan2.2 [35] model on the MSB benchmark. Since VstoryGen integrates all reference images beforehand, it produces video results that align more closely with the textual descriptions. We compute the CLIP score between each generated video segment and its corresponding text. In Tab. 6, our method achieves higher CLIP scores, indicating that it generates videos that are semantically closer to the given descriptions.

8.2. Ablation on Visual Reference Retrieval.

We ablate Visual Reference Retrieval with varying numbers of previous reference images, μ. As shown in Tab. 7, using only the most recent reference frame achieves the best performance, while incorporating earlier frames slightly degrades results. We hypothesize that multiple temporally distant references introduce conflicting or less relevant visual cues. As model must align its prediction with all references, excessive conditioning may overwhelm the model and dilute its focus on the most informative frame, I_{t-1}. Consequently, relying on a single temporally adjacent reference allows the model to maintain a clearer and more stable temporal signal, leading to improved consistency.

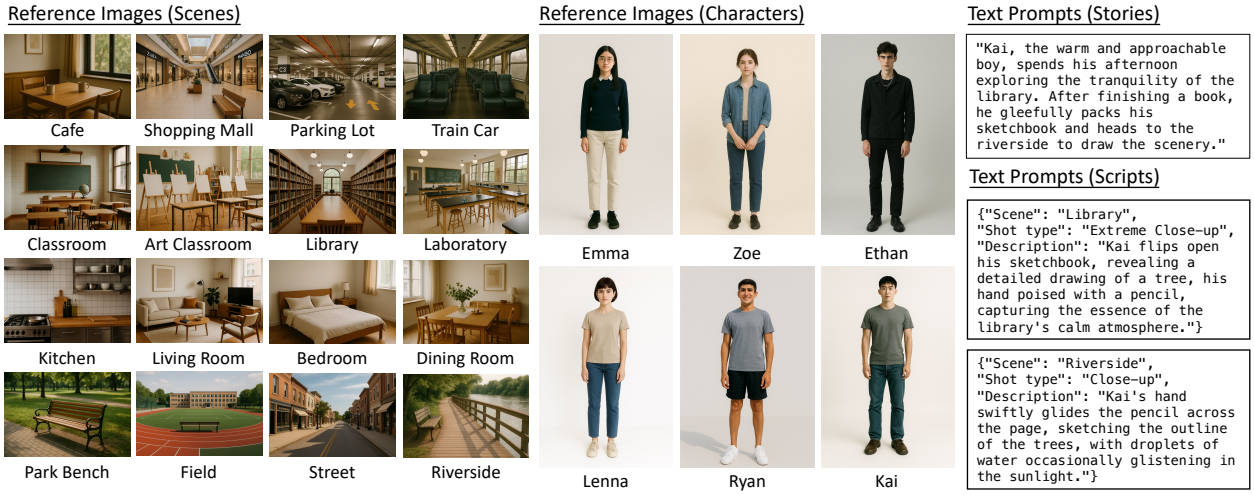

Figure 7. Overview of Multimodal Story Benchmark (MSB)

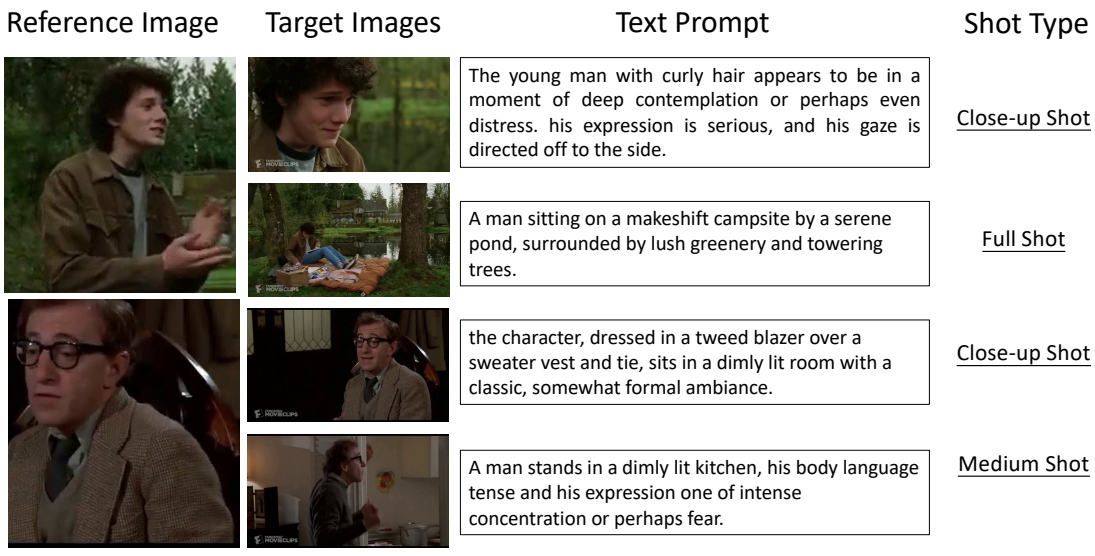

Figure 8. Dataset for shot-type control from CMD