DreamShot: Personalized Storyboard Synthesis with Video Diffusion Prior Supplementary Materials

Junjia Huang[1,2*] Binbin Yang[3*†] Pengxiang Yan[3] Jiyang Liu[3] Bin Xia[3]
Zhao Wang[3] Yitong Wang[3] Liang Lin[1,2,4] Guanbin Li[1,2,4‡]
[1]Sun Yat-sen University, [2]Peng Cheng Laboratory, [3]ByteDance Intelligent Creation
[4]Guangdong Key Laboratory of Big Data Analysis and Processing
huangjj77@mail2.sysu.edu.cn, wantong1017@163.com
linliang@ieee.org, liguanbin@mail.sysu.edu.cn
{yangbinbin.3, yanpengxiang.ai, liujiyang.liu, xiabin.zj, zhaoxu.bit}@bytedance.com
https://ll3rd.github.io/DreamShot/

1. Storyboard Dataset

1.1. More Details about the Data Construction

In this section, we provide a detailed description of the four stages in our storyboard data collection pipeline, along with the specific procedures applied at each stage.

Data Curation. To ensure the quality of raw videos collected from online sources, we apply multiple filtering criteria. Specifically, we discard videos released before 2015, those with a resolution below 720p, or a bitrate lower than 800 kbps, retaining only clips with minimal motion artifacts and high spatial clarity.

Scene Detect & Keyframe Extract. To extract high-quality keyframes from videos, we first segment each video into shots using PySceneDetect [1]. Within each shot, we compute the Laplacian score for all frames to assess sharpness and rank them accordingly, while also estimating optical flow to measure motion magnitude. Based on the combined rankings of sharpness and motion, we select the clearest and most distinct frames as keyframes.

Quality Filter. For each extracted keyframe, we apply multiple evaluation methods,including AES scoring [4], VLM-based assessment [6, 9], and image quality metrics, to ensure visual fidelity. We additionally use watermark and subtitle detection tools [2] to filter out any keyframes containing such artifacts.

Scene-Wise Grouping. After extracting keyframes, we group those belonging to the same scene and sharing narrative continuity. Since feeding too many frames into a VLM can degrade its accuracy, we adopt a sliding-window strategy, as shown in Fig. 1: keyframes are processed in small batches, and the VLM identifies those that form a coherent narrative sequence within each scene. Overlapping frames between windows are then used to determine the storyboard group to which each keyframe belongs, ensuring both temporal order and grouping accuracy.

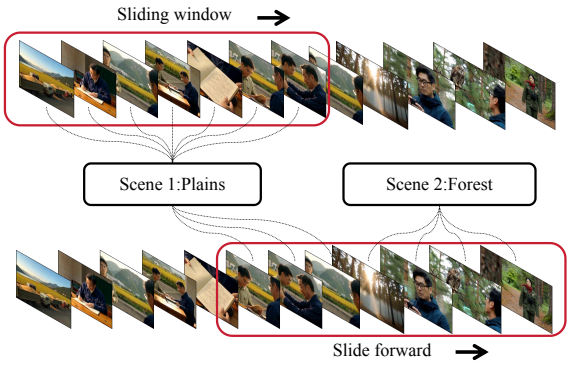


Figure 1. Using a sliding-window strategy with a VLM to extract storyboard groups from the same scene.

Role Extraction. When extracting roles from each storyboard group, roles often appear repeatedly across shots. Directly applying an instance segmentation model to each shot would therefore produce many duplicate detections. To address this, we further use a VLM to merge and deduplicate the extracted character regions. We avoid face-embedding–based clustering [3], as face recognition methods are unreliable for animated content or non-human characters.

Multi-View Role Generation. Directly using roles extracted from the current shot as reference images can lead to severe copy–paste artifacts, causing the model to simply replicate the reference appearance. To increase ref-

*Equal Contribution.
†Project Lead.
‡Corresponding Author.

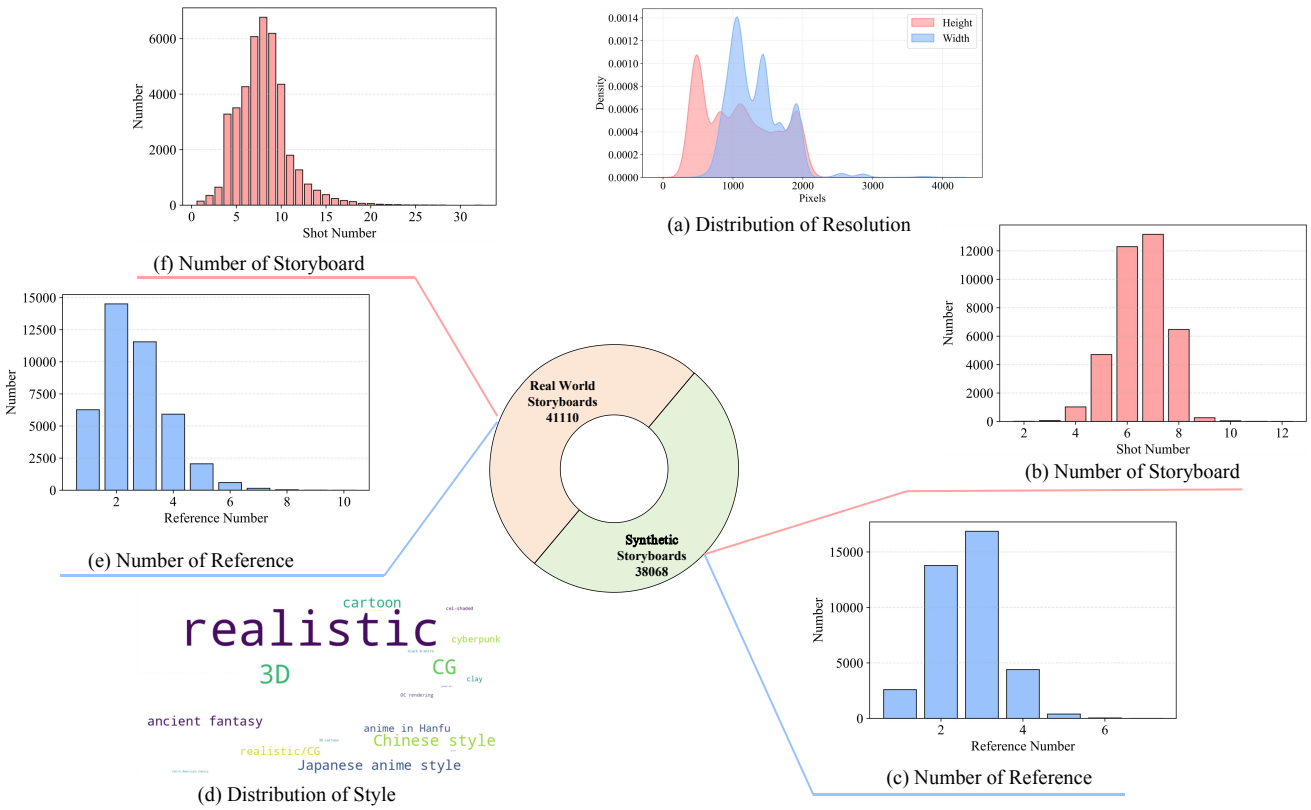


Figure 2. Data statistics of the DreamShot storyboard dataset. We report statistics for both data sources (real videos and AIGC-generated videos), including: the total number of storyboard samples; (a) resolution distribution of all samples; (b) shot-count distribution for synthetic storyboards; (c) reference-count distribution for synthetic storyboards; (d) overall style distribution; (e) reference-count distribution for real storyboards; and (f) shot-count distribution for real storyboards.

erence diversity, we first apply cross-scene matching to retrieve the same characters from other scenes and use them as additional references. We further employ character augmentation techniques—such as rotating characters with video models [5] to obtain front, side, and half-body views, or generating diverse portraits with reference-image models [7]. These generated candidates are further filtered using ArcFace [3] identity comparison to ensure that only images depicting the same role are retained. These augmented references provide richer appearance variations and effectively mitigate copy–paste behavior during generation.

Story Annotation. To ensure the narrative quality of the resulting storyboards, we use a VLM to generate structured annotations for all retained high-quality shots. The VLM provides shot-level descriptions covering perspective, environment, character actions, and stylistic attributes. For each reference roles, we further annotate detailed identity information, including gender, ethnicity, appearance, hairstyle, and clothing.

Final Quality Control. For the final storyboard data, we perform an additional quality check using both a VLM and human review to assess narrative coherence, annotation accuracy, and visual clarity.

1.2. Data Statistics

Through the above pipeline, we obtain approximately 41K real-world storyboard samples and 38K AI-generated samples. We conduct a detailed analysis of the dataset, as shown in Fig. 2, including the distribution of shot counts, reference-role counts, and resolutions. Real-world storyboards exhibit a broader range of shot lengths, from 2 to 30 shots, with most falling between 5 and 12, while AI-generated storyboards are primarily concentrated between 4 and 8 shots. Similarly, most storyboard groups contain 2 to 4 reference characters. All samples maintain high visual resolution, with many reaching up to 2K. As shown in Fig. 2 (d), the dataset also covers a wide variety of visual styles, including realistic, 3D, and anime aesthetics.

Based on this diverse dataset, we train our DreamShot model leveraging video-model priors to fully unlock its storyboard generation capability, achieving more consistent and more accurate shot synthesis.

DreamShot	CIDS(Character)		CSD(Style)		AES	Alignment
	Self↑	Cross↑	Self↑	Cross↑	Score↑	Score↑
Dataset source	Reference-to-Shot					
w/ Real Data	48.6	**52.2**	**64.7**	44.6	5.24	3.33
w/ Synthetic Data	46.4	49.3	63.2	43.5	**5.57**	3.38
Ours	**48.7**	51.6	64.5	**45.1**	5.50	**3.39**

Table 1. Effectiveness of Different Data Source.

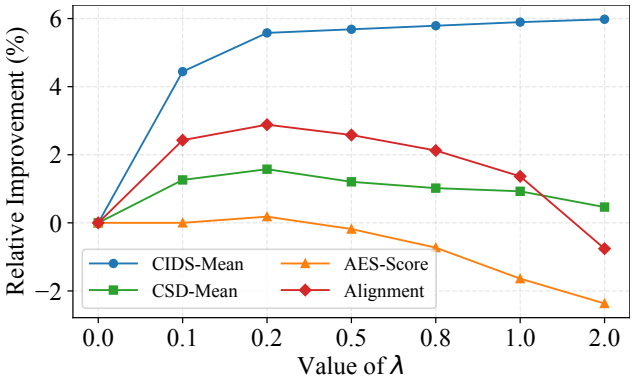
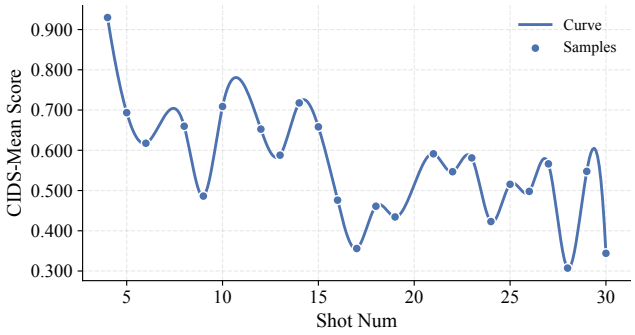
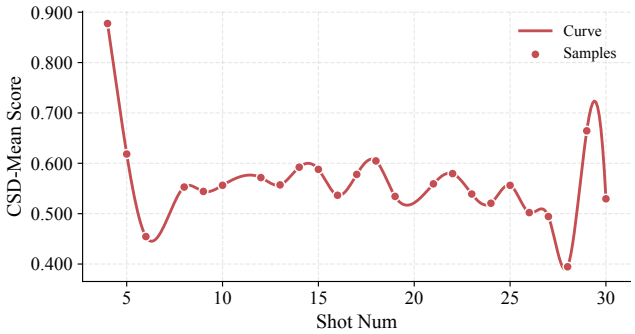


Figure 3. Effectiveness of λ in Training Objective.



(a) CIDS-Mean scores for different numbers of scenes.



(b) CSD-Mean scores for different numbers of scenes.

Figure 4. Effectiveness of Storyboard Number.

1.3. Effectiveness of Different Data Source

Tab. 1 summarizes the performance of models trained on different data sources. As shown, real storyboard data yield more consistent shot generation, particularly achieving a CDIS-Cross score of 52.2. This indicates that character identity and appearance in real storyboards align more faithfully with the reference characters, leading to improved cross-shot consistency. In contrast, synthetic storyboard data achieve higher aesthetic scores and better alignment score, suggesting that their enhanced visual quality can further boost the fidelity of generated shots. Therefore, combining real and synthetic data during training provides the best of both worlds—improving both the consistency and the overall quality of storyboard generation.

Fig. 7 presents storyboard samples from different data sources. Real data exhibit stronger consistency and more expressive shot composition, although the overall aesthetic quality is sometimes lower and the lighting tends to be darker. In contrast, synthetic storyboard data show higher aesthetic quality and brighter, more visually appealing lighting conditions.

2. Experiments

2.1. Effectiveness of λ in Training Objective

In Fig. 3, we further investigate the effectiveness of the coefficient λ in the training objective. The table reports the relative improvement compared to the setting without L_{RAC} (i.e. $\lambda = 0$). We observe that as λ gradually increases, the CIDS score consistently improves, reaching up to a 6% gain. However, when λ becomes as large as 0.5, the aesthetic score starts to decline. This is mainly because the increasing weight of L_{RAC} dilutes the influence of L_{diff}, introducing interference that ultimately reduces visual quality. Considering both consistency and aesthetics, we select $\lambda = 0.2$, which provides a noticeable CIDS improvement while maintaining high visual quality.

2.2. Performance under Different Numbers of Storyboards

Fig. 4 illustrates the performance of DreamShot under different storyboard number. We observe that the CSD-Mean score remains largely stable across 4 to 30 shots, indicating that our method maintains strong scene-style consistency even in very long storyboards. In contrast, the CIDS-Mean score gradually decreases as the number of shots increases, suggesting that character consistency becomes more challenging in ultra-long sequences. This limitation is primarily due to the distribution of our training data, where most storyboards contain only 6–10 shots.

2.3. More Qualitative Results

Qualitative Results in VistoryBench. Fig. 5 presents the quantitative comparison between our method and UNO [8] on VistoryBench [10]. VistoryBench primarily contains real reference images and each storyboard sequence may in-

Figure 5. Qualitative results between our method and UNO [8] in VistoryBench [10]. Our method demonstrates superior character consistency, as reflected in the first storyboard group where fine-grained details such as the girl's hair accessories are better preserved. In the case of long storyboard sequences, shown in the second group, our approach also exhibits strong generalization ability. The entire sequence maintains a more coherent style and scene layout compared with UNO.

clude multiple reference shots. The results show that our approach achieves strong performance in both character consistency and scene consistency. Notably, when generating ultra-long storyboards, our method generalizes well despite the limited number of long-storyboard samples in the training data. The overall style and scene coherence remain consistently stable throughout the storyboards. This demonstrates both the strong generalization ability of our approach and its robustness in preserving character identity.

Quantitative results under Different Modes. Fig. 8 presents the quantitative results of our method under different generation modes. The results show that our approach performs well both in maintaining consistency with the reference images and in responding accurately to the shot-level textual descriptions. At the same time, DreamShot preserves strong scene coherence across shots, retaining the original environment even after camera transitions and supporting smooth narrative progression.

2.4. Storyboards to Long Video

We further investigate DreamShot's ability to support long-video creation through storyboard generation. Specifically, we employ a image-to-video (I2V) model [5] to convert each generated shot into a 5-second video clip, and then concatenate these clips to form videos longer than 30 seconds. The resulting long videos are provided in the supplementary folder. The long video composed from our generated storyboards further demonstrates that the produced shots exhibit strong narrative coherence and consistent scene and character representation. Our method effec-

Figure 6. Examples of dynamic combat shots. DreamShot tends to generate clear and static storyboard frames, making it difficult to capture the sense of motion typically present in fighting storyboards.

tively decomposes long-form narratives into well-structured storyboard sequences, reducing the redundancy that often arises when generating long videos directly. Moreover, the storyboard representation offers clear advantages for editing and post-processing, as modifications to the final video can be achieved by editing individual shots rather than regenerating the entire sequence.

3. Limitation

3.1. Performance under Extremely Long Storyboards

In terms of storyboard number, DreamShot is primarily trained on storyboards containing around 6 to 10 shots. Although Dreamshot demonstrates strong generalization ability and can extend to longer sequences, its performance still degrades when facing ultra-long storyboards. As shown in Fig. 4a, the CIDS-Mean score gradually decreases as the number of shots increases. This limitation is largely attributed to the scarcity of long-storyboard samples in the training data. In future work, we plan to explore strategies for improving character consistency in ultra-long storyboards. Possible directions include expanding data collection to incorporate more long storyboards, or adopting self-forcing or other autoregressive techniques to reduce error accumulation across extended shot sequences.

3.2. Motion Storyboard

Another limitation lies in the realism of action-oriented or combat-related shots. In such scenes, character motions are typically accompanied by dynamic blur, which conveys a stronger sense of movement, impact, and physical intensity. However, in cinematic data, it is often difficult to extract clean and representative frames from fast-motion sequences. Moreover, as discussed in Sec. 1.1, our preprocessing pipeline employs a Laplacian-based filter to remove blurry frames. As a result, motion-blurred frames, which are essential for expressing dynamic actions, are frequently filtered out. Consequently, most of the training samples become static shots, and DreamShot struggles to generate storyboard frames with strong dynamic motion cues or the visual intensity characteristic of combat scenes, as shown in Fig. 6.

3.3. Resolution

Another limitation lies in the resolution of the generated outputs. This constraint primarily arises because existing video-based foundation models typically operate at around 480p or 720p, whereas image generation models can produce outputs at much higher resolutions, such as 1K. As a result, storyboards generated by video models may score lower on metrics such as aesthetic quality compared with those produced by high-resolution image models. However, storyboard frames are often used as a guiding structure for downstream video production or as preliminary visual references for creators. In these scenarios, maintaining strong consistency across shots is more important and can provide more reliable creative guidance. In future work, we plan to explore higher-resolution storyboard generation to further enhance visual quality.

References

[1] Pyscenedetect. https://github.com/Breakthrough/PySceneDetect, 2025. 1

[2] watermark-detection. https://github.com/boomb0om/watermark-detection, 2025. 1

[3] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019. 1, 2

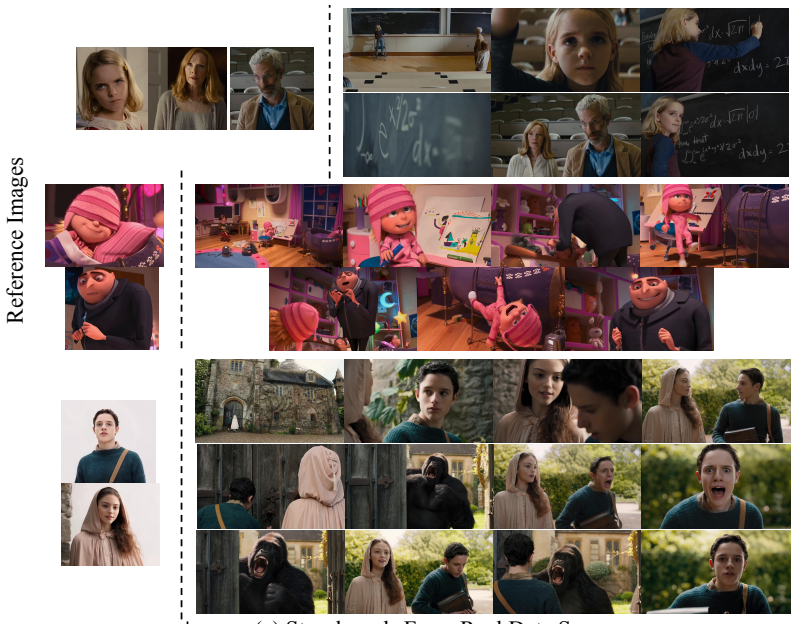
[4] Discus0434. Aesthetic predictor v2.5. https://github.com/discus0434/aesthetic-predictor-v2-5, 2024. 1

[5] Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025. 2, 4

[6] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025. 1

[7] Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025. 2

[8] Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. In *ICCV*, 2025. 3, 4



(a) Storyboards From Real Data Source

(b) Storyboards From Synthetic Data Source

Figure 7. Storyboard samples from different data sources.

[9] Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. *arXiv preprint arXiv:2412.21059*, 2024. 1

[10] Cailin Zhuang, Ailin Huang, Wei Cheng, Jingwei Wu, Yaoqi Hu, Jiaqi Liao, Hongyuan Wang, Xinyao Liao, Weiwei Cai, Hengyuan Xu, et al. Vistorybench: Comprehensive benchmark suite for story visualization. *arXiv preprint arXiv:2505.24862*, 2025. 3, 4

(a) Reference-to-Shot

(b) Text-to-Shot

(c) Shot-to-Shot

Figure 8. Quantitative results under Different Modes. (a) Reference-to-Shot Mode. Our method effectively preserves the character identity presented in the reference images. Moreover, for storyboards involving multiple reference roles, we demonstrates strong discriminative ability, with no noticeable confusion between different roles. (b) Text-to-Shot Mode. Our method responds well to prompts regarding shot transitions and viewpoint changes, while consistently maintaining scene coherence throughout the sequence. (c) Shot-to-Shot Mode. Our method maintains strong stylistic consistency with the previous shot and preserves the continuity of the appearing characters.