CI-VID: A Coherent Interleaved Text-Video Dataset

Supplementary Material

6. Usability Verification

Sequence Usability. To mitigate single-model bias, our pipeline adopts a dual-VLM cross-checking mechanism. We manually inspect 200 randomly sampled sequences and evaluate whether (1) the sequence maintains thematic and narrative coherence across clips and (2) inter-clip transitions are natural and semantically distinguishable. The error rate decreases from **14.8%** before cross-checking to **3.1%** after cross-checking.

Caption Usability. We further evaluate caption quality on 200 randomly sampled captions along two aspects: (1) main entity and scene description accuracy and (2) fine-grained detail accuracy (e.g., color, direction, and entity relations). We find that **92.6%** of captions correctly describe the main entities and scenes, while **74.1%** accurately capture fine-grained details, indicating room for improvement for GPT-4o-generated captions.

7. Multi-shot Video Generation.

Recent studies explore multi-shot video generation, aiming to produce temporally coherent videos composed of multiple scenes while maintaining narrative consistency. CineTrans [42] models long-range dependencies across multiple shots using a transformer-based framework. HoloCine [25] studies structured multi-shot generation by modeling shot-level relationships and global narrative structure. StoryDiffusion [47] focuses on maintaining character identity and visual consistency across sequential scenes. Related directions such as controllable video generation [40, 41, 46] also study temporal consistency across generated segments.

8. Comparison with CineTrans and HoloCine

CineTrans [42] and HoloCine [25] are recent works on multi-shot video generation. CineTrans proposes a transformer-based framework that leverages the diffusion model's capability to maintain temporal consistency across multiple shots without relying on additional training datasets. HoloCine focuses on structured multi-shot generation by modeling shot-level relationships and global narrative coherence, although the training data used by HoloCine has not been publicly released.

In contrast, CI-VID aims to provide a publicly available dataset and benchmark to facilitate research on text-and-video-to-video generation and multi-shot narrative modeling. Figure 7 presents a qualitative comparison on a six-scene example. The results show that a 0.6B model trained on CI-VID produces more coherent multi-shot transitions than CineTrans 1.3B, and achieves performance comparable to HoloCine 14B in this example.

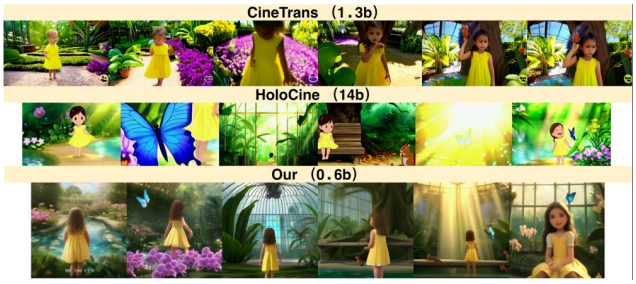

Figure 7. A case study comparing CineTrans and HoloCine on a six-scene generation example.