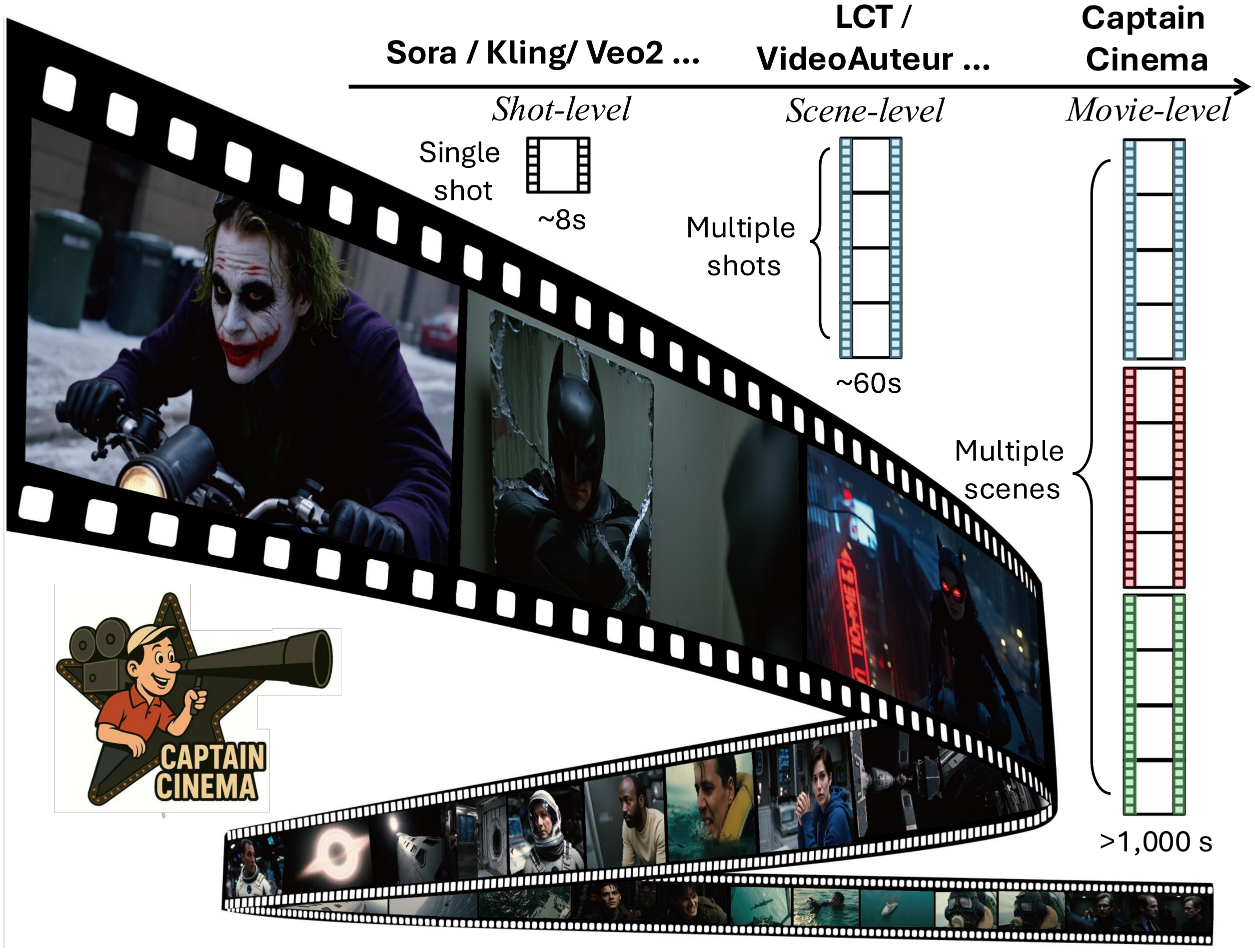

Figure 1 Captain Cinema: "I can film this all day!" Captain Cinema bridges top-down interleaved keyframe planning with bottom-up interleaved-conditioning video generation, taking a step toward the first multi-scene, whole-movie generation, preserving high visual consistency in scenes and identities. All the movie frames here are **generated**.

propose **Captain Cinema**, a framework tailored for story-driven movie synthesis.

`CaptainCinema` balances global plot structure with local visual fidelity through two complementary modules. A top-down planner first produces a sequence of key narrative frames that outline the storyboard, ensuring coherent high-level guidance. A bottom-up video synthesizer then interpolates full motion conditioned on these keyframes, maintaining both narrative flow and visual detail. Central to this design is `GoldenMem`, a memory mechanism that selectively retains and compresses contextual information from past keyframes. By summarizing long histories without exceeding memory budgets, `GoldenMem` preserves character and scene consistency across multiple acts, enabling scalable generation of multi-scene videos.

Additionally, we build a specialized data processing pipeline for processing long video data for movie generation and introduce progressive long-context tuning strategies tailored for Multimodal Diffusion Transformers (MM-DiT). These techniques enable stable and efficient fine-tuning on large-scale, long-form cinematic datasets, addressing the challenges of multi-scene video generation. Extensive experiments and ablation studies demonstrate that `Captain Cinema` not only achieves strong performance in long-form narrative video synthesis but also enables the automated creation of visually consistent short films that significantly exceed the duration of existing works, setting a new milestone in multimodal video generation capabilities.