CineMobile: On-Device Image-to-Video Diffusion for Cinematic Camera Motion Generation

Xuyao Huang[1,*], **Zelai Deng**[2,*], **Xu Wang**[1], **Xizhong Xiao**[3], **Zhijie Deng**[1]

[1]Shanghai Jiao Tong University, [2]Nankai University, [3]Transsion

The growing demand for image-to-video creation on mobile devices has increasingly focused on cinematic motion effects like *bullet time*, *dolly zoom*, *slow motion*, etc. While Diffusion Transformers (DiTs) exhibit strong performance in video generation, their large parameter sizes and multi-step iterative denoising processes lead to substantial computational overhead, making efficient generation on mobile devices challenging. We propose **CineMobile** to bridge the gap. In particular, CineMobile adopts a three-fold optimization strategy: (1) leveraging a distillation-guided pruning approach to derive a compact yet efficient model that retains the essential video generation capabilities required for cinematic effects; (2) optimizing the compressed model into a 4-step generator via a combination of diffusion distillation and reinforcement learning; (3) employing a hybrid post-training quantization strategy to compress the model footprint to under 1 GB. Experimental results show that compared to the teacher model with the Wan 2.1 architecture, CineMobile achieves a **40×** **speedup** in generation while maintaining comparable visual quality. Specifically, CineMobile generates 49-frame 480p videos with a per-step denoising latency of **0.6s** on an NVIDIA H200 GPU and **20s** on the MediaTek Dimensity 8400 Ultimate 5G platform, with a peak memory usage of **1.8 GB**, demonstrating its practical applicability for mobile-based image-to-video creation.

Correspondence: Zhijie Deng: zhijied@sjtu.edu.cn

Date: July 7, 2026

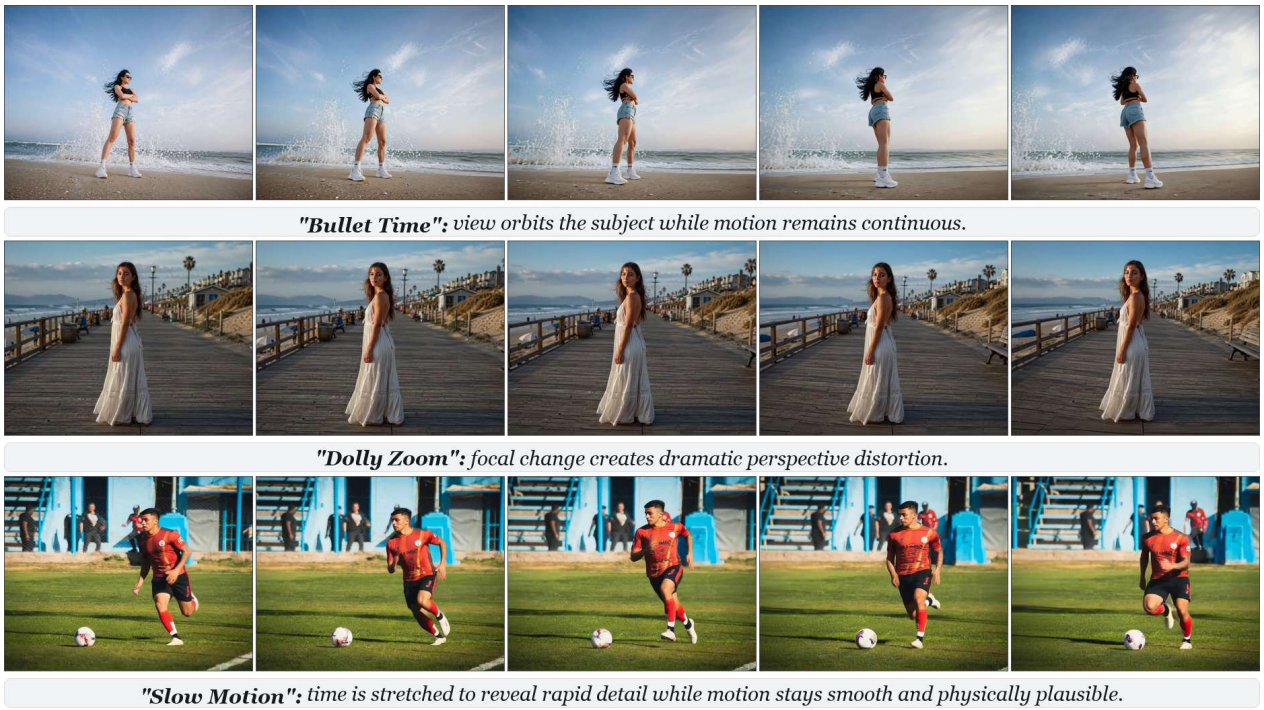

Figure 1 Bullet time, dolly zoom, and slow motion videos generated by CineMobile. CineMobile can produce continuous cinematic camera motion while preserving subject identity and scene consistency.

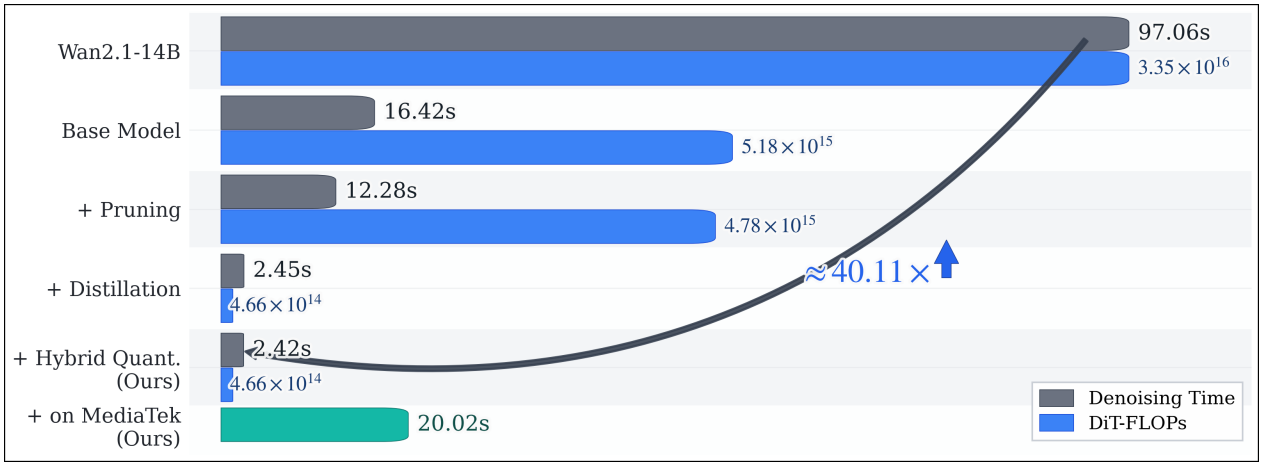

Figure 2 Denoising time and DiT-FLOPs comparison of CineMobile across different acceleration stages. Wan2.1-14B is used as the teacher model, requiring 97.06s denoising time and 3.35×10^{16} DiT-FLOPs. Starting from the base model Wan2.1-v1.1-Fun-1.5B, pruning reduces the model complexity and brings a moderate speedup, reducing the denoising time from 16.42s to 12.28s. Distillation further enables 4-step generation, leading to the main efficiency improvement with 2.45s denoising time. After applying hybrid precision weight quantization, CineMobile achieves **40.11**× faster DiT denoising and **71.89**× lower DiT-FLOPs than the teacher model. On the MediaTek Dimensity 8400 Ultimate 5G platform, CineMobile achieves a per-step denoising latency of 20.02s.

1 Introduction

The widespread adoption of smartphones for visual content creation is driving an increasing demand for mobile image-to-video (I2V) generation. Meanwhile, recent video generation models—such as Kling 2.5 Turbo (Team et al., 2025), Wan2.7 (Wan et al., 2025), Vidu Q3 (Bao et al., 2024), and Seedance 2.0 (Seedance et al., 2026)—have demonstrated remarkable capabilities in generating high-quality, realistic videos, with Diffusion Transformers (DiTs) emerging as the leading architecture underpinning these advances.

Despite their strong performance, DiTs suffer from huge parameter sizes and slow generation speed, limiting their practicality for cost-effective video creation (Kahatapitiya et al., 2025). For example, Wan2.1-I2V-14B (Wan et al., 2025) requires 240 seconds to generate a 49-frame video at 832 × 480 resolution with 40 inference steps on an NVIDIA H200 GPU, making mobile deployment infeasible. Furthermore, existing acceleration methods for deploying video generation models on mobile devices mainly focus on U-Net architectures (Ronneberger et al., 2015), leaving DiT models largely underexplored (Zhang et al., 2025a; Wu et al., 2025b). Considering that I2V models struggle to faithfully preserve prompt semantics under aggressive acceleration (Lv et al., 2025), we focus on camera motion effects with more structured and controllable motion targets.

This paper proposes **CineMobile** to enable DiT-based cinematic camera motion generation on mobile devices, based on a combination of structured pruning, step distillation, and hybrid precision quantization. Concretely, we first extend the Pluggable Pruning with Contiguous Layer Distillation (PPCL) (Ma et al., 2026) method, originally designed for text-to-image models, to prune image-to-video DiTs. Considering our prioritization of preserving fine portrait details and stabilizing camera control, we identify that it is better to perform structured depth pruning while maintaining the hidden dimension of the backbone unchanged.

After pruning, we introduce a supervised fine-tuning warm-up stage to restore the student's fundamental image-to-video generation capability and adapt it to the target camera-motion distribution. For step distillation, we adapt AdvDMD (Wang et al., 2026a), which integrates reinforcement learning (RL) (Li, 2018) into distribution matching distillation (DMD) (Yin et al., 2024a) for text-to-image generation, to I2V DiTs.

The involved RL process enables CineMobile to generate high-quality videos with a small number of denoising steps. To reduce the on-device memory cost, we adopt a hybrid precision quantization strategy. By quantizing FFN weights to 4-bit and the remaining components to 8-bit (Contributors, 2025), CineMobile reduces the

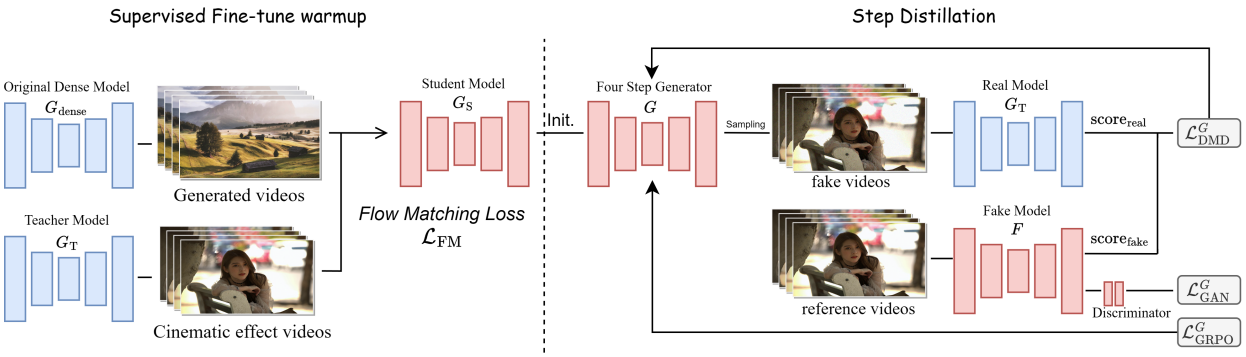

Figure 3 Overview of the step distillation pipeline. In the supervised fine-tuning warm-up, the student model is trained with the flow-matching objective on videos generated by the original dense model and cinematic-effect videos produced by the teacher model. The warm-up checkpoint then initializes a 4-step generator, which is further refined through adversarial distillation. In this stage, real model G_T and an online fake model F provide distribution-matching signals, while a lightweight discriminator supplies adversarial supervision and rewards for GRPO.

model memory footprint to under 1 GB while preserving visual quality.

To assess the capability of CineMobile in generating cinematic camera and temporal effects, we evaluate it on three representative cinematic motion effects: bullet time, dolly zoom, and slow motion. These effects cover distinct forms of cinematic motion control with clear motion patterns, making them suitable for studying controllable and efficient on-device I2V generation. We evaluate the generated videos on high-quality portrait and motion data using the VBench (Huang et al., 2024) evaluation protocol and human evaluation. As shown in Figure 2, with only 4 denoising steps and a model footprint of less than 1 GB, CineMobile achieves a 40× speedup over the teacher model in denoising time, and a per-step latency of 20s on the MediaTek Dimensity 8400 Ultimate 5G platform. In Table 1, following the VBench evaluation protocol, CineMobile maintains generation quality comparable to the teacher model across all three cinematic motion-effect scenarios, with total scores of 88.35, 89.30, and 88.05 on bullet time, dolly zoom, and slow motion, respectively, closely matching the teacher model's 89.27, 89.96, and 88.51.

2 Related Work

Video Diffusion Models. Video generation has rapidly evolved from early video diffusion and latent video models (Ho et al., 2022; He et al., 2023; Zhou et al., 2023) to stronger large-scale foundations such as Stable Video Diffusion (Blattmann et al., 2023), Lumiere (Bar-Tal et al., 2024), CogVideoX (Yang et al., 2025), Open-Sora (Zheng et al., 2024), LTX-Video (HaCohen et al., 2024), and Wan (Wan et al., 2025). These models improve visual fidelity, motion coherence, and scalability through advances in temporal modeling, latent compression, and backbone design (Wang et al., 2026b). From the perspective of task formulation, existing video diffusion models can be broadly organized into text-to-video generation (Gupta et al., 2024; Hong et al., 2023; Jin et al., 2025; Li et al., 2024; Hassan et al., 2026; Zhang et al., 2025b), image-to-video generation (Blattmann et al., 2023; Zhou et al., 2024; Shi et al., 2024; Ren et al., 2024a; Namekata et al., 2025), and specific motion-controlled generation (Ren et al., 2024b; Wu et al., 2024; Zeng et al., 2023; Zhao et al., 2023; Geng et al., 2025; Wang et al., 2025). Text-to-video models emphasize semantic alignment and open-ended video synthesis at scale (Singer et al., 2022; Li et al., 2023), while image-to-video models focus more on preserving subject identity, scene layout, and appearance consistency under temporal evolution (Guo et al., 2024; Shi et al., 2024). Despite this progress, most high-quality video diffusion models still depend on large backbones, server-class hardware, and multi-step sampling, limiting their practical deployment on edge devices (Zheng, 2025). This limitation becomes more critical in image-to-video settings with structured camera effects, where controllability and efficiency must be achieved simultaneously (Zheng, 2025; Shao et al., 2026; Wu et al., 2025b).

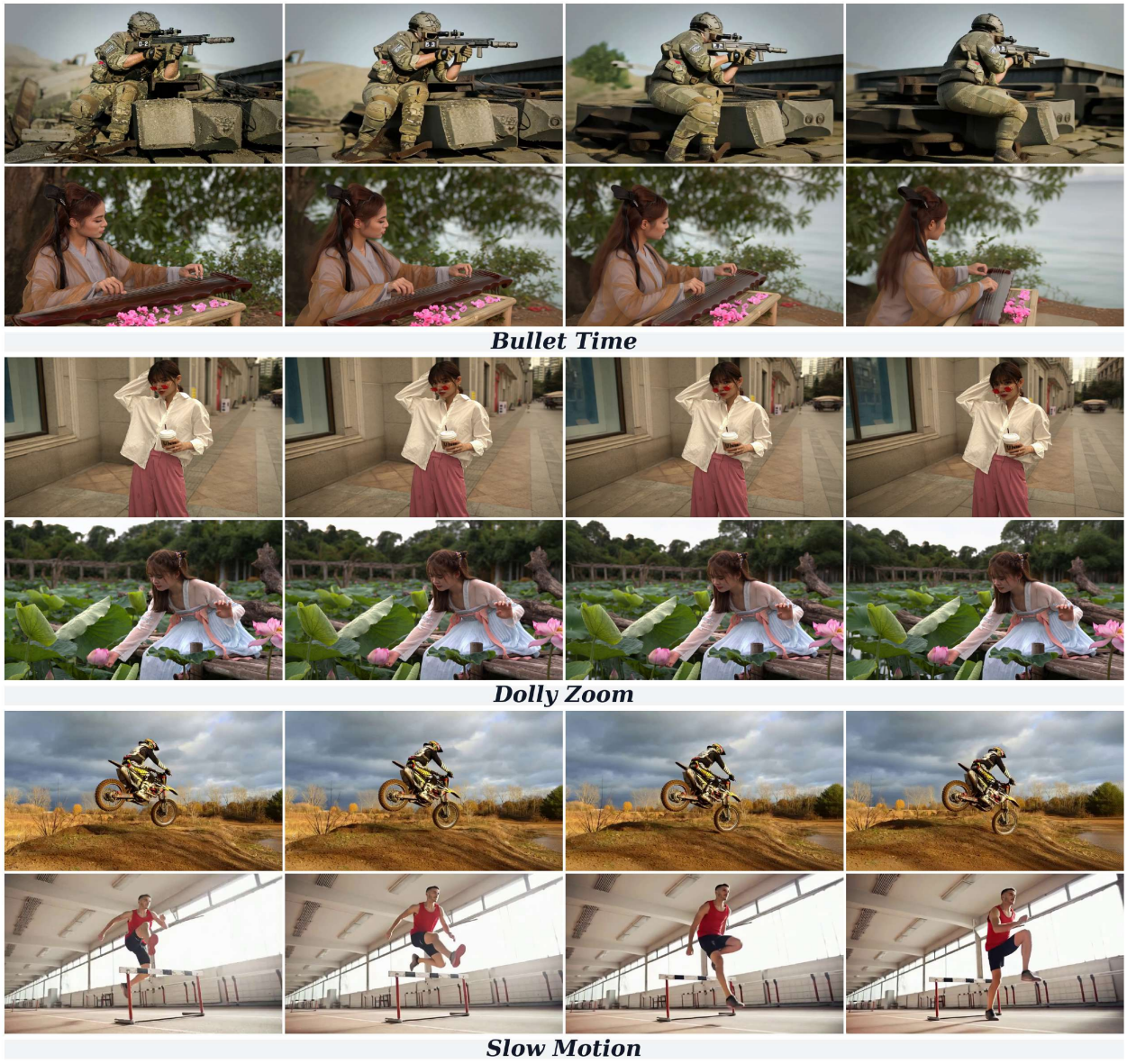

Figure 4 Qualitative examples of CineMobile across three cinematic camera effects.

On-Device Models. Recent work has started to bring video generation to mobile devices (Kim et al., 2025; Wu et al., 2025c; Zhang et al., 2025a; Yahia et al., 2024; Wu et al., 2025b). Mobile Video Diffusion compresses image-to-video generation through resolution reduction, temporal multi-scaling, pruning, and adversarial fine-tuning (Yahia et al., 2024). On-device Sora improves mobile deployment with denoising reduction, temporal token merging, and dynamic loading (Kim et al., 2025). SnapGen-V shows that interactive mobile generation can be achieved by combining an efficient spatial backbone, mobile-oriented temporal design, and adversarial low-step training (Wu et al., 2025c). Taming Diffusion Transformer for Efficient Mobile Video Generation in Seconds further extends this line to DiT-based models with highly compressed VAE design, sensitivity-aware pruning, and mobile-oriented step distillation (Wu et al., 2025b). These studies indicate that practical on-device video generation depends on joint optimization of architecture, compression, memory, and sampling.

Step Distillation. Reducing denoising steps is a central direction for efficient diffusion generation (Du et al., 2025; Salimans and Ho, 2022). Early acceleration methods include DDIM (Song et al., 2022) and Progressive Distillation (Salimans and Ho, 2022), followed by stronger few-step paradigms such as Latent Consistency

Models (Luo et al., 2023), Adversarial Diffusion Distillation (Sauer et al., 2024), DMD (Yin et al., 2024a), and DMD2 (Yin et al., 2024b). In video generation, VideoLCM (Wang et al., 2023), AnimateLCM (Wang et al., 2024), T2V-Turbo (Li et al., 2024), and T2V-Turbo-v2 (Li et al., 2025) adapt these ideas to the temporal setting and achieve competitive quality with only a few sampling steps. Recent mobile systems combine step distillation with hardware-aware model design: Mobile Video Diffusion approaches single-step inference through adversarial fine-tuning (Yahia et al., 2024; Kim et al., 2025), SnapGen-V reduces generation to 4 steps (Wu et al., 2025c), and Taming Diffusion Transformer develops adversarial step distillation for compressed mobile DiTs (Wu et al., 2025b). These works are especially relevant to our setting, where efficient on-device image-to-video generation requires few-step sampling together with aggressive model compression.

3 Method

Our objective is to build an I2V model that can efficiently generate different camera effect videos on mobile devices. To support lightweight deployment and diverse effects, CineMobile uses a shared backbone with effect-specific LoRA modules. To reduce the memory footprint, we first prune the shared DiT backbone in Section 3.1. After pruning, we restore the generative capability through a two-stage fine-tuning procedure, and then distill the model into a 4-step generator in Section 3.2. Finally, we obtain the final model through hybrid post-training quantization in Section 3.3.

3.1 Structured Depth Pruning

Directly pruning video DiTs can easily disrupt temporal modeling, leading to identity drift and severe degradation of fine-grained details, especially for cinematic motion generation that requires strict temporal consistency and camera trajectory smoothness. Given the strong performance of PPCL on text-to-image DiTs, we adapt it to video DiTs with specific designs for video temporal modeling and portrait detail preservation. PPCL (Ma et al., 2026) first detects redundant contiguous layer intervals in DiTs, and then compresses the model through depth-wise and width-wise pruning with distillation. Empirically, we find that width pruning can disturb the pretrained temporal representations of video DiTs by changing their internal feature dimensionality, as detailed in Appendix Section C. We adopt a structured depth pruning strategy. Specifically, our pruning procedure consists of two stages: redundant interval detection and interval distillation.

Redundant interval detection. We first train a residual linear probe for each transformer block to estimate whether its transformation can be approximated by a lightweight substitute. For a teacher video DiT with blocks $T = \{T_1, T_2, \ldots, T_M\}$, we denote the activation after block T_i as h_i^T. We use residual linear probes as lightweight diagnostics and CKA similarity (Kornblith et al., 2019) to rank candidate intervals by representation similarity and removable depth.

To adapt interval detection to video generation, we evaluate candidate spans on latent video sequences. We use more than 5,000 calibration samples to obtain stable span estimates that reflect redundancy under temporal motion and spatial appearance interactions.

Interval distillation. For each selected span $[u, v] \in \mathcal{P}$, we replace the whole span with one trainable surrogate transformer block D_u. It maintains the same block structure and hidden dimensionality as the base model blocks, and is initialized from the middle block $T_{\lfloor (u+v)/2 \rfloor}$ for stable training.

During distillation, the base model is frozen, and only the surrogate blocks are updated. We use portrait and motion data synthesized by the frozen base model (Wan et al., 2025). These samples cover facial identity and motion patterns, helping the surrogate blocks retain identity consistency and motion controllability during interval pruning. For each span, the surrogate block takes the teacher's hidden state at the span input and is trained to reproduce the teacher's hidden state at the span output through normalized hidden-state alignment. After training, we keep D_u at position u and remove blocks $T_{u+1}, \ldots, T_v$.

3.2 Step Distillation

We adopt a two-stage fine-tuning strategy for the pruned model. Directly applying step distillation to the pruned model can lead to unstable training and even generation collapse. Therefore, we first perform supervised fine-tuning as a warm-up stage, which restores the model's ability to reliably generate motion-effect videos with multi-step sampling. For step distillation, we build upon AdvDMD (Wang et al., 2026a) to improve the quality and stability of few-step generation. We extend AdvDMD to video DiTs and use it to distill our model as a 4-step I2V generator.

Supervised fine-tuning warm-up. We begin with a supervised fine-tuning warm-up to provide the compressed student with a stable initialization before adversarial optimization. This warm-up consists of two steps under the standard flow-matching objective (Duan et al., 2024). We start by fine-tuning the student model on videos generated by the original dense model to restore its fundamental video generation capability. We then use cinematic-effect videos generated by the teacher model to adapt the student model to the distributions of the target motion effects. Both steps share the same flow-matching training objective.

Adversarial distillation. After the supervised warm-up, we further refine the student through an adversarial distillation stage. The student is initialized from the warm-up checkpoint and updated with LoRA adapters. As illustrated in Figure 3, we use a frozen real score estimator G_T and an online fake score estimator F. G_T is initialized from the teacher model with motion-effect LoRA, and provides the teacher-induced score direction. F shares the student's base backbone, with randomly initialized LoRA adapters updated online to track the evolving student distribution. To provide adversarial realism feedback, we attach a lightweight discriminator head to the fake model F. Given an intermediate video latent and its associated I2V condition, F extracts multi-layer spatio-temporal DiT features, which are mapped by the discriminator heads to realism logits.

For each rollout batch, the 4-step generator G produces fake videos conditioned on the input frame. Reference videos are generated by the teacher model under the same conditions and serve as motion-effect targets. The frozen real model G_T and the online fake model F estimate the real and fake score directions on generated samples, and their discrepancy forms the DMD loss $\mathcal{L}^G_{\text{DMD}}$ for updating the generator. Meanwhile, the discriminator attached to F provides an adversarial loss $\mathcal{L}^G_{\text{GAN}}$, which encourages generated videos to be judged as reference-like. The discriminator scores on intermediate latents are further used as GRPO rewards, and $\mathcal{L}^G_{\text{GRPO}}$ is optimized to refine the 4-step trajectory.

We adopt an alternating training schedule: the generator is updated once every five steps, while the fake model and discriminator are updated in the remaining steps to maintain stable score estimation and realism feedback.

3.3 Hybrid Post-training Quantization

After obtaining a generator capable of producing videos in 4 denoising steps, we apply a hybrid post-training quantization strategy to further reduce the runtime memory footprint and latency for on-device generation (Contributors, 2025). Instead of assigning a uniform precision to all layers, we adopt a layer-wise hybrid quantization scheme for the DiT backbone. We keep activations in 16-bit precision for numerical stability and use FP8 weights for most quantized linear layers, including attention projections, text-embedding projections, and time-embedding projections. Since feed-forward networks contain a large fraction of the DiT parameters, the two linear projections in each block use 4-bit weights with 16-bit activations.

4 Experiments

In this section, we systematically evaluate the efficiency and performance of CineMobile on three cinematic motion effects.

4.1 Setup

Training Details. For the initial pruning stage, we use more than 5,000 samples to estimate layer redundancy and identify redundant block intervals. Based on this analysis, we prune the original DiT backbone to obtain a

Table 1 Systematic comparison of video generation models on VBench with different parameter sizes and inference steps, including representative publicly available models such as HunyuanVideo-1.5 (Wu et al., 2025a), LTX2.3 (HaCohen et al., 2026), and Open-Sora2.0 (Zheng et al., 2026). All models are evaluated on the same test set of 200 portrait images with five videos generated for each image, following the official VBench protocol to compare video generation quality. The General panel reports representative publicly available video generation models. Wan2.1 is used as the teacher model and Wan2.1-v1.1-Fun as the base model; Ours-Prune, Ours-SFT, and Ours correspond to the pruned-and-finetuned model, the second-stage SFT model, and the final distilled and hybrid-PTQ accelerated model, respectively. With only 1.2B parameters and 4 denoising steps, Ours achieves VBench scores close to the teacher model, demonstrating comparable generation quality with substantially higher efficiency.

Model	Steps	Params (B)	Subject Consis.	Background Consis.	Motion Smooth.	Aesthetic Quality	Imaging Quality	I2V Subject	I2V Background	Total Score
General										
HunyuanVideo-1.5	12	8.3	98.46	96.49	99.61	64.57	70.78	99.39	99.49	89.83
LTX2.3	12	22	98.39	96.08	99.38	63.06	71.54	98.87	98.60	89.42
LTX-2	60	13	98.21	95.78	99.01	62.86	69.31	98.83	98.42	88.92
Open-Sora2.0	50	11	97.35	94.23	98.46	63.03	68.46	99.48	98.76	88.54
Wan2.1-v1.1-Fun	20	1.5	94.12	94.78	99.02	66.72	69.55	98.18	99.31	88.81
Ours-Prune	20	1.2	93.92	94.55	98.96	65.25	68.35	98.10	99.22	88.34
Bullet Time										
Wan2.1	20	14	95.33	95.79	99.09	66.65	70.56	98.16	99.30	89.27
Ours-SFT	20	1.2	94.01	94.86	99.00	66.85	69.29	98.15	99.34	88.79
Ours	4	1.2	93.46	94.06	98.76	64.76	70.71	98.15	98.58	88.35
Dolly Zoom										
Wan2.1	20	14	98.56	97.82	99.36	65.45	70.45	98.54	99.52	89.96
Ours-SFT	20	1.2	98.22	96.18	99.11	63.90	69.22	98.21	99.12	89.14
Ours	4	1.2	98.19	95.96	99.28	64.18	69.84	98.30	99.33	89.30
Slow Motion										
Wan2.1	20	14	96.48	96.01	99.20	63.42	68.15	97.36	98.97	88.51
Ours-SFT	20	1.2	96.65	95.57	98.95	62.29	67.89	97.55	99.25	88.31
Ours	4	1.2	95.82	94.98	99.03	61.52	68.75	97.28	99.00	88.05

27-layer student model, followed by 8k training steps to restore its generation capability. For the subsequent supervised fine-tuning warm-up, we follow the training implementation of DiffSynth-Studio (ModelScope Team, 2026) and train the student model for 5k steps. We use a learning rate of 1×10^{-4} and set the LoRA rank to 128. During distillation, we use over 10K training samples to optimize the student model, which follows a 4-step denoising trajectory. The student is trained with a learning rate of 1×10^{-4}. We set $\lambda_{\text{adv}} = 0.5$ and $\lambda_{\text{D}} = 0.005$, and train the discriminator head and fake model with a learning rate of 1×10^{-5}. All experiments are conducted on 8 NVIDIA H200 GPUs.

Model & Datasets To obtain reliable supervision for cinematic motion effect generation, we build teacher models by combining a shared Wan2.1-I2V-14B image-to-video backbone with effect-specific LoRA adapters. For bullet time and dolly zoom, we use publicly available LoRA adapters tailored to these effects. For slow motion, we construct the adapter by fine-tuning on slow motion videos derived from motion-intensive base-model outputs through 2× frame interpolation (Reda et al., 2022). For the student model, we adopt a Wan2.1-v1.1-Fun-I2V model (Wan et al., 2025) with 1.5B parameters as the base model. For data collection, we use the high-quality portrait dataset PPR10K (Liang et al., 2021) as the image source for bullet time and dolly zoom. For slow motion, we collect motion-oriented images from public platforms such as Pexels and Pixabay. Each effect-specific training set contains over 10K samples.

Baselines & Metrics We conduct a systematic evaluation of CineMobile in terms of video generation performance and efficiency. For generation performance, we combine VBench (Huang et al., 2024) and human evaluation to assess the generated videos from complementary perspectives. Specifically, we use VBench to

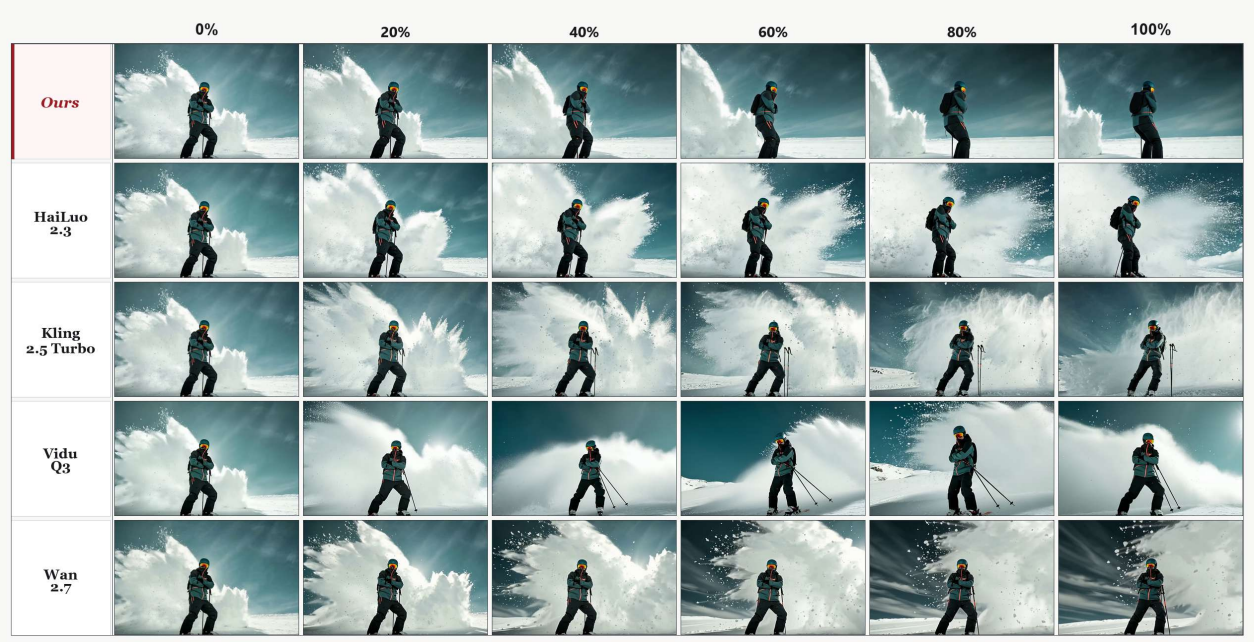

Figure 5 Qualitative comparison with commercial video generation models on cinematic motion effects. Frames are uniformly sampled at different temporal ratios from generated videos. Our method shows more stable camera progression and stronger subject-background consistency.

evaluate general video quality and compare CineMobile with representative open-source video generation models, including HunyuanVideo-1.5 (Wu et al., 2025a), LTX2.3 (HaCohen et al., 2026), and Open-Sora2.0 (Zheng et al., 2026). Since cinematic motion effects require effect-specific motion patterns that are not fully captured by standard automated metrics, we further conduct a human evaluation. Specifically, we collect 50 video samples of bullet time and ask 10 participants to evaluate them. All videos are anonymized, randomly shuffled, and presented without revealing the identity of the corresponding generation model. In this evaluation, CineMobile is compared with strong commercial video generation systems, including Kling, Hailuo, and Vidu, based on human preference scores. For efficiency, we analyze the reductions in denoising time and FLOPs achieved by CineMobile across different acceleration stages.

Deployment Platform We evaluate on-device deployment on an Infinix NOTE 60 Ultra equipped with the MediaTek Dimensity 8400 Ultimate 5G platform Infinix Mobility. This SoC integrates an all-big-core Arm Cortex-A725 CPU, an Arm Mali-G720 MC7/MP7 GPU, and a MediaTek NPU 880 MediaTek. The Mali-G720 MP7 GPU has a reported theoretical FP32 throughput of 2329.6 GFLOPS, approximately 2.33 TFLOPS NanoReview. For reference, this theoretical FP32 peak is close to that of an NVIDIA GeForce GTX 960, whose FP32 throughput can be estimated as approximately 2.41 TFLOPS from its official CUDA core count and boost clock NVIDIA. All mobile latency and memory measurements are conducted on this device unless otherwise specified.

4.2 Main Results

Quantitative analysis. We evaluate CineMobile with VBench (Huang et al., 2024) under three cinematic scenarios. As shown in Table 1, CineMobile achieves video quality comparable to the teacher model and other larger models, despite using only 4 denoising steps and a compact model size. Compared with the 20-step Wan2.1-I2V-14B teacher model, CineMobile uses only 10% of the parameters and 20% of the denoising steps, while limiting the absolute total-score gap to 0.92 on bullet time, 0.66 on dolly zoom, and 0.46 on slow motion. A closer look at individual metrics shows that CineMobile preserves motion smoothness and I2V consistency particularly well, suggesting that the 4-step student maintains reliable temporal dynamics and image-condition alignment. Overall, these results demonstrate that CineMobile substantially reduces

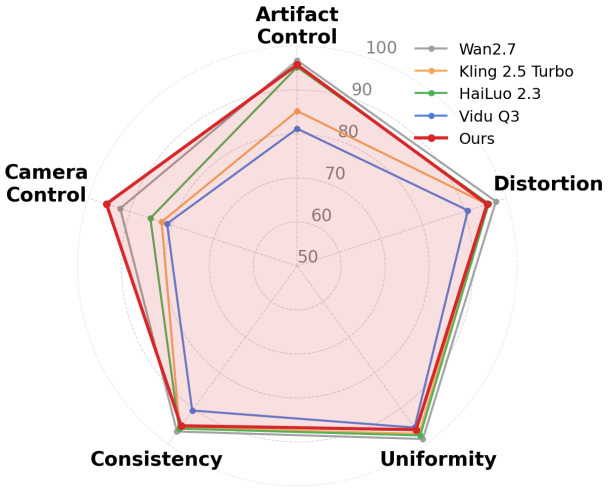

Figure 6 Human evaluation of bullet time video generation against representative commercial models. CineMobile shows balanced quality with a clear advantage in camera control, while remaining competitive in overall quality.

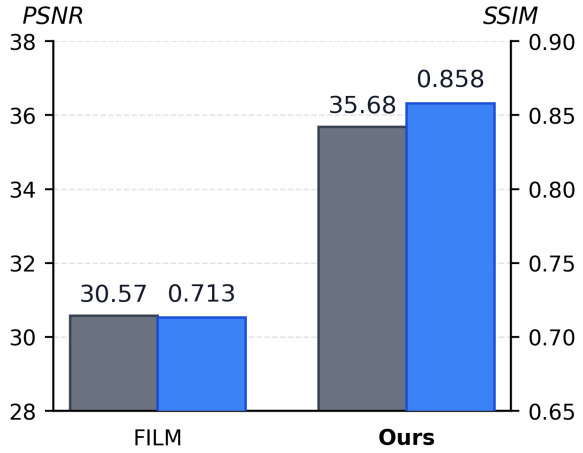

Figure 7 Comparison with FILM (Reda et al., 2022) on 20× slow motion synthesis. Given the first and last frames, CineMobile achieves consistently higher PSNR and SSIM than FILM, showing better reconstruction over large temporal gaps.

inference cost while retaining competitive cinematic video generation quality.

Qualitative analysis. Figure 4 and Figure 5 presents a qualitative comparison using uniformly sampled frames across the generated video. Our method produces smoother camera progression and more coherent temporal evolution throughout the sequence. By contrast, several commercial models exhibit stronger temporal fluctuation, less stable subject scale, or less consistent background motion, especially at later temporal ratios.

We further compare CineMobile with commercial models through human evaluation, as shown in Figure 6. Overall, CineMobile performs competitively across multiple perceptual dimensions, including artifact control, distortion, uniformity, and temporal consistency. It also achieves the strongest result in camera control, showing its advantage in controllable bullet-time motion. These results indicate that CineMobile remains on par with representative commercial models in overall perceptual quality, while providing better camera controllability.

Overall, our model remains competitive across multiple perceptual dimensions, including artifact control, distortion, uniformity, and temporal consistency. Most notably, CineMobile achieves the best score on camera control, which is consistent with our objective of optimizing cinematic motion effects rather than only static image quality. Although some commercial models obtain slightly stronger scores on individual dimensions, our method provides a favorable overall trade-off between controllable motion and perceptual quality. This is particularly notable given that CineMobile is designed for efficient on-device deployment rather than cloud-scale generation.

Efficiency analysis. We further analyze the efficiency of CineMobile across different acceleration stages in Figure 2. Structured pruning decreases the runtime to 12.28 s by removing redundant DiT blocks. The most significant gain comes from step distillation, which reduces the denoising time to 2.45 s and the DiT-FLOPs to 4.66×10^{14}, showing that reducing the number of denoising steps is the key factor for acceleration. Hybrid precision quantization brings a slight additional runtime improvement and mainly improves deployment efficiency by reducing memory cost. Overall, CineMobile achieves 40.11× faster DiT denoising and 71.89× lower DiT-FLOPs than the teacher.

Table 2 Ablation study of GRPO on bullet time generation. We compare models trained with and without GRPO on both the base model and CineMobile under the same sampling setting. GRPO benefits the distillation process and improves aesthetic quality and imaging quality.

Model	Subject Consis.	Background Consis.	Motion Smooth.	Aesthetic Quality	Imaging Quality	I2V Subject	I2V Background	Total Score	Δ Total
Base model (w/o GRPO)	94.46	94.91	98.84	64.18	68.69	98.11	98.58	88.25	–
Base model (w/ GRPO)	**95.01**	94.36	**99.04**	**66.45**	**69.29**	**98.15**	**99.10**	**88.77**	**+0.52**
CineMobile (w/o GRPO)	93.54	93.21	**99.00**	64.35	68.90	97.36	**98.77**	87.88	–
CineMobile (w/ GRPO)	93.46	**94.06**	98.76	**64.76**	**70.71**	**98.15**	98.58	**88.35**	**+0.47**

4.3 Ablation & Discussion

Effect of GRPO refinement. We evaluate the effect of GRPO under the same 4-step sampling setting. As shown in Table 2, GRPO improves the total score for both the base model and CineMobile, increasing the score by +0.52 and +0.47, respectively. The gains mainly come from perceptual-related metrics, including aesthetic quality and imaging quality, while other consistency metrics remain largely comparable. This trend suggests that GRPO helps refine the distilled generator by providing additional reward guidance beyond supervised distillation. Overall, GRPO improves the perceptual quality of 4-step generation without noticeably degrading image-conditioned consistency.

Comparison with frame interpolation. We compare CineMobile with FILM (Reda et al., 2022) under a 20× slow-motion setting, where only the first and last frames are provided as inputs. As shown in Figure 7, CineMobile achieves higher reconstruction quality, improving PSNR from 30.57 dB to 35.68 dB and SSIM from 0.713 to 0.858. This suggests that directly synthesizing slow-motion dynamics is more effective under large temporal gaps than conventional frame interpolation. We also observe that FILM can suffer from object interpenetration and motion inconsistency in this setting, while CineMobile produces more coherent intermediate motion.

5 Conclusion

In this paper, we presented CineMobile, an efficient on-device image-to-video generation framework for cinematic motion effects. Experiments on bullet time, dolly zoom, and slow motion demonstrate that CineMobile substantially improves inference efficiency while maintaining competitive visual quality and temporal consistency relative to the teacher model, and showing favorable camera-motion control against commercial references in human evaluation. Overall, CineMobile offers a new and practical solution for on-device deployment of image-to-video generation models. In future work, CineMobile can be further extended to a broader range of video generation models and cinematic effect generation tasks.

References

Fan Bao, Chendong Xiang, Gang Yue, Guande He, Hongzhou Zhu, Kaiwen Zheng, Min Zhao, Shilong Liu, Yaole Wang, and Jun Zhu. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models, 2024. URL https://arxiv.org/abs/2405.04233.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, SA '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400711312. doi: 10.1145/3680528.3687614. URL https://doi.org/10.1145/3680528.3687614.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. URL https://arxiv.org/abs/2311.15127.

LightX2V Contributors. Lightx2v: Light video generation inference framework. https://github.com/ModelTC/lightx2v, 2025.

Zhenbang Du, Yonggan Fu, Lifu Wang, Jiayi Qian, Xiao Luo, and Yingyan Celine Lin. Fewer denoising steps or cheaper per-step inference: Towards compute-optimal diffusion model deployment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3001–3010, October 2025.

Zhongjie Duan, Lizhou You, Chengyu Wang, Cen Chen, Ziheng Wu, Weining Qian, and Jun Huang. Diffsynth: Latent in-iteration deflickering for realistic video synthesis. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track: European Conference, ECML PKDD 2024, Vilnius, Lithuania, September 9–13, 2024, Proceedings, Part X*, page 332–347, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-70380-5. doi: 10.1007/978-3-031-70381-2_21. URL https://doi.org/10.1007/978-3-031-70381-2_21.

Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Yusuf Aytar, Michael Rubinstein, Chen Sun, Oliver Wang, Andrew Owens, and Deqing Sun. Motion prompting: Controlling video generation with motion trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–12, June 2025.

Xun Guo, Mingwu Zheng, Liang Hou, Yuan Gao, Yufan Deng, Pengfei Wan, Di Zhang, Yufan Liu, Weiming Hu, Zhengjun Zha, Haibin Huang, and Chongyang Ma. I2v-adapter: A general image-to-video adapter for diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, SIGGRAPH '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705250. doi: 10.1145/3641519.3657407. URL https://doi.org/10.1145/3641519.3657407.

Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Fei-Fei Li, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXIX*, page 393–411, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-72985-0. doi: 10.1007/978-3-031-72986-7_23. URL https://doi.org/10.1007/978-3-031-72986-7_23.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion, 2024. URL https://arxiv.org/abs/2501.00103.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. Ltx-2: Efficient joint audio-visual foundation model, 2026. URL https://arxiv.org/abs/2601.03233.

Mariam Hassan, Bastien Van Delft, Wuyang Li, and Alexandre Alahi. Anchored video generation: Decoupling scene construction and temporal synthesis in text-to-video diffusion models, 2026. URL https://arxiv.org/abs/2512.16371.

Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation, 2023. URL https://arxiv.org/abs/2211.13221.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David Fleet. Video diffusion models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 8633–8646. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/39235c56aef13fb05a6adc95eb9d8d66-Paper-Conference.pdf.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=rB6TpjAuSRy.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21807–21818, June 2024.

Infinix Mobility. Infinix NOTE 60 Ultra Specifications. URL https://wap.infinixmobility.com/specs/note-60-ultra. Accessed: 2026-05-06.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong MU, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *International Conference on Learning Representations*, volume 2025, pages 23378–23402, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/3ab228c4703c4459b1a600ebadc5732c-Paper-Conference.pdf.

Kumara Kahatapitiya, Haozhe Liu, Sen He, Ding Liu, Menglin Jia, Chenyang Zhang, Michael S. Ryoo, and Tian Xie. Adaptive caching for faster video generation with diffusion transformers. In *Proceedings of the IEEE/CVF*

International Conference on Computer Vision (ICCV), pages 15240–15252, October 2025.

Bosung Kim, Kyuhwan Lee, Isu Jeong, Jungmin Cheon, Yeojin Lee, and Seulki Lee. On-device sora: Enabling training-free diffusion-based text-to-video generation for mobile devices, 2025. URL https://arxiv.org/abs/2502.04363.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Jiachen Li, Weixi Feng, Tsu-Jui Fu, Xinyi Wang, Sugato Basu, Wenhu Chen, and William Yang Wang. T2v-turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 75692–75726. Curran Associates, Inc., 2024. doi: 10.52202/079017-2410. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/8a57aa8e8b57e64a42e95f7dceb0adb9-Paper-Conference.pdf.

Jiachen Li, Qian Long, Jian (Skyler) Zheng, Xiaofeng Gao, Robinson Piramuthu, Wenhu Chen, and William Wang. T2v-turbo-v2: Enhancing video model post-training through data, reward, and conditional guidance design. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *International Conference on Learning Representations*, volume 2025, pages 92279–92305, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/e68af7d8a44bc1964f6be4de464e38f9-Paper-Conference.pdf.

Xin Li, Wenqing Chu, Ye Wu, Weihang Yuan, Fanglong Liu, Qi Zhang, Fu Li, Haocheng Feng, Errui Ding, and Jingdong Wang. Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation, 2023. URL https://arxiv.org/abs/2309.00398.

Yuxi Li. Deep reinforcement learning: An overview, 2018. URL https://arxiv.org/abs/1701.07274.

Jie Liang, Hui Zeng, Miaomiao Cui, Xuansong Xie, and Lei Zhang. Ppr10k: A large-scale portrait photo retouching dataset with human-region mask and group-level consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference, 2023. URL https://arxiv.org/abs/2310.04378.

Zhengyao Lv, Chenyang Si, Tianlin Pan, Zhaoxi Chen, Kwan-Yee K. Wong, Yu Qiao, and Ziwei Liu. Dual-expert consistency model for efficient and high-quality video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14983–14993, October 2025.

Jian Ma, Qirong Peng, Xujie Zhu, Peixing Xie, Chen Chen, and Haonan Lu. Pluggable pruning with contiguous layer distillation for diffusion transformers, 2026. URL https://arxiv.org/abs/2511.16156.

MediaTek. MediaTek Dimensity 8400. URL https://www.mediatek.com/products/smartphones/mediatek-dimensity-8400. Accessed: May 6, 2026.

ModelScope Team. DiffSynth-Studio: Enjoy the magic of diffusion models, 2026. URL https://github.com/modelscope/DiffSynth-Studio. Accessed: May 7, 2026.

Koichi Namekata, Sherwin Bahmani, Ziyi Wu, Yash Kant, Igor Gilitschenski, and David Lindell. Sg-i2v: Self-guided trajectory control in image-to-video generation. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *International Conference on Learning Representations*, volume 2025, pages 38483–38505, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/5fba70900a84a8fb755c48ba99420c95-Paper-Conference.pdf.

NanoReview. MediaTek Dimensity 8400: Specs and Benchmarks. https://nanoreview.net/en/soc/mediatek-dimensity-8400. Accessed: May 6, 2026.

NVIDIA. GeForce GTX 900 Series Graphics Cards. https://www.nvidia.com/en-us/geforce/900-series/. Accessed: 2026-05-06.

Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. Film: Frame interpolation for large motion. In *European Conference on Computer Vision (ECCV)*, 2022.

Weiming Ren, Huan Yang, Ge Zhang, Cong Wei, Xinrun Du, Wenhao Huang, and Wenhu Chen. Consisti2v: Enhancing visual consistency for image-to-video generation. *Transactions on Machine Learning Research*, 2024a. ISSN 2835-8856. URL https://openreview.net/forum?id=vqniLmUDvj.

Yixuan Ren, Yang Zhou, Jimei Yang, Jing Shi, Difan Liu, Feng Liu, Mingi Kwon, and Abhinav Shrivastava. Customize-a-video: One-shot motion customization of text-to-video diffusion models. In *European Conference on Computer Vision*, pages 332–349. Springer, 2024b.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmen-

tation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TIdIXIpzhoI.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXXVI*, page 87–103, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-73015-3. doi: 10.1007/978-3-031-73016-0_6. URL https://doi.org/10.1007/978-3-031-73016-0_6.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, Mojie Chi, Xuyan Chi, Jian Cong, Qinpeng Cui, Fei Ding, Qide Dong, Yujiao Du, Haojie Duanmu, Junliang Fan, Jiarui Fang, Jing Fang, Zetao Fang, Chengjian Feng, Yu Gao, Diandian Gu, Dong Guo, Hanzhong Guo, Qiushan Guo, Boyang Hao, Hongxiang Hao, Haoxun He, Jiaao He, Qian He, Tuyen Hoang, Heng Hu, Ruoqing Hu, Yuxiang Hu, Jiancheng Huang, Weilin Huang, Zhaoyang Huang, Zhongyi Huang, Jishuo Jin, Ming Jing, Ashley Kim, Shanshan Lao, Yichong Leng, Bingchuan Li, Gen Li, Haifeng Li, Huixia Li, Jiashi Li, Ming Li, Xiaojie Li, Xingxing Li, Yameng Li, Yiying Li, Yu Li, Yueyan Li, Chao Liang, Han Liang, Jianzhong Liang, Ying Liang, Wang Liao, J. H. Lien, Shanchuan Lin, Xi Lin, Feng Ling, Yue Ling, Fangfang Liu, Jiawei Liu, Jihao Liu, Jingtuo Liu, Shu Liu, Sichao Liu, Wei Liu, Xue Liu, Zuxi Liu, Ruijie Lu, Lecheng Lyu, Jingting Ma, Tianxiang Ma, Xiaonan Nie, Jingzhe Ning, Junjie Pan, Xitong Pan, Ronggui Peng, Xueqiong Qu, Yuxi Ren, Yuchen Shen, Guang Shi, Lei Shi, Yinglong Song, Fan Sun, Li Sun, Renfei Sun, Wenjing Tang, Boyang Tao, Zirui Tao, Dongliang Wang, Feng Wang, Hulin Wang, Ke Wang, Qingyi Wang, Rui Wang, Shuai Wang, Shulei Wang, Weichen Wang, Xuanda Wang, Yanhui Wang, Yue Wang, Yuping Wang, Yuxuan Wang, Zijie Wang, Ziyu Wang, Guoqiang Wei, Meng Wei, Di Wu, Guohong Wu, Hanjie Wu, Huachao Wu, Jian Wu, Jie Wu, Ruolan Wu, Shaojin Wu, Xiaohu Wu, Xinglong Wu, Yonghui Wu, Ruiqi Xia, Xin Xia, Xuefeng Xiao, Shuang Xu, Bangbang Yang, Jiaqi Yang, Runkai Yang, Tao Yang, Yihang Yang, Zhixian Yang, Ziyan Yang, Fulong Ye, Bingqian Yi, Xing Yin, Yongbin You, Linxiao Yuan, Weihong Zeng, Xuejiao Zeng, Yan Zeng, Siyu Zhai, Zhonghua Zhai, Bowen Zhang, Chenlin Zhang, Heng Zhang, Jun Zhang, Manlin Zhang, Peiyuan Zhang, Shuo Zhang, Xiaohe Zhang, Xiaoying Zhang, Xinyan Zhang, Xinyi Zhang, Yichi Zhang, Zixiang Zhang, Haiyu Zhao, Huating Zhao, Liming Zhao, Yian Zhao, Guangcong Zheng, Jianbin Zheng, Xiaozheng Zheng, Zerong Zheng, Kuan Zhu, and Feilong Zuo. Seedance 2.0: Advancing video generation for world complexity, 2026. URL https://arxiv.org/abs/2604.14148.

Shitong Shao, Lichen Bai, Pengfei Wan, James Kwok, and Zeke Xie. Efficient video diffusion models: Advancements and challenges, 2026. URL https://arxiv.org/abs/2604.15911.

Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In *ACM SIGGRAPH 2024 Conference Papers*, SIGGRAPH '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705250. doi: 10.1145/3641519.3657497. URL https://doi.org/10.1145/3641519.3657497.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data, 2022. URL https://arxiv.org/abs/2209.14792.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. URL https://arxiv.org/abs/2010.02502.

Kling Team, Jialu Chen, Yuanzheng Ci, Xiangyu Du, Zipeng Feng, Kun Gai, Sainan Guo, Feng Han, Jingbin He, Kang He, Xiao Hu, Xiaohua Hu, Boyuan Jiang, Fangyuan Kong, Hang Li, Jie Li, Qingyu Li, Shen Li, Xiaohan Li, Yan Li, Jiajun Liang, Borui Liao, Yiqiao Liao, Weihong Lin, Quande Liu, Xiaokun Liu, Yilun Liu, Yuliang Liu, Shun Lu, Hangyu Mao, Yunyao Mao, Haodong Ouyang, Wenyu Qin, Wanqi Shi, Xiaoyu Shi, Lianghao Su, Haozhi Sun, Peiqin Sun, Pengfei Wan, Chao Wang, Chenyu Wang, Meng Wang, Qiulin Wang, Runqi Wang, Xintao Wang, Xuebo Wang, Zekun Wang, Min Wei, Tiancheng Wen, Guohao Wu, Xiaoshi Wu, Zhenhua Wu, Da Xie, Yingtong Xiong, Yulong Xu, Sile Yang, Zikang Yang, Weicai Ye, Ziyang Yuan, Shenglong Zhang, Shuaiyu Zhang, Yuanxing Zhang, Yufan Zhang, Wenzheng Zhao, Ruiliang Zhou, Yan Zhou, Guosheng Zhu, and Yongjie Zhu. Kling-omni technical report, 2025. URL https://arxiv.org/abs/2512.16776.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang,

Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

Angtian Wang, Haibin Huang, Jacob Zhiyuan Fang, Yiding Yang, and Chongyang Ma. Ati: Any trajectory instruction for controllable video generation, 2025. URL https://arxiv.org/abs/2505.22944.

Fu-Yun Wang, Zhaoyang Huang, Weikang Bian, Xiaoyu Shi, Keqiang Sun, Guanglu Song, Yu Liu, and Hongsheng Li. Animatelcm: Computation-efficient personalized style video generation without personalized video data, 2024. URL https://arxiv.org/abs/2402.00769.

Xiang Wang, Shiwei Zhang, Han Zhang, Yu Liu, Yingya Zhang, Changxin Gao, and Nong Sang. Videolcm: Video latent consistency model, 2023. URL https://arxiv.org/abs/2312.09109.

Xu Wang, Zexian Li, Litong Gong, Tiezheng Ge, and Zhijie Deng. Advdmd: Adversarial reward meets dmd for high-quality few-step generation, 2026a. URL https://arxiv.org/abs/2604.28126.

Yimu Wang, Xuye Liu, Wei Pang, Li Ma, Shuai Yuan, Paul Debevec, and Ning Yu. Survey of video diffusion models: Foundations, implementations, and applications, 2026b. URL https://arxiv.org/abs/2504.16081.

Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, Linus, Patrol, Peizhen Zhang, Peng Chen, Penghao Zhao, Qi Tian, Songtao Liu, Weijie Kong, Weiyan Wang, Xiao He, Xin Li, Xinchi Deng, Xuefei Zhe, Yang Li, Yanxin Long, Yuanbo Peng, Yue Wu, Yuhong Liu, Zhenyu Wang, Zuozhuo Dai, Bo Peng, Coopers Li, Gu Gong, Guojian Xiao, Jiahe Tian, Jiaxin Lin, Jie Liu, Jihong Zhang, Jiesong Lian, Kaihang Pan, Lei Wang, Lin Niu, Mingtao Chen, Mingyang Chen, Mingzhe Zheng, Miles Yang, Qiangqiang Hu, Qi Yang, Qiuyong Xiao, Runzhou Wu, Ryan Xu, Rui Yuan, Shanshan Sang, Shisheng Huang, Siruis Gong, Shuo Huang, Weiting Guo, Xiang Yuan, Xiaojia Chen, Xiawei Hu, Wenzhi Sun, Xiele Wu, Xianshun Ren, Xiaoyan Yuan, Xiaoyue Mi, Yepeng Zhang, Yifu Sun, Yiting Lu, Yitong Li, You Huang, Yu Tang, Yixuan Li, Yuhang Deng, Yuan Zhou, Zhichao Hu, Zhiguang Liu, Zhihe Yang, Zilin Yang, Zhenzhi Lu, Zixiang Zhou, and Zhao Zhong. Hunyuanvideo 1.5 technical report, 2025a. URL https://arxiv.org/abs/2511.18870.

Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. Motionbooth: Motion-aware customized text-to-video generation, 2024. URL https://arxiv.org/abs/2406.17758.

Yushu Wu, Yanyu Li, Anil Kag, Ivan Skorokhodov, Willi Menapace, Ke Ma, Arpit Sahni, Ju Hu, Aliaksandr Siarohin, Dhritiman Sagar, Yanzhi Wang, and Sergey Tulyakov. Taming diffusion transformer for efficient mobile video generation in seconds, 2025b. URL https://arxiv.org/abs/2507.13343.

Yushu Wu, Zhixing Zhang, Yanyu Li, Yanwu Xu, Anil Kag, Yang Sui, Huseyin Coskun, Ke Ma, Aleksei Lebedev, Ju Hu, Dimitris Metaxas, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Snapgen-v: Generating a five-second video within five seconds on a mobile device, 2025c. URL https://arxiv.org/abs/2412.10494.

Haitam Ben Yahia, Denis Korzhenkov, Ioannis Lelekas, Amir Ghodrati, and Amirhossein Habibian. Mobile video diffusion, 2024. URL https://arxiv.org/abs/2412.07583.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer, 2025. URL https://arxiv.org/abs/2408.06072.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6613–6623, June 2024a.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis, 2024b. URL https://arxiv.org/abs/2405.14867.

Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. Make pixels dance: High-dynamic video generation, 2023. URL https://arxiv.org/abs/2311.10982.

Shuai Zhang, Bao Tang, Siyuan Yu, Yueting Zhu, Jingfeng Yao, Ya Zou, Shanglin Yuan, Li Yu, Wenyu Liu, and Xinggang Wang. Mobilei2v: Fast and high-resolution image-to-video on mobile devices, 2025a. URL https://arxiv.org/abs/2511.21475.

Xiangjun Zhang, Litong Gong, Yinglin Zheng, Yansong Liu, Wentao Jiang, Mingyi Xu, Biao Wang, Tiezheng Ge, and Ming Zeng. Rise-t2v: Rephrasing and injecting semantics with llm for expansive text-to-video generation, 2025b. URL https://arxiv.org/abs/2511.04317.

Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jiawei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models, 2023. URL https://arxiv.org/abs/2310.08465.

Dongqi Zheng. Diffusion models on the edge: Challenges, optimizations, and applications, 2025. URL https://arxiv.org/abs/2504.15298.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024. URL https://arxiv.org/abs/2412.20404.

Zangwei Zheng, Xiangyu Peng, Yuxuan Lou, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, Wenjun Li, Yuhui Wang, Anbang Ye, Gang Ren, Qianran Ma, Wanying Liang, Xiang Lian, Xiwen Wu, Yuting Zhong, Zhuangyan Li, Chaoyu Gong, Guojun Lei, Leijun Cheng, Limin Zhang, Minghao Li, Ruijie Zhang, Silan Hu, Shijie Huang, Xiaokang Wang, Yuanheng Zhao, Yuqi Wang, Ziang Wei, and Yang You. Open-sora 2.0: Training a commercial-level video generation model in $200k, 2026. URL https://arxiv.org/abs/2503.09642.

Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models, 2023. URL https://arxiv.org/abs/2211.11018.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation, 2024. URL https://arxiv.org/abs/2405.01434.

A Effect Definitions

In this work, we focus on three representative cinematic motion effects: bullet time, dolly zoom, and slow motion. We briefly define them below and clarify the visual properties that are emphasized in our experiments.

Bullet Time. Bullet time refers to a cinematic effect in which the viewpoint appears to move rapidly around a relatively frozen subject or scene, producing a strong sense of spatial immersion and three-dimensional geometry. In video generation, this effect requires coherent camera motion, stable subject identity, and consistent scene structure across viewpoints. A successful bullet-time video should preserve the foreground subject while producing a smooth orbital or sweeping camera trajectory around it.

Dolly Zoom. The Hitchcock effect, also known as dolly zoom, is a classic camera effect created by simultaneously changing camera position and focal scale so that the subject remains relatively stable in size while the background perspective changes dramatically. In generated videos, this effect is characterized by a distinct push-pull perception: the foreground subject remains visually anchored, while the background expands or contracts in a way that induces strong spatial tension. This motion pattern places high demands on foreground-background consistency and controllable camera transformation.

Slow Motion. Slow motion refers to the visual effect of temporally stretching motion so that dynamic events appear to unfold more slowly than in real time. In our setting, this effect emphasizes smooth temporal interpolation, stable object appearance, and reduced motion artifacts across consecutive frames. Compared with bullet time and dolly zoom, slow motion is less dependent on explicit camera trajectory control and more sensitive to temporal continuity, detail preservation, and motion realism.

B Evaluation Criteria

For automatic evaluation, we report the dimensions most relevant to cinematic motion effects: subject consistency, background consistency, motion smoothness, aesthetic quality, imaging quality, I2V subject, and I2V background. Together they cover the core requirements of our task, namely preserving subject identity, maintaining background coherence, and producing temporally smooth, visually plausible motion.

We omit several automatic metrics that are poorly aligned with our setting. The dynamic degree metric, for example, measures the overall magnitude of motion, whereas our target effects are not defined by motion amplitude. Bullet time and dolly zoom rely mainly on coordinated camera transformation and foreground–background geometric consistency, while slow motion emphasizes temporal smoothness and motion realism at reduced apparent speed. A higher dynamic degree therefore does not imply better effect quality, and can bias evaluation toward motion amplitude rather than controllable cinematic behavior.

We also exclude the VBench camera-control metric from the automatic comparison with base I2V models. There, we compare CineMobile mainly with Wan2.1-based models that are not optimized for bullet time, dolly zoom, or slow motion. In this case, the camera-control score is hard to read as a fair measure of effect quality, since it largely reflects whether a model has been explicitly adapted to these motion patterns. We therefore rely on the selected VBench dimensions for automatic comparison, and assess effect-specific camera behavior through qualitative results and human evaluation.

For human evaluation against commercial models, we rate five dimensions. *Artifact control* captures the absence of visible rendering artifacts, such as flicker, ghosting, unstable texture, or implausible structure. *Distortion* measures whether the video preserves plausible scene geometry and avoids unnatural deformation during motion. *Uniformity* measures appearance stability across frames, covering style, illumination, color tone, and overall presentation. *Consistency* measures broader temporal coherence: stable subject identity, continuous motion, and preserved foreground–background relations over time. *Camera control* measures whether the video follows the intended motion pattern, such as smooth orbital motion for bullet time, coordinated perspective change for dolly zoom, or temporally stretched but natural motion for slow motion. We report the average over these five dimensions as the total human-evaluation score.

Figure 8 Width pruning may lead to identity shifts and deformation issues.

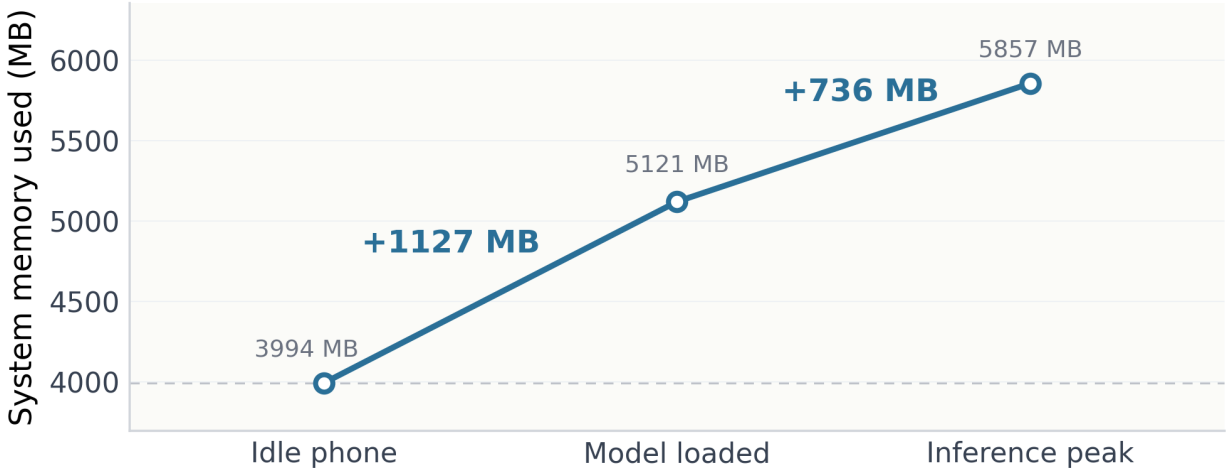

Figure 9 Memory footprint of CineMobile on an Infinix NOTE 60 Ultra. When the phone is idle under normal operation, the system memory usage is 3994 MB. After loading CineMobile, memory usage increases by 1127 MB to 5121 MB. During inference on 480p video, the peak memory usage reaches 5857 MB, requiring an additional 736 MB over the loaded-model state. Overall, CineMobile introduces only 1863 MB of additional runtime memory over the idle baseline, demonstrating that it can run efficiently on mobile devices with a modest memory footprint.

C Compare with Width Pruning

As illustrated in Figure 8, the width-pruned method exhibits noticeable identity deviation and detail degradation. Therefore, instead of reducing the hidden dimensionality, we keep the original backbone width unchanged and focus on structured depth pruning. This design confines compression to redundant layer transformations while allowing the pruned model to remain aligned with the original backbone's feature space.

D Additional Qualitative Results

We provide additional qualitative results for the three cinematic motion effects considered in this work, namely bullet time, dolly zoom, and slow motion. All videos in this section are generated on a MediaTek Dimensity 8400 Ultimate 5G platform. For each example, frames are uniformly sampled at different temporal ratios from the generated video. These visualizations complement the main-paper comparisons by showing that CineMobile maintains stable subject identity, coherent foreground-background relations, and smooth temporal progression across diverse portrait scenes.

E Data Curation and Preprocessing

Our training images are collected from high-quality portrait datasets, including PPR10K and Pixel. Since our task focuses on cinematic motion effects for portrait-centric image-to-video generation, we further apply a dedicated data curation pipeline to improve both visual quality and scene suitability. In particular, we use Qwen3-VL-30B-A3B-Instruct to automatically filter out images with overly simple or uninformative backgrounds, extreme overexposure or underexposure, insufficient spatial resolution, multiple prominent persons, overly tight face-dominant close-ups, missing human subjects, or very low aesthetic quality. We additionally remove samples with poor composition, ambiguous foreground-background structure, or limited motion potential, since such cases are less suitable for evaluating camera-driven effects such as bullet time and dolly zoom. This preprocessing procedure yields a cleaner and more task-aligned portrait dataset for both training and evaluation. All images used for both training and inference are sourced from publicly accessible datasets.

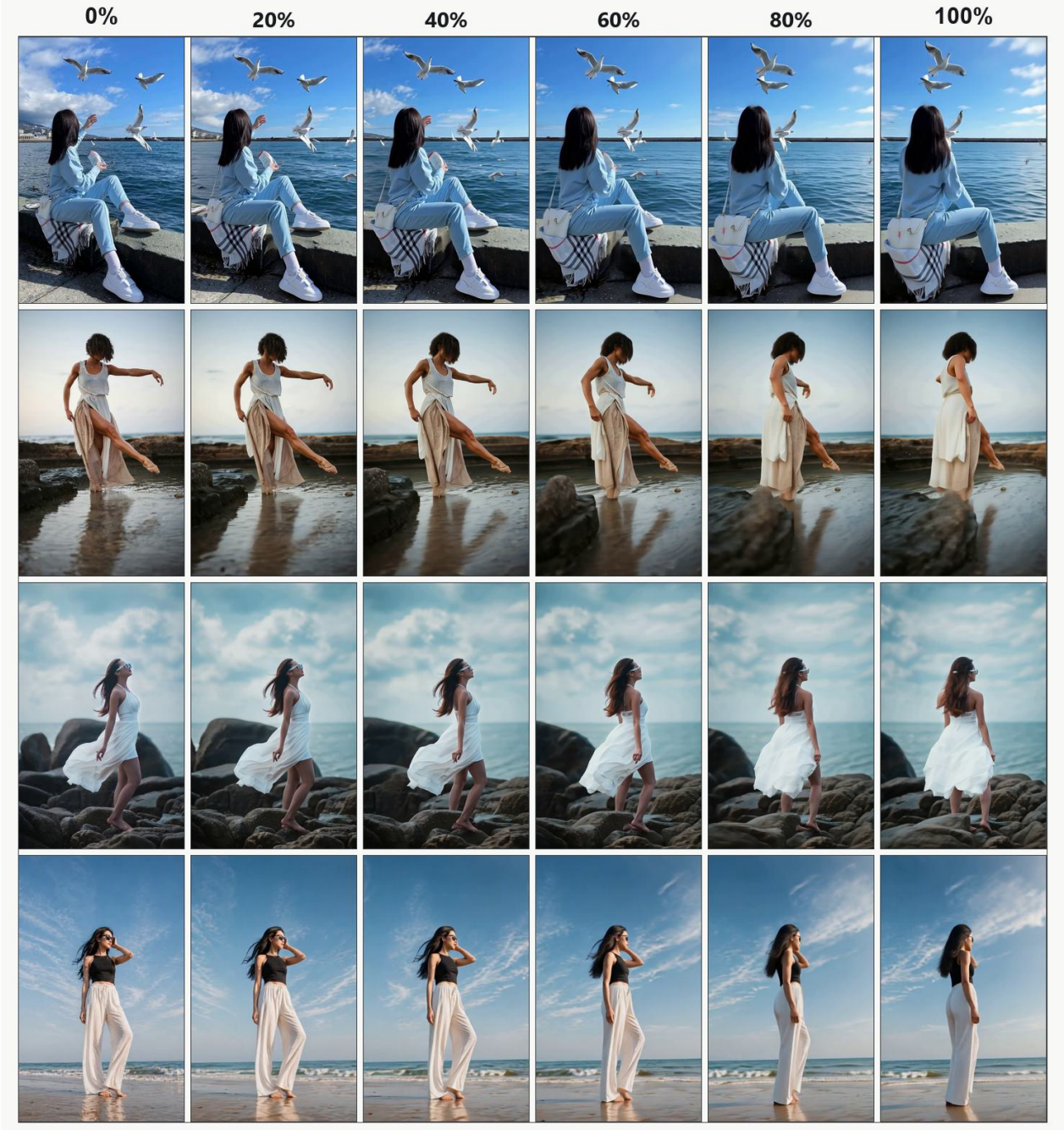

Figure 10 Additional qualitative results on **bullet time**. Each row shows one generated example, and columns correspond to uniformly sampled frames at different temporal ratios. CineMobile produces smooth viewpoint transition while preserving subject appearance and scene geometry throughout the sequence.

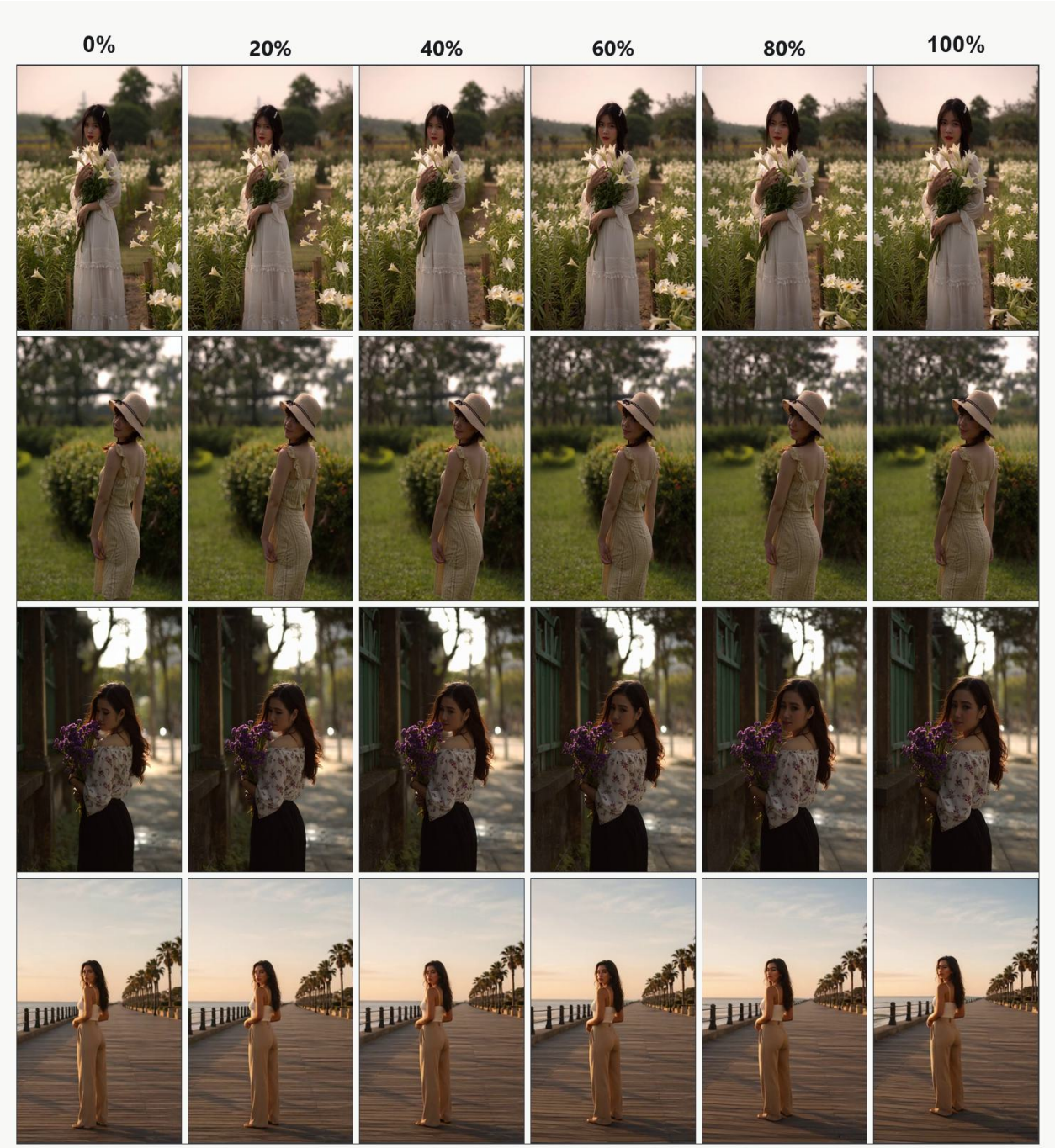

Figure 11 Additional qualitative results on **dolly zoom**. The sampled frames illustrate consistent foreground stabilization together with coordinated background perspective variation, which is the key visual characteristic of the Hitchcock dolly-zoom effect.

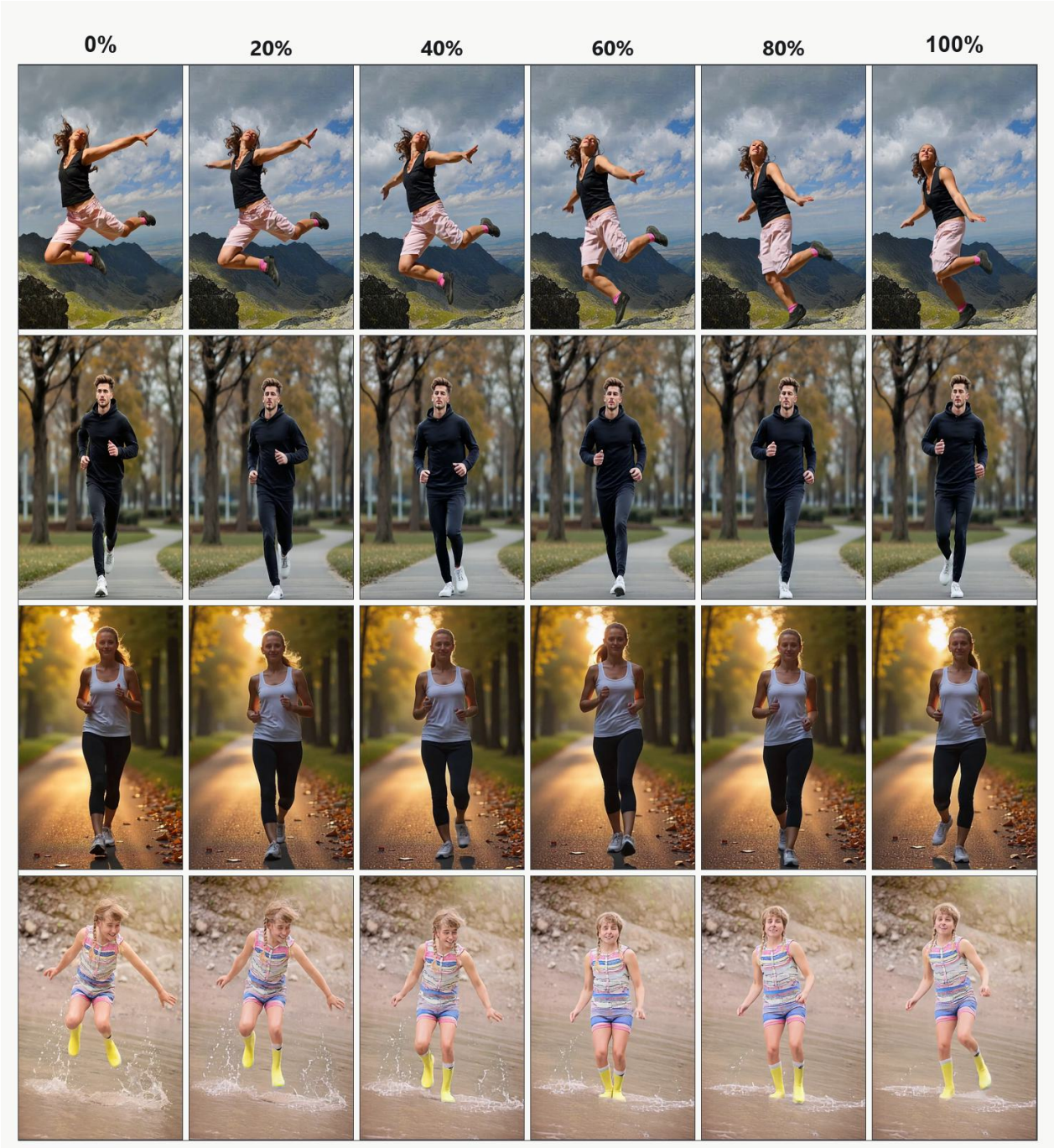

Figure 12 Additional qualitative results on **slow motion**. CineMobile preserves temporal smoothness and subject consistency while reducing abrupt motion changes, yielding visually plausible slow-motion sequences across different scenes.