REWARD-FORCING: AUTOREGRESSIVE VIDEO GENERATION WITH REWARD FEEDBACK

Jingran Zhang[1], **Ning Li**[1], **Yuanhao Ban**[2], **Andrew Bai**[2], **Justin Cui**[2]
[1]University of California, San Diego [2]University of California, Los Angeles
{jiz330, nil024}@ucsd.edu, banyh2000@gmail.com
andrewbai@cs.ucla.edu, justincui@ucla.edu

ABSTRACT

While most prior work in video generation relies on bidirectional architectures, recent efforts have sought to adapt these models into autoregressive variants to support near real-time generation. However, such adaptations often depend heavily on teacher models, which can limit performance, particularly in the absence of a strong autoregressive teacher, resulting in output quality that typically lags behind their bidirectional counterparts. In this paper, we explore an alternative approach that uses reward signals to guide the generation process, enabling more efficient and scalable autoregressive generation. By using reward signals to guide the model, our method simplifies training while preserving high visual fidelity and temporal consistency. Through extensive experiments on standard benchmarks, we find that our approach performs comparably to existing autoregressive models and, in some cases, surpasses similarly sized bidirectional models by avoiding constraints imposed by teacher architectures. For example, on VBench, our method achieves a total score of 84.92, closely matching state-of-the-art autoregressive methods that score 84.31 but require significant heterogeneous distillation.

1 INTRODUCTION

Diffusion models (Ho et al., 2020; Liu et al., 2023) have emerged as a powerful class of generative models, achieving state-of-the-art results in a wide range of domains, including image synthesis, audio generation, and molecular modeling. By iteratively denoising data from a simple noise distribution, these models are capable of producing high-fidelity samples that capture complex data distributions. Their theoretical foundation rooted in stochastic differential equations, combined with their empirical robustness, has made diffusion models a dominant approach in the generative modeling landscape.

Building on this success, video diffusion models (Ho et al., 2022b; Brooks et al., 2024) extend the capabilities of diffusion-based generation to the temporal domain. Unlike images, videos require the model to generate spatially coherent and temporally consistent sequences of frames, which significantly increases the modeling complexity. Recent advancements have adapted diffusion models to handle temporal correlations by incorporating spatiotemporal attention, motion priors, and multi-frame conditioning mechanisms. These models have demonstrated promising results in generating short to medium-length video clips with high perceptual quality.

However, most existing video diffusion models are built upon bidirectional architectures (Wang et al., 2025; Kong et al., 2024), where the generation of each frame depends on information from both past and future timesteps. While effective for high quality video generation, bidirectional models are inherently unsuitable for streaming or real-time video generation, taking minutes to generate a single video with 5 seconds. Recent efforts have explored transforming bidirectional video diffusion models into autoregressive formulations (Yin et al., 2025; Kim et al., 2024) to enable low-latency and scalable generation. These methods typically rely on distillation from powerful bidirectional teachers, but their performance is often limited by the quality of the teacher model and the challenges of maintaining temporal coherence in the absence of future context.

In parallel, reinforcement learning (RL) has gained traction as a post-training strategy for enhancing the quality and alignment of diffusion-based image generation. By optimizing generation policies directly with respect to reward signals—such as human preferences, aesthetic scores, or perceptual quality—RL enables fine-grained control over generative behavior beyond supervised objectives. Techniques such as Reinforcement Learning with Human Feedback (RLHF) (Ouyang et al., 2022; Liu et al., 2025a) have demonstrated success in aligning language and vision-language models, and recent works have begun applying RL to tune diffusion models for improved fidelity, diversity, or alignment with desired styles.

In this work, we build upon these directions and present a novel approach that leverages autoregressive video diffusion modeling combined with reinforcement learning to achieve efficient, scalable, and high-quality streaming video generation. In summary, our contributions are as follows:

- We show that the performance of existing methods for converting bidirectional video diffusion models into autoregressive models are bounded by the teacher's performance.
- We observe that consistent motions are learned before texture which aligns with previous studies on image generation and propose a framework that uses pure reward signals to guide autoregressive generation of high quality videos.
- Extensive experiments on VBench show that our proposed method is able to generate high quality videos, comparable to baseline autoregressive video diffusion models on quality and total score.

2 Related work

Diffusion Models for Image Generation Denoising Diffusion Probabilistic Models (DDPM) pioneered the modern image generation by casting generative modeling as a Markovian noising–denoising process whose reverse dynamics are learned with a simple, noise-conditional score network (Ho et al., 2020). While DDPMs achieve impressive sample quality, they rely on hundreds or thousands of sequential denoising steps. Denoising Diffusion Implicit Models (DDIM) alleviate this inefficiency by interpreting the learned stochastic process as a non-Markovian deterministic ordinary differential equation, enabling much faster inference without retraining and preserving visual fidelity (Song et al., 2021). Subsequent flow-matching methods such as the Flow Matching framework further unify diffusion and continuous-normalizing-flow viewpoints by directly training vector fields that map noise to data in a single pass, often reducing both training variance and sampling cost (Lipman et al., 2022; Liu et al., 2023). The most widely used architecture for such tasks was UNet (Ronneberger et al., 2015), which was later revolutionized by the introduction of DiT (Peebles & Xie, 2023). Due to the high computational cost of operating directly in pixel space, modern architectures typically first project the input into a latent space (Rombach et al., 2022; Blattmann et al., 2023a), where subsequent computations are performed. The final results are then decoded back into pixel space using a decoder (Kingma & Welling, 2014).

Video Diffusion Models Video diffusion models represent a significant advancement in generative AI, extending the principles of image diffusion models to handle temporal dynamics for tasks such as text-to-video generation, image-to-video synthesis, and video editing. Early extensions to video, such as Video Diffusion Models (Ho et al., 2022b), introduced 3D U-Net architectures with factorized spatio-temporal attention to ensure temporal coherence in video output with relative position embeddings (Shaw et al., 2018). Subsequent models like Imagen Video (Ho et al., 2022a) employed cascaded diffusion pipelines for high-definition videos, while Make-A-Video (Singer et al., 2023) leveraged pretrained text-to-image models and unsupervised video data to learn visual appearance and motion without requiring paired text-video data. By introducing novel spatial-temporal modules and a multi-stage video generation pipeline, it achieves state-of-the-art results in video quality, temporal consistency, and alignment with textual input. Latent space approaches, exemplified by models such as Stable Video Diffusion (Blattmann et al., 2023a;b), enhanced efficiency by operating on compressed representations and incorporating temporal attention mechanisms. More recent innovations include zero-shot methods like Text2Video-Zero (Khachatryan et al., 2023), which enable video generation without video-specific training, and Sora (OpenAI, 2023), utilizing Diffusion Transformers on spacetime patches for scalable, high-fidelity outputs. These developments address challenges in temporal consistency and computational demands, paving the way for applications in

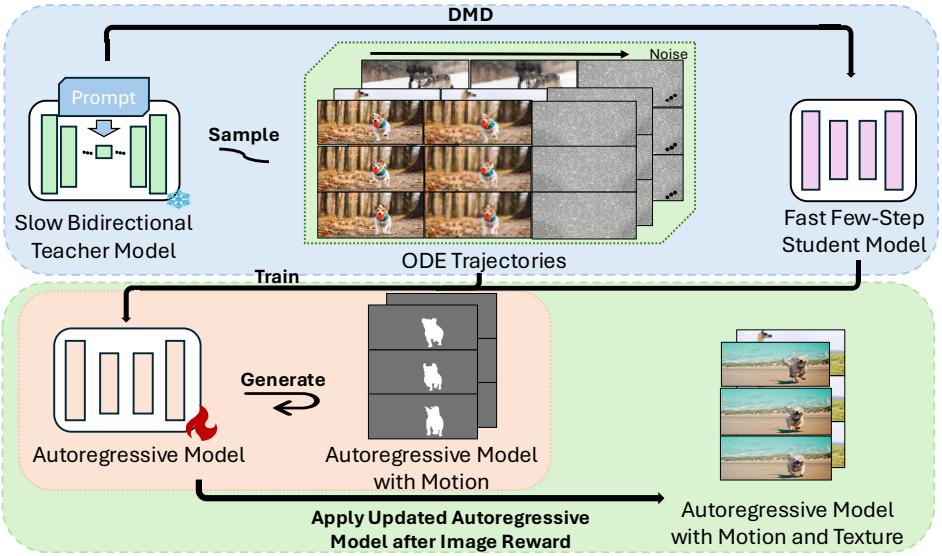

Figure 1: Overview of our proposed pipeline. The method first leverages a small set of ODE-based trajectories generated by a teacher model to guide the learning of motion dynamics. Subsequently, a reward model is employed to enhance the generation with fine-grained texture details.

entertainment, simulation, and content creation. The quality of generation is further enhanced by following models such as Wan (Wan et al., 2025), HunyuanVideo (Kong et al., 2024) and Veo (Google, 2025).

Diffusion Acceleration Although diffusion models are able to perform high quality image and video synthesis, they suffer from extremely high denoising steps with high cost. In order to solve the problem, one direction is to design better ODE solvers which enable few-step sampling such as DPM-Solver (Lu et al., 2022), GENIE (Dockhorn et al., 2022) and S4S (Frankel et al., 2025). An alternative approach involves distillation techniques to compress multi-step diffusion processes into fewer or single-step generations, including Score Distillation Sampling (SDS) for leveraging pre-trained diffusion priors in optimization tasks (Luo et al., 2024), Distribution Matching Distillation (Yin et al., 2024b) which transforms diffusion models into efficient one-step generators with minimal quality loss, its enhanced successor DMD2 (Yin et al., 2024a) that further improves training efficiency and performance, and Consistency Models that enable high-quality one- or few-step sampling by learning consistent mappings from noise to data (Song et al., 2023).

Autoregressive and Streaming Generation While most diffusion models are trained with full-sequence denoising, they usually require substantially long time to generate videos up to several seconds. Thus autoregressive and streaming variants have been proposed to enable fast frame-by-frame or chunk by chunk generation (Yin et al., 2025; Huang et al., 2025). Among the work, FIFO-Diffusion (Kim et al., 2024) proposes a training-free method that uses a queue to hold frames of different noise levels with the frames closer to the beginning of the queue having lower noise level which requires less denoising and frames closer to the end of the queue have high noise levels which requires more denoising steps. The method is able to generate substantially longer videos than the original model. However, due to the heterogeneous denoising steps in the input, the generated videos often show degraded quality and inconsistency. Ouroboros-Diffusion (Chen et al., 2025b) tries to solve the inconsistency problem by introducing a novel latent sampling technique at the end of the queue with subject-aware cross-frame attention mechanism. The heterogeneous noise level used by Kim et al. (2024); Chen et al. (2025b) shares the same idea as Diffusion-Forcing (Chen et al., 2024; Song et al., 2025) where the models are trained to perform denoising on noisy frames with different noise levels. However, due to the large combination of noise levels, recent models (Chen et al., 2025a) usually first train the model with same noise level, then finetune it to handle different noise levels. Later work (Kodaira et al., 2025) tries to distill the model into few-step generators which alleviates the problem. Another line of works such as CausVid (Yin et al., 2025) and Self

Forcing (Huang et al., 2025) adopts the same autoregressive manner but with same noise levels which show improved video quality.

Reinforcement Learning for Generative Models Reinforcement learning has been widely applied in Large Language Models OpenAI (2023); Dubey et al. (2024); Comanici et al. (2025) to finetune models to align better with user preferences Christiano et al. (2017); Rafailov et al. (2023); Liu et al. (2024). Such techniques have also been generalized to diffusion models. ImageReward Xu et al. (2023) builds the first general-purpose text-to-image human preference reward model which can be used in the Reward Feedback Learning (ReFL) framework to effectively align image generation with human preferences. VisionReward Xu et al. (2024) extends it to both images and videos by designing a fine-grained, multi-dimensional reward model, achieving new state-of-the-art performances on various benchmark datasets. VideoAlign Liu et al. (2025a) introduces a VLM-based reward model to address three critical dimensions including Visual Quality, Motion Quality and Text Alignment. ROCM Shekhar & Zhang (2025) proposes a direct reward optimization framework for applying reinforcement learning from human feedback (RLHF) to consistency models, enabling efficient training without the need for policy gradients. Reward-Instruct Luo et al. (2025) achieves fast image synthesis through reward-centric approaches.

3 METHODOLOGY

In this section, we will first introduce the background of diffusion models that distillation technique used for converting the model into a few step generator. Then we introduce how our method works. The overall pipeline is shown in Fig. 1.

3.1 PRELIMINARIES

Diffusion Model Diffusion-based generative models treat data generation as running time backwards from a simple prior toward the data distribution. In the widely-used denoising diffusion probabilistic model (DDPM), one defines a forward Gaussian noise process

$$q(x_t \mid x_0) = \mathcal{N}\big(\alpha_t x_0,\ \sigma_t^2 \mathbf{I}\big),$$
$$\alpha_t = \prod_{s=1}^{t} (1-\beta_s)^{\frac{1}{2}}, \quad \sigma_t^2 = 1 - \alpha_t^2.$$

where $x_0 \sim p_{\text{data}}$ and $\{\beta_t\}_{t=1}^{T}$ is a small variance schedule. Generating a sample amounts to integrating the reverse-time score SDE

$$\mathrm{d}x_t = \big[f(x_t,t) - g(t)^2\, \nabla_{x_t} \log q_t(x_t)\big]\mathrm{d}t + g(t)\, \mathrm{d}\bar{w}_t.$$

or, in practice, its deterministic probability-flow ODE, using a neural score estimator $s_\theta(x,t) \approx \nabla_x \log q_t(x)$.

While DDPMs rely on stochastic trajectories, flow matching (Lipman et al., 2022) learns a deterministic velocity field that continuously transports a tractable prior p_0 to the data distribution p_1. Given a coupling $(x_0, x_1) \sim p_0 \times p_1$, define linear interpolates $x_t = (1-t)\, x_0 + t\, x_1$. The ground-truth velocity $v^\star(x_t,t)$ is

$$v^\star(x_t,t) = \frac{dx_t}{dt} = x_1 - x_0.$$

and one trains a neural field $v_\theta(x,t)$ by the flow-matching loss

$$\mathcal{L}(\theta) = \mathbb{E}_{t\sim\mathcal{U}(0,1)}\, \mathbb{E}_{(x_0,x_1)} \Big[\big\| v_\theta(x_t,t) - v^\star(x_t,t) \big\|^2 \Big].$$

At inference, samples evolve deterministically via the ODE $\dot{x}_t = v_\theta(x_t,t)$, often requiring far fewer steps than stochastic diffusion samplers while maintaining high generative fidelity. In this work, we use Wan2.1 (Wang et al., 2025) which is trained using the flow matching objective.

Video Diffusion Distillation Same as CausVid (Yin et al., 2025) and SelfForcing (Huang et al., 2025), we also utilize the distilled few-step model as the starting model for faster generation. It is achieved by minimizing the reverse KL divergence between data distribution and stduent generator's output distribution which can be formulated as

$$\begin{aligned}\nabla_\theta \mathcal{L}_{\mathrm{DMD}} &= \mathbb{E}_t\big[\nabla_\theta \,\mathrm{KL}\big(p_{\mathrm{fake},t} \| p_{\mathrm{real},t}\big)\big] \\ &= -\,\mathbb{E}_t\Big[\int \big(s_{\mathrm{real}}\big(\Phi(G_\theta(z),t),t\big) \\ &\qquad -\; s_{\mathrm{fake}}\big(\Phi(G_\theta(z),t),t\big)\big)\,\frac{dG_\theta(z)}{d\theta}\,dz\Big].\end{aligned}$$

where Φ is the forward diffusion process and $z \sim \mathcal{N}(0, \mathbf{I})$ is a random Gaussian noise input.

3.2 Training with ODE and Reward Guidance

Recent studies have shown that diffusion models typically generate a global coarse structure before progressively refining texture details (Chen et al., 2023; Jiang et al., 2024; Sun et al., 2025; Liu et al., 2025b). We empirically observe a similar phenomenon when converting bidirectional video diffusion teachers into autoregressive models which can be seen in Fig. 3 here. After distilling the model into few-step models, similar to CausVid (Yin et al., 2025), we use sample ODE trajectories for early training. The process is done by first sample noise inputs $\{x_T^i\}_{i=1}^L$ from $\mathcal{N}(0, I)$ and then use an ODE solver with the pretrained teacher to generate reverse trajectories $\{x_t^i\}_{i=1}^L$ across all timesteps. The model is then trained with the following objective:

$$\mathcal{L}_{\text{ode}} = \mathbb{E}_{x,t^i} \left\| G_\phi(\{x_{t^i}^i\}, \{t^i\}) - \{x_0^i\} \right\|^2.$$

where G_ϕ is the student trained from the teacher. After this step, we empirically observe that the model has learned to generate consistent motions without much texture info.

Optimization with Reward Feedback After initializing the model with an ODE-based process that teaches motion synthesis, we introduce a reward-guided optimization stage to enhance video quality. Specifically, we adopt ImageReward (Xu et al., 2023) as our reward model and incorporate it into training in a differentiable manner to directly guide the video diffusion model.

Let the generated video be denoted as

$$\hat{x}_{1:T} = G_\theta(z)$$

, where G_θ is the autoregressive video generator parameterized by θ, and z is the latent input. The reward model $\mathcal{R}(\cdot)$ assigns a scalar reward indicating perceptual quality. We define the reward-guided objective as:

$$\mathcal{L}_{\text{reward}}(\theta) = -\mathbb{E}_{z\sim\mathcal{Z}}\left[\mathcal{R}(\hat{x}_T)\right].$$

where $\hat{x}_T$ is the last frame of the generated video. We choose to apply supervision to the last frame because we find that supervising more frames encourages static content with reduced motion, while supervising the last frame preserves motion better likely due to its proximity to the end of the generation trajectory in the autoregressive process. Results of supervising random frames are shown in ablation study.

3.3 Baseline Methods

We benchmark three classes of approaches: (i) standard bidirectional diffusion models, (ii) autoregressive methods that render videos one (latent) frame at a time, and (iii) autoregressive methods that synthesize videos in temporal chunks. For standard diffusion models, we include LTX-Video (HaCohen et al., 2024) and Wan2.1 (Wan et al., 2025) which is the base model used by Yin et al. (2025); Huang et al. (2025) and our method. For frame-wise autoregressive models, we include NOVA (Deng et al., 2025) which formulates the video generation problem as non-quantized autoregressive modeling of temporal frame-by-frame prediction and spatial set-by-set prediction, Pyramid Flow (Jin et al., 2025) that interprets the original denoising trajectory as a series of pyramid stages and Self Forcing (Huang et al., 2025) that trains the autoregressive models by relying on

Model	#Params	Resolution	Throughput (FPS) ↑	Latency (s) ↓	Evaluation scores ↑		
					Total Score	Quality Score	Semantic Score
Diffusion models							
LTX-Video	1.9B	768×512	8.98	13.5	80.00	82.30	70.79
Wan2.1	1.3B	832×480	0.78	103	84.26	85.30	80.09
Frame-wise Autoregressive models							
NOVA	0.6B	768×480	0.88	4.1	80.12	80.39	79.05
Pyramid Flow	2B	640×384	6.7	2.5	81.72	84.74	69.62
Self Forcing (frame-wise)	1.3B	832×480	8.9	0.45	84.26	85.25	80.30
Chunk-wise autoregressive models							
SkyReels-V2	1.3B	960×540	0.49	112	82.67	84.70	74.53
MAGI-1	4.5B	832×480	0.19	282	79.18	82.04	67.74
CausVid	1.3B	832×480	17.0	0.69	81.20	84.05	69.80
Self Forcing	1.3B	832×480	17.0	0.69	84.31	85.07	**81.28**
ODE Only (chunk-wise)	1.3B	832×480	17.0	0.69	68.77	73.27	50.81
Ours (chunk-wise)	1.3B	832×480	17.0	0.69	**84.92**	**85.91**	80.97

Table 1: Overall performance comparison with previous baseline methods using VBench.

self-generated frames instead of ground-truth ones. For Chunk-wise models, we include SkyReels-V2 (Chen et al., 2025a), MAGI-1(Sand-AI, 2025) that utilizes transformer based VAE and increased noise levels among chunks that are being generated, CausVid (Yin et al., 2025) that uses causal attention at training time and self-rollout at inference time and Self Forcing (Huang et al., 2025) which utilizes similar architectures as its frame-wise counterpart.

3.4 Implementation Details

Similar to Yin et al. (2025); Huang et al. (2025), our model is based on Wan2.1-T2V-1.3B(Wan et al., 2025). The model is first distilled into a 4-step model using DMD. For ODE initialization, we also directly sample 1.4K trajectories from the bidirectional model. Our work uses Self-Rollout during training which is the same as SelfForcing (Huang et al., 2025). For our main method, we only employ loss from the reward models. For our reward plus distillation setting, we train the model using DMD together with reward loss as shown in the previous section. Our method also works for both chunk-wise and frame-wise autoregressive generation. In our implementation, since we only need to sample the ODE trajectory from the teacher model, we don't need any real training dataset. Thus our overall approach remains data free which is similar to that of Self Forcing (Huang et al., 2025). Similar to previous works (Huang et al., 2025), we enable EMA (Hunter, 1986) during the training process.

Note that our training process eliminates the second distillation stage used in CausVid (Yin et al., 2025) and Self Forcing (Huang et al., 2025). As a result, there is no need to load the teacher and critic models or train the critic model to approximate the generator distribution, making the training process after ODE initialization significantly more lightweight and efficient.

4 Experimental Results

4.1 Evaluation Metrics

We evaluate our method on VBench (Huang et al., 2024), following the protocol in Yin et al. (2025); Huang et al. (2025). VBench comprises 16 evaluation dimensions, which are aggregated into three scores: quality, semantic, and total. The quality score is a weighted average of metrics such as subject consistency, background consistency, temporal flickering, motion smoothness, aesthetic quality, and dynamic degree, reflecting the technical fidelity of generated videos. The semantic score averages dimensions like object class, multiple objects, human action, color, spatial relationships, and overall consistency, capturing content relevance and semantic coherence. Since we adopt a similar KV cache strategy, our inference runtime matches that of Self Forcing. In addition to the aggregated

scores, we also report per-dimension results to provide a more fine-grained comparison of visual quality and semantic consistency across different methods.

4.2 Main Results

Following prior works (Yin et al., 2025; Huang et al., 2025), we report aggregated scores in Tab. 1, with selected metrics plot in Fig. 2.

Overall, standard bidirectional teacher models tend to perform better than autoregressive models in many aspects. For instance, the bidirectional model Wan2.1 (Wan et al., 2025) achieves a total score of 84.26, whereas CausVid (Yin et al., 2025) attains a lower score of 81.20, likely due to architectural differences and a mismatch between training and inference procedures. Self Forcing (Huang et al., 2025) attempts to address this issue by incorporating self-rollout during training, which helps reduce the discrepancy and leads to improved results. However, since it is distilled from the teacher using DMD (Yin et al., 2024a), its performance remains largely comparable to that of the bidirectional model with same size.

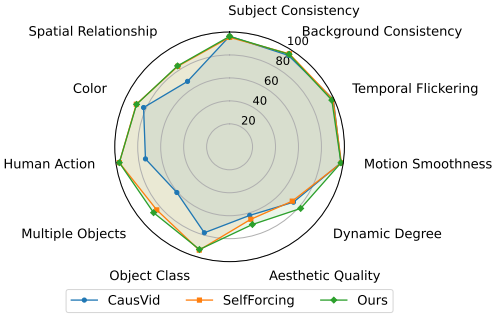

Figure 2: Comparison between our methods and baseline methods on selected VBench metrics. Our method shows competitive performances without extensive heterogeneous distillation.

Our method, despite being relatively simple and not relying on extensive heterogeneous distillation between bidirectional teacher and autoregressive student model, achieves competitive results and shows encouraging signs in comparison to both the bidirectional model and recent autoregressive baselines. For example, it achieves a quality score of 85.81, slightly higher than the 85.25 obtained by state-of-the-art frame-wise Self Forcing. It also reaches a total score of 84.92, outperforming the second-best score achieved by the chunk-wise Self Forcing model. As shown in Fig. 2, our method performs favorably on several individual dimensions, such as aesthetic quality and dynamic degree.

One interesting observation is that as the distillation process progresses, the dynamic degree of motion tends to diminish. While more analysis is needed, this may suggest a trade-off between distillation and motion richness, which we hope to explore further in future work. Additionally, although our model is only initialized with a small number of ODE trajectories, Tab. 1 shows meaningful improvements over the initial model in both quality and semantic scores. Taken together, these findings suggest that with careful design, autoregressive models may be able to close the performance gap with teacher models, possibly even without relying heavily on distillation. We believe this points toward a promising direction worth further investigation.

4.3 Qualitative Comparison

Here we show the generated videos in Fig. 3 and compare with two other state-of-the-art models including CausVid (Yin et al., 2025) and Self Forcing (Huang et al., 2025) and the original Wan2.1 1.3B model. It can be seen that our model is able to generate high quality videos compared to the baseline model. Note that since the texture of our videos are generated with the guidance of an external reward model, the overall style will look different with that of the teacher and baseline models.

5 Ablation Study

5.1 Combining Reward with Distillation

One alternative we explored is to directly combine the reward loss with the distribution matching loss and jointly optimize both objectives during training. While this approach seems intuitive, balancing the teacher's guidance with task-specific reward supervision, we observed that it actually results in degraded performance. Specifically, when training under a comparable computational budget, the final model only achieves a total score of 82.55, which is notably lower than both the original

Figure 3: Comparison of videos generated by our method and other baseline methods, using the prompt: "A joyful, playful Corgi running and frolicking in a vibrant park during sunset." The second row displays the optical flow of the video generated by the model trained solely with ODE trajectories, highlighting the motion patterns the model has learned.

SelfForcing baseline and our proposed method. We believe this performance drop is rooted in a fundamental conflict between the objectives of distribution matching and reward-based optimization. The distillation framework is explicitly designed to align the output distribution of the student model with that of the teacher, ensuring that the student faithfully mimics the teacher's behavior. However, by introducing a reward loss, especially one derived from models such as ImageReward, the student is incentivized to generate outputs that may diverge from those of the teacher in pursuit of higher reward scores. This divergence undermines the distribution alignment objective, weakening the distillation process and leading to overall lower performance. These results highlight the importance of carefully decoupling teacher supervision from reward optimization. The results can be found in Tab. 2

Model	Total Score	Quality Score	Semantic Score
CausVid	81.20	84.05	69.80
Self Forcing	84.31	85.07	**81.28**
Reward + Distill	82.55	83.20	79.96
Ours	**84.92**	**85.91**	80.97

Table 2: Performance comparison between our method and one combining both reward and distillation losses. Results are evaluated under similar compute. No further gains observed with more computation

5.2 Random Steps vs Last Step

Since a video is inherently a temporal sequence of image frames, there are multiple strategies for applying reward models during training. Common approaches include: (1) applying the reward only

to the last frame of the video versus (2) applying the reward to randomly selected frames throughout the sequence. In our experiments, we empirically find that supervising randomly selected frames results in a degradation in motion quality, with the motion degree dropping by more than 10 percentage points. We believe this decline is primarily due to the nature of the ImageReward model, which is designed to evaluate individual frames based on visual-textual alignment and focuses primarily on texture-level details. Because it lacks any notion of temporal continuity or cross-frame motion dynamics, it fails to capture the quality of motion across frames. As a result, applying it to random frames can encourage the model to prioritize static visual features over coherent motion, ultimately harming the motion of the generated videos.

5.3 Different Teacher Models

Previously, we observe that incorporating both distillation and reward loss into the training objective does not necessarily improve performance. This raises the question of whether using a different teacher model could yield different results. We experiment with two teacher configurations: the larger Wan2.1-14B model and the smaller Wan2.1-1.3B model, both combined with the same image-based reward loss to guide the student. Across multiple runs and evaluation metrics, we do not observe substantial differences in performance. We hypothesize that this is due to the limited capacity of the 1.3B student model, which may reach its performance ceiling regardless of their size or expressiveness. Additionally, the fundamental difference between the bidirectional teacher and the autoregressive student model may also contribute to the performance gap.

6 Conclusion & Future Work

In this work, we introduce an alternative approach for converting bidirectional video diffusion models into autoregressive counterparts by leveraging the guidance of reward models. Unlike prior methods that rely heavily on heterogeneous distillation from pre-trained bidirectional teachers, our framework eliminates the need for this additional training stage. We show that it is feasible to initialize the autoregressive model directly using the ODE-based training of the original diffusion model, and then progressively refine its generation capabilities through reward-guided optimization. By avoiding the constraints of strict teacher-student alignment, our method offers the potential to unlock more expressive and higher-quality video generation.

One promising direction for future research is to explore alternative reward models, including both image- and video-based rewards, to further improve the quality of generated videos. Our current approach builds on the findings (Jiang et al., 2024; Sun et al., 2025; Liu et al., 2025b; Materzyńska et al., 2024), which suggest that global structures in images and motions in videos tend to emerge early in the generative process. Directly modeling motion dynamics could potentially further enhance generation quality (Shi et al., 2024; Chefer et al., 2025). In addition, Reward-Instruct (Luo et al., 2025) shows that high-quality image synthesis can be achieved solely through reward-guided training while bypassing the heavy distillation process. Inspired by this, future work could investigate the possibility of eliminating the initial ODE training phase entirely in favor of a reward-driven learning paradigm for video generation.

7 Limitations

Although our method demonstrates competitive performance relative to baseline approaches, video generation inherently involves more complex and multi-dimensional judgments than image generation. Our framework relies entirely on open-source reward models, and users may select reward functions that best suit their downstream applications. We propose an alternative pathway for converting bidirectional teacher models into autoregressive student models; however, since bidirectional video diffusion models already provide strong guidance, we do not apply our method to standard non-autoregressive settings, which we leave for future work. Due to limited teacher supervision, we observe occasional inconsistencies in generated videos, which could potentially be mitigated by stronger or more expressive reward models. From a practical standpoint, improving the stability and efficiency of reward-based training for large-scale video models—through better optimization strategies, variance reduction, and scalable reward modeling—remains an important direction for future research.

REFERENCES

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023a.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023b.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL `https://openai.com/research/video-generation-models-as-world-simulators`.

Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. Videojam: Joint appearance-motion representations for enhanced motion generation in video models. *arXiv preprint arXiv:2502.02492*, 2025.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025a.

Jingyuan Chen, Fuchen Long, Jie An, Zhaofan Qiu, Ting Yao, Jiebo Luo, and Tao Mei. Ouroboros-diffusion: Exploring consistent content generation in tuning-free long video diffusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2079–2087, 2025b.

Tsai-Shien Chen, Chieh Hubert Lin, Hung-Yu Tseng, Tsung-Yi Lin, and Ming-Hsuan Yang. Motion-conditioned diffusion model for controllable video synthesis. *arXiv preprint arXiv:2304.14404*, 2023.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. In *ICLR*, 2025.

Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Genie: Higher-order denoising diffusion solvers. *Advances in Neural Information Processing Systems*, 35:30150–30166, 2022.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

Eric Frankel, Sitan Chen, Jerry Li, Pang Wei Koh, Lillian J Ratliff, and Sewoong Oh. S4s: Solving for a diffusion model solver. *arXiv preprint arXiv:2502.17423*, 2025.

Google. Veo, 2025. URL `https://deepmind.google/models/veo/`.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *NeurIPS*, 2022b.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

J Stuart Hunter. The exponentially weighted moving average. *Journal of quality technology*, 18(4): 203–210, 1986.

Jing Jiang, Yiran Ling, Binzhu Li, Pengxiang Li, Junming Piao, and Yu Zhang. Poetry2image: An iterative correction framework for images generated from chinese classical poetry. *arXiv preprint arXiv:2407.06196*, 2024.

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 *ICLR*, 2025.

Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15954–15964, 2023.

Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han. Fifo-diffusion: Generating infinite videos from text without training. In *NIPS*, 2024.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

Akio Kodaira, Tingbo Hou, Ji Hou, Masayoshi Tomizuka, and Yue Zhao. Streamdit: Real-time streaming text-to-video generation. *arXiv preprint arXiv:2507.03745*, 2025.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025a.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.

Yiheng Liu, Liao Qu, Huichao Zhang, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Xian Li, Shuai Wang, Daniel K Du, et al. Detailflow: 1d coarse-to-fine autoregressive image generation via next-detail prediction. *arXiv preprint arXiv:2505.21473*, 2025b.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems*, 35:5775–5787, 2022.

Weijian Luo, Zemin Huang, Zhengyang Geng, J Zico Kolter, and Guo-jun Qi. One-step diffusion distillation through score implicit matching. *Advances in Neural Information Processing Systems*, 37:115377–115408, 2024.

Yihong Luo, Tianyang Hu, Weijian Luo, Kenji Kawaguchi, and Jing Tang. Reward-instruct: A reward-centric approach to fast photo-realistic image generation. *arXiv preprint arXiv:2503.13070*, 2025.

Joanna Materzyńska, Josef Sivic, Eli Shechtman, Antonio Torralba, Richard Zhang, and Bryan Russell. Newmove: Customizing text-to-video models with novel motions. In *Proceedings of the Asian Conference on Computer Vision*, pp. 1634–1651, 2024.

OpenAI. Openai. `https://www.openai.com`, 2023. Mar 14 version.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Sand-AI. Magi-1: Autoregressive video generation at scale, 2025. URL `https://static.magi.world/static/files/MAGI_1.pdf`.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.

Shivanshu Shekhar and Tong Zhang. Rocm: Rlhf on consistency models. *arXiv preprint arXiv:2503.06171*, 2025.

Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, et al. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. In *ICLR*, 2023.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021.

Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. *arXiv preprint arXiv:2502.06764*, 2025.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *ICML*, 2023.

Xianbing Sun, Yan Hong, Jiahui Zhan, Jun Lan, Huijia Zhu, Weiqiang Wang, Liqing Zhang, and Jianfu Zhang. Ds-vton: High-quality virtual try-on via disentangled dual-scale generation. *arXiv preprint arXiv:2506.00908*, 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

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.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo 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*, pp. 6613–6623, 2024b.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *CVPR*, 2025.

A Appendix

A.1 More related works

Motion in Video Diffusion Models Motion modeling remains a central challenge in video diffusion. Many early video diffusion models, such as Video Diffusion Models (Ho et al., 2022b) primarily focus on spatiotemporal attention mechanisms but treat motion implicitly. More recent efforts like MCDiff (Chen et al., 2023) and VideoJAM (Chefer et al., 2025) attempt to model motion more explicitly, using a flow completion model or joint appearance-motion representations. And Motion-I2V (Shi et al., 2024) further explores motion disentanglement in the image-to-video task, suggesting a broader trend toward treating motion and appearance as separable generative components.

B Implementation Details

B.1 Computing Infrastructure

All experiments were conducted on 8 H100 GPUs with 80 GB of memory, running on a Linux operating system. The versions of all relevant libraries and frameworks will be detailed in the later released GitHub repository.

B.2 Hyperparameters

Here we list the hyperparameters we used for our experiment. We use a batch size of 8 after training with the ODE trajectory. For the optimizer, we use AdamW with $\beta_1 = 0, \beta_2 = 0.999, \epsilon = 1e-8, weight_decay = 0.01$ with a learning rate of $2e-6$. For EMA, we use a decay of 0.99. When combing distillation loss with image reward loss, we normalize the losses so that they are on the same scale. When comparing with baseline methdos such as DMD or Causvid, we use their author-provided checkpoints.

C Quality of Generated Videos

It is important to recognize that our autoregressive model has not been fully distilled from the bidirectional teacher model, which processes information from both directions for a holistic view. Consequently, the videos generated by our model may differ substantially from those of the bidirectional teacher. These trajectories guide the video's temporal dynamics, and deviations could lead to unique visual patterns or flows.

Additionally, our model's texture generation is directed by a reward model, which provides evaluative feedback to align outputs with desired criteria. This allows downstream users to select or customize their own reward models, enabling the creation of videos tailored to specific styles, such as artistic animations or realistic footage, enhancing personalization.

This design also means our final model's performance is not strictly limited by the teacher. Instead, the autoregressive model can potentially outperform the teacher, guided by reward models, in areas like motion quality (e.g., smoother transitions) or aesthetic quality (e.g., richer details and harmony) if strong reward models are used.

D Generalization to Long Videos

Since our model is directly built upon CausVid (Yin et al., 2025) and Self Forcing (Huang et al., 2025), thus the model can be used to generate long videos as well. E.g. Causvid generates long videos by conditioning on previous generated frames which could be utilized to generate long videos by our method too.

Metric	Score ↑		
	CausVid	Self Forcing	Ours
Subject Consistency	96.32	95.32	96.15
Background Consistency	94.65	96.30	95.83
Temporal Flickering	99.38	99.05	97.88
Motion Smoothness	98.00	98.35	97.99
Dynamic Degree	73.61	72.22	81.94
Aesthetic Quality	62.24	65.75	70.51
Imaging Quality	67.82	69.02	68.84
Object Class	78.01	93.75	93.20
Multiple Objects	60.75	83.92	87.58
Human Action	74.00	97.00	97.00
Color	82.34	89.28	89.04
Spatial Relationship	67.74	83.38	83.78
Scene	14.03	57.34	54.36
Appearance Style	21.23	24.40	24.06
Temporal Style	18.85	20.48	21.19
Overall Consistency	22.50	26.72	26.14

Table 3: Complete VBench evaluation metrics in all 16 dimensions. Our method shows improvements in motion-related metrics (Dynamic Degree) and perceptual quality (Aesthetic Quality), while maintaining competitive performance on other dimensions.

E Full Evaluation Metrics

Tab. 3 presents the comprehensive evaluation across all 16 dimensions of VBench. The results show our method maintains competitive performance across most metrics while exhibiting particular strengths in certain areas. Our approach achieves a dynamic degree of 81.94, suggesting the ODE-based motion initialization contributes to motion representation. We also observe an aesthetic quality score of 70.51, aligning with the reward-guided refinement strategy. Additionally, the method yields scores of 87.58 in multiple objects and 83.78 in spatial relationship metrics, indicating its capability in preserving compositional relationships.

E.1 More Generated Samples

(a) In a still frame, a stop sign

(b) In a still frame, within the desolate desert, an oasis unfolded, characterized by the stoic presence of palm trees and a motionless, glassy pool of water

(c) A panda drinking coffee in a cafe in Paris, featuring a steady and smooth perspective

(d) A person is digging

(e) A person is crawling baby

(f) A person is ice skating

Figure 4: More generated samples. Prompts are randomly sampled from VBench of various scenes to benchmark the overall capability of the model.