BAgger: Backwards Aggregation for Mitigating Drift in Autoregressive Video Diffusion Models

Ryan Po   Eric Ryan Chan   Changan Chen   Gordon Wetzstein

Stanford University

https://ryanpo.com/bagger

Figure 1. In standard autoregressive video diffusion, exposure bias causes compounding errors ("drift") as generated frames become the context for future steps. This is typically observed as degrading image quality, contrast, and disintegration of content (top row). We reverse the model's own rollouts to form corrective trajectories that exemplify how to undo drifted states. A DAgger-style aggregation loop fine-tunes the model on these corrective samples using the original diffusion objective—no bi-directional teacher or distribution matching objectives required. The resulting model exhibits more stable, diverse long-horizon generations (bottom row).

Abstract

Autoregressive video models are promising for world modeling via next-frame prediction, but they suffer from exposure bias: a mismatch between training on clean contexts and inference on self-generated frames, causing errors to compound and quality to drift over time. We introduce Backwards Aggregation (BAgger), a self-supervised scheme that constructs corrective trajectories from the model's own rollouts, teaching it to recover from its mistakes. Unlike prior approaches that rely on few-step distillation and distribution-matching losses, which can hurt quality and diversity, BAgger trains with standard score or flow matching objectives, avoiding large teachers and long-chain back-propagation through time. We instantiate BAgger on causal diffusion transformers and evaluate on text-to-video, video extension, and multi-prompt generation, observing more stable long-horizon motion and better visual consistency with reduced drift.

1. Introduction

Video world models are causal generative models designed to predict how the visual world evolves given a set of actions or prompts [2, 4, 9, 22, 36, 46, 47, 49, 67]. Recently, video diffusion models have proven to be a promising ap-proach for world modeling [28]. While earlier video dif-fusion models focus on generating fixed-length clips in a non-causal, i.e., bi-directional, manner [62, 71], the task of world modeling inherently requires causality, leading to a shift in adoption of autoregressive architectures [3, 18, 26, 34, 69, 75].

However, common to all autoregressive generative mod-els, causal video diffusion suffers from exposure bias [45, 54], an effect that leads to rapid deterioration of video quality over time [29, 63, 74, 76]. Exposure bias, also known as drift, stems from a fundamental mismatch be-tween the training and inference objectives. During train-ing, the model learns to condition on *ground truth* context frames from the previous step, providing clean and reli-able context for the next step of generation. In contrast, during inference the model is given *generated* frames from the previous step to generate the next video frames. If the model makes an error during generation, these errors are compounded into subsequent steps, leading to a cascading effect that degrades video quality over time [74].

Prior works that tackle exposure bias usually aim at clos-ing the training–inference gap by aligning the generated dis-tribution from the autoregressive model with a pre-trained bidirectional teacher [14, 29, 42, 55, 70]. However, this im-poses three key limitations: (1) the reliance on a pre-trained
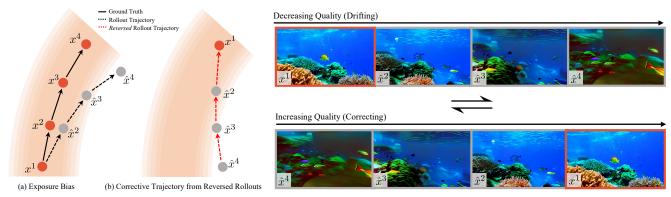
Figure 2. **Backwards Model Rollouts as Corrective Trajectories.** (a) Exposure bias arises from a mismatch between training and inference objectives. During inference, small errors accumulate and cause generations $\hat{x}^{2:4}$ to drift from the real distribution $x^{2:4}$. The orange contours depict regions of higher probability under the video data distribution. (b) Given an AR rollout that may exhibit drift, we treat the reversed clip as a supervision signal for recovering from drifted contexts. This converts the model's own mistakes into training data without requiring an external expert or teacher network.

bi-directional teacher diffusion model [62], (2) the need for back-propagation through time (BPTT) over the entire autoregressive generation process, and (3) the distribution-matching loss itself [58, 72, 73], which is known to be mode seeking and can suppress diversity. Other methods also seek to mitigate drift by injecting noise into context frames [11, 51, 56, 59, 76], but this often degrades temporal consistency and fails to address exposure bias at its core.

We propose a method for overcoming drift by guiding a model to learn how to correct its own mistakes. Drawing inspiration from Dataset Aggregation (DAgger) [53], we iteratively construct and fine-tune a model on a dataset of "corrective trajectories". This teaches the model to be robust to its own mistakes, correcting correcting errors made during generation instead of allowing them to compound to subsequent steps. In the field of imitation learning, where DAgger is often used to great effect, these "corrective trajectories" are gathered through expert intervention, where a human manually corrects mistakes made by the trained policy. However, in the case of video generation, "corrective trajectories", i.e. smooth, continuous transitions from low-quality clips to high-quality clips, cannot be obtained via human intervention. In our work, we leverage the following observation: by reversing a model's own generated rollouts, which start from high-quality real videos and degrade autoregressively in time, we are naturally provided with examples of how the model can correct its own errors. This is uniquely supported by the fact that time-reversed videos remain in-distribution for text-to-video models as long as the prompt is updated to describe the reversed motion.

We show that by repeatedly aggregating more "corrective" samples, the resulting model learns to correct errors made during its own generation process, leading to better generation quality over long horizons compared to teacher/diffusion-forcing training. We also identify how the characteristics of errors change across multiple rounds of

training, and propose a method for modifying the training regime, preventing quality degradation due to training on generated data. We compare our method against prior works performing autoregressive video generation, showing qualitative and quantitative improvements. We summarize our contributions as follows:

- We introduce *Backwards Aggregation (BAgger)*, which leverages reversed model rollouts as self-corrective signals to enhance stability over long-horizon generations.
- Inspired by Dataset Aggregation [53], we iteratively construct and train on a dataset of "corrective trajectories", gradually teaching a model to learn stable generative trajectories while conditioned on drifted states.
- We evaluate our model on diverse tasks—including text-to-video, video extension, and multi-prompt generation—and show improved long-horizon performance in both qualitative and quantitative evaluations.

2. Related Work

Video Diffusion Models. Current state-of-the-art in video generation models have mostly been set by large-scale bidirectional diffusion transformers [8, 48], demonstrating remarkable capabilities in synthesizing high-quality, complex video clips. The core mechanism relies on full spatiotemporal attention applied to all tokens [8], as every video token is denoised simultaneously [5–8, 10, 21, 23–26, 35, 37, 50, 61, 62, 71], generating videos of fixed lengths. However, because these models are inherently non-causal and constrained to fixed-length generation, they are not well suited to serve as true world models.

Autoregressive Video Models. Autoregressive approaches in video generation factorizes the joint distribution over all frames into, $p(x^{1:N}) = \prod_{i=1}^{N} p(x^i | x^{<i})$. This formulation naturally aligns with the causality of time, as video frames are generated sequentially, making autoregres-

(a) Long Text-to-Video Generation

"A detailed portrait of a reindeer in a snowy forest at sunset. The lighting is cinematic and gorgeous... "

(b) Video Extension

Input

"A close-up shot of a young woman driving a car, lost in thought as she gazes ahead. Raindrops blur..."

(c) Multi-Prompt Video

"... a white and orange tabby cat joyfully darting through a dense garden..."   *"... pauses briefly near a patch of tall grass, lowering its body as if to pounce"*

"the cat leaps forward playfully, scattering petals and dry leaves into the air..."   *"... cat slows down, turning its head curiously toward a fluttering butterfly"*

Figure 3. **Applications of BAgger Trained AR Video Model**. (a) Text-to-video generation from a static text prompt, producing a 30-second clip that maintains subject identity and cinematic appearance over time. (b) Video extension, where our model continues an input clip while preserving the input's appearance and motion dynamics. (c) Multi-prompt generation, in which the text conditioning is updated over time, enabling fine-grained, temporally coherent control over the evolving video.

sive models well suited for interactive simulation [29, 55] and world modeling tasks [4, 36, 67]. Conventional autoregressive video models rely on direct next-token prediction of discrete video tokens [9, 34, 52, 64, 65, 69]. However, the performance of discretized AR models often lag behind diffusion models. To address this, recent works have explored training strategies combining autoregression and diffusion [16, 19, 20, 27, 32, 39, 41, 43, 66, 77, 78]. Some works train conditional diffusion models that denoise next frames condition on past clean frames [29, 76], while other approaches introduce per-frame independent noise levels during training [11, 56, 59], allowing for AR inference.

Mitigating Exposure Bias. Several training strategies have been proposed to mitigate exposure bias. Diffusion Forcing (DF) [11] trains a diffusion model with varying per-frame independent noise levels. With DF, adding noise to context frames during training has shown to mitigate drift [59], as the model becomes more robust to perturbations in the context. However, DF models still suffer from error accumulation, since the noised context frames re-

main mismatched from the inference-time context distribution [59, 74]. History Guidance [56] proposes a conditioning mechanism that guides the generation process during inference to remain consistent with its past. Although effective along short horizons, History Guidance can worsen error accumulation, as the model is guided to stay consistent even with drifted frames. Self Forcing [29] takes a more principled approach towards closing the train–test gap by performing autoregressive self-rollout during training. Aligning the distribution of its own rollouts against the distribution of a pre-trained bidirectional teacher, the resulting model is robust to error accumulation for dozens of seconds, with follow-up works extending this to minutes [14, 42, 70]. While effective, this approach fundamentally relies on distribution-matching losses such as score distillation [72, 73] or adversarial objectives [58], which are known to be less stable and accurate for density estimation compared to flow or score-matching objectives [40]. SVI [38] explores this problem through noise injection, but only for the non-causal image-conditioned setting.
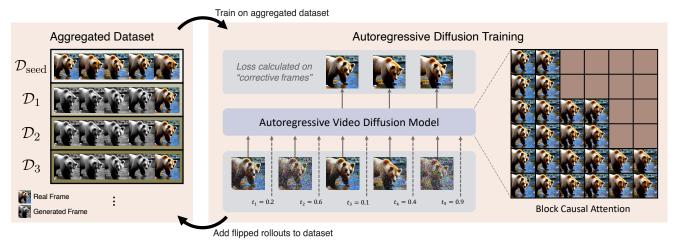
Figure 4. **BAgger Training Loop.** Starting from an initial dataset of ground truth video data $\mathcal{D}_{\text{seed}}$, each round: (1) train a Diffusion Forcing model (policy) on the dataset; (2) sample on-policy rollouts starting from a ground truth frame; (3) reverse them to form corrective trajectories; (4) aggregate with the dataset; (5) continue training with the original diffusion objective. Iterating this process closes the train–test gap by teaching the model to act on its own drifted states.

Dataset Aggregation. Our approach is fundamentally inspired by Dataset Aggregation (DAgger) [53], an established algorithm from the field of imitation learning for addressing exposure bias. Originally designed to mitigate distributional shift in behavioral cloning, DAgger teaches a model to "learn from its own mistakes" through an iterative process. A single round of DAgger, (1) trains a policy with an initial dataset of observations, (2) rolls out the model to collect on-policy states, (3) queries an expert for correct actions in those states, and (4) aggregates the new state-action pairs with the existing dataset to re-train the policy. Our method follows a similar approach, with the key insight that reversing an AR generated video can be viewed as an oracle-free method for obtaining "corrective" actions. Over many such cycles, DAgger progressively builds a dataset that approximates the model's own trajectory distribution in the limit, effectively resolving the train–test mismatch at the heart of the exposure bias problem.

3. Method

3.1. Preliminary: Autoregressive Video Diffusion

Autoregressive video diffusion models combine autoregressive factorization with denoising diffusion models, leveraging the benefits of both [11]. Specifically, they make the factorization $p(x^{1:N}) = \prod_{i=1}^{N} p(x^i|x^{<i})$, and model the conditional distribution of each frame $p(x^i|x^{<i})$ as a diffusion process. This requires training a denoiser network, $\hat{\epsilon}_\theta$ to predict and remove noise ϵ^i from a corrupted frame x_t^i at timestep t. The distinction from non-causal models is that this prediction is conditioned on preceding context frames $x^{<i}$. During training, this context can be kept clean (Teacher Forcing) or have independently sampled noise injected (Diffusion Forcing). Specifically, Diffusion Forc-

ing works by sampling noise levels independently for each frame in a training sequence. A ground truth sequence $x_0^{1:N}$ is corrupted such that each from x_0^i is noised to a different, randomly sampled timestep t_i:

$$x_{t_i}^i = \alpha_{t_i} x_0^i + \sigma_{t_i} \epsilon^i. \tag{1}$$

The denoiser $\hat{\epsilon}_\theta$ is then trained to reverse the diffusion process, predicting the noise e^i conditioned on all previous noisy context frames $x_{t_j}^{j<i}$. The model is trained on a frame-wise denoising objective:

$$\mathcal{L}_{DF}(\theta) = \mathbb{E}_{t_i, x_0, \epsilon^i} \left[||\hat{\epsilon}_\theta(x_{t_i}^i, t_i, x_{t_j}^{j<i}) - \epsilon^i||_2^2 \right] \tag{2}$$

The above training objective naturally enables autoregressive inference, as denoising a noisy frame x_t^i given clean context frames $x_0^{<i}$ is simply a special case covered by the training distribution.

In this work, we train our autoregressive models using the Diffusion Forcing [11] objective. We primarily focus on using a transformer-based architecture [48] and enable autoregressive factorization through the use of causal attention [3, 29, 49, 74], allowing consistent video generation capabilities across extended temporal horizons.

3.2. Backwards Aggregation (BAgger)

Learning Corrective Trajectories. As previously discussed, exposure bias originates from a fundamental mismatch between the model's training and inference-time distributions. During training, the model is exposed exclusively to ground-truth data. It learns to model the distribution $p(x^i|x^{<i})$, where $x^{<i}$ is always drawn from the real data manifold. During inference, the model must sample from $p(x^i|\hat{x}^{<i})$, conditioned on its own imperfect, self-generated outputs $\hat{x}^{<i}$.

To close this train–test gap, the model must be trained to be robust to its own imperfect outputs, learning to correctly model the test-time distribution $p(x^i|\hat{x}^{<i})$. While autoregressive rollout provides the drifted context states $\hat{x}^{<i}$, collecting the corrective examples to train the desired inference time policy $p(x^i|\hat{x}^{<i})$ is non-trivial. In classic behavioral cloning settings [57], the DAgger algorithm collects these examples by querying an expert oracle. For example, in autonomous driving [2, 13, 17, 46], if the policy drifts away from the center of the lane, a human expert can intervene and provide a specific corrective action (e.g. steering towards the center) for the drifted state. However, this data collection process is not directly applicable to the setting of AR video generation. Error accumulation in videos is not a low-dimensional, discrete error, but a subtle, high-dimensional degradation in the latent space, manifested in forms such as progressive over-saturation, over-smoothing, or loss in motion diversity [74]. A human expert cannot feasibly intervene and manually create a set of "corrective" frames that reverses the compounded errors. Therefore, we must rely on a different source to provide expert supervision needed to generated these corrective trajectories.

Reversed Rollouts as Corrective Trajectories. We make the observation that reversing an AR video model's own rollouts provides effective corrective trajectories for training a model that is robust to its own mistakes. Specifically, given some initial ground truth frame $x^1 \sim p_{\text{data}}(x)$, we autoregressively sample a video

$$(x^1, \hat{x}^{2:N}) \sim p_\theta(x^{2:N}|x^1, c) = \prod_{i=2}^{N} p_\theta(x^i|x^{<i}, c), \quad (3)$$

where c is the conditioning text-prompt. Note that each subsequent frame $\hat{x}^{>1}$ is drawn from the distribution of potentially drifted states. By reversing the sampled video $(\hat{x}^{N:2}, x^1)$, we arrive at an example of a corrective video trajectory that provides a ground truth next frame given a set of potential drifted frames as shown in Fig. 7.

A key consideration is that these collected trajectories $(\hat{x}^{N:2}, x^1)$ are time-reversed. However, these corrective states are still valid for two primary reasons. First, we rely on the assumption that the manifold of text-conditioned video data is closed under time reversal. For instance, a video of a person walking backwards is just as valid as a video of someone walking forwards. Consider $\mathcal{M}_{\text{data}}$ as the manifold of video sequences corresponding to high-density regions of the distribution p_{data}. We make the assumption that, if $x^{1:N} \in \mathcal{M}_{\text{data}}$, then $x^{N:1} \in \mathcal{M}_{\text{data}}$. Second, to make $(\hat{x}^{N:2}, x^1)$ a valid training sample, we simply need to modify the original text prompt c into c', where the modified text prompt reflects the time-reversed nature of the sample. For example, the text prompt "a person walking" would instead be replaced by "*a reversed video of* a person walking".

Algorithm 1 BAgger Algorithm

Require: Pre-trained bidirectional diffusion model p_θ
Require: Pre-trained AR diffusion model p_{θ_0}
Require: Aggregated dataset $\mathcal{D}_{\text{agg}} \leftarrow \mathcal{D}_{\text{seed}}$
Require: DAgger rounds K, Videos per round M
1: **for** $k = 0$ to $K - 1$ **do**
2:     $\mathcal{D}_k \leftarrow \emptyset$ ▷ Initialize dataset for round k
3:     **for** $m = 1$ to M **do**   ▷ Sample M drifted videos
4:         Sample $x^1 \sim p_{\text{data}}(x)$, c
5:         $(x^1, \hat{x}^{2:N}) \sim \prod_{i=2}^{N} p_{\theta_k}(x^i|x^{<i}, c)$
6:         $\text{video}_{\text{corr}} \leftarrow (\hat{x}^{N:2}, x^1)$
7:         $c' \leftarrow \text{ModifyPrompt}(c, \text{"in reverse"})$
8:         $\mathcal{D}_k \leftarrow \mathcal{D}_k \cup \{(\text{video}_{\text{corr}}, c')\}$
9:     **end for**
10:     $\mathcal{D}_{\text{agg}} \leftarrow \mathcal{D}_{\text{agg}} \cup \mathcal{D}_k$ ▷ Aggregate corrective traj.
11:     $p_{\theta_{k+1}} \leftarrow \text{DF}(p_\theta, \mathcal{D}_{\text{agg}})$ ▷ Train on aggredage data
12: **end for**
Ensure: Optimized model p_{θ_K}

BAgger Training Loop. Leveraging these corrective trajectories, we implement the BAgger training loop as an iterative training procedure [53]. We begin with an autoregressive video diffusion model p_{θ_0}, pre-trained using a standard Diffusion Forcing [11] objective on a dataset of ground truth video clips $\mathcal{D}_{\text{seed}}$. Then, for each round k of BAgger training we: (1) sample a set of drifted video rollouts under the current AR diffusion model, (2) reverse these sampled rollouts and combine them into an aggregated dataset $\mathcal{D}_{\text{agg}}$, and (3) train a new model on the aggregated dataset following the original diffusion forcing objective. We formalize this iterative training procedure in Algorithm 1.

Architecture. We focus on a diffusion transformer architecture [48] and train our models using the Diffusion Forcing objective (Eq. 2) with a block causal mask [3, 29, 49, 74]. Data drawn from $\mathcal{D}_{\text{seed}}$ is trained with the DF objective on every frame without modification. Corrective data drawn from $\mathcal{D}_k$ is trained with a modified objective, where a set number of frames are chosen as a prefix. These prefix frames are kept clean when fed as input into the diffusion model, and the resulting losses are not calculated on these frames. We treat these prefix frames as the "drifted" states, therefore the model should not learn their distribution, and only learn the distribution conditioned on these states.

4. Experiments

Implementation. We implement BAgger on a pre-trained bi-directional diffusion transformer, namely Wan2.1 1.3B [62]. We modify the model to have block-causal attention [3, 29, 49, 74] and train using the Diffusion Forcing objective. The model is trained on 5s long videos at 16FPS and 832×480 resolution [62]. The model is trained on the
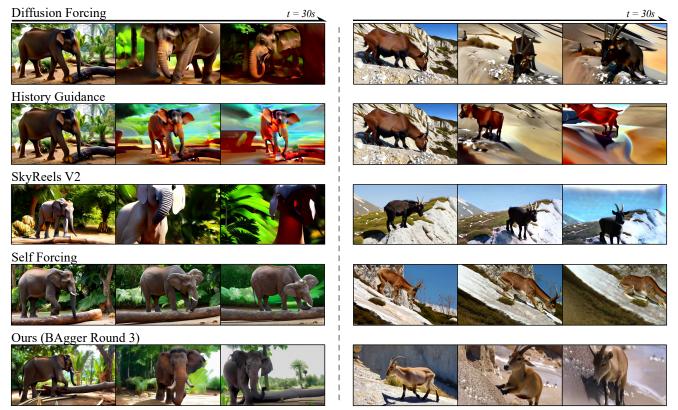
Figure 5. **Qualitative Comparisons.** Side-by-side long-horizon text-to-video results. We compare Diffusion Forcing, History Guidance, SkyReels-V2, Self Forcing, and our BAgger (Round 3). Baseline AR methods exhibit exposure bias artifacts such as over-smoothing, over-saturation, and reduced motion diversity, whereas BAgger better preserves subject identity, scene layout, and local contrast over time. All models share the same architecture with 1.3B parameters. Results show generated frames at 0/15/30s time steps.

latent space produced by a 3D VAE [44, 60], mapping 81 frames per video to 21 latent frames. We opt for a block-causal mask of 3 latent frames per chunk (total of 7 chunks) for our causal diffusion forcing model. Efficient training is also enabled through the use of FlexAttention [15].

Data and Training. We build $\mathcal{D}_{\text{seed}}$ from 55K high-quality clips from the Pexels [1] video dataset, with extended captions sourced from MiraData [33]. The initial diffusion forcing model is trained on this dataset for 16K iterations and a batch size of 96. Please refer to the supplementary materials for a detailed discussion of training configurations and hyperparameters.

For each round of BAgger, we generate 27K corrective trajectories (50% of seed dataset), and directly aggregate all new videos into the next round of training. For each corrective trajectory, we populate the first frame chunk in the model with ground truth frames and perform video extension using the model trained in the last round to generate the 6 leftover chunks. Samples are decoded into pixel space, reversed, then encoded back into the latent space to form corrective trajectories. During training, clean ground-truth clips are trained with standard diffusion objective, where each frame chunk is given independent noise levels. For

corrective clips, the first 4 chunks are kept clean, and the DF objective is only calculated against the remaining 3 chunks. For each round of BAgger, the resulting model p_{θ_k} is fine-tuned from the base Wan2.1 1.3B model for 16K steps.

Long Video Inference. While the base model only supports up to 5 seconds of video generation, long video generation can be enabled through the use of a sliding window. Prior works have explored the use of a rolling KV-cache [14, 29]. However, without explicitly modifying training to accommodate this inference scheme, it may introduce artifacts due to out-of-distribution KV-cache values [29]. To ensure this is not a confounding factor during evaluations, we opt to recompute the KV-cache for every sliding window [3, 74].

4.1. Text-to-Video Generation

Baselines. We compare against prior methods for training autoregressive diffusion video models. Specifically, we make comparisons to Diffusion Forcing [11] and History Guidance [56]. We compare out method against these baselines trained using the seed dataset and equal number of training iterations as our method. For diffusion forcing, we evaluate on two settings, one where all context frames

Method	Global Metrics ↑						Drifting Metrics ↓	
	Subject	Bg. Cons.	Smooth.	Dynamic	Aesthetic	Imaging	$\Delta^{\text{Aesthetic}}$	Δ^{Imaging}
Diffusion Forcing [11] ($\sigma_{\text{test}} = 0$)	80.70	87.74	98.11	**81.25**	53.05	59.97	5.12	7.34
Diffusion Forcing [11] ($\sigma_{\text{test}} = 0.2$)	82.13	88.76	98.37	75.78	53.98	57.49	7.83	12.27
History Guidance [56]	79.15	86.20	98.65	59.38	49.24	54.33	8.83	14.56
Our Method (Multiple Rounds)								
BAgger Round 1	82.69	89.02	**99.03**	61.72	53.84	53.67	4.84	11.83
BAgger Round 2	82.29	88.70	98.49	80.47	54.68	59.98	3.79	5.92
BAgger Round 3	**84.05**	**89.58**	98.61	76.56	**55.35**	**63.41**	**3.29**	**3.57**

Table 1. **Quantitative Comparisons.** VBench evaluation on 50s text-to-video generations averaged over a diverse prompt set. "Global Metrics" (↑) report subject/background consistency, motion smoothness/dynamic degree, and per-frame aesthetic/imaging quality. Because per-frame means can mask drift, we additionally report "Drifting Metrics" (↓): the change (first 20% - average) for aesthetic and imaging scores, lower indicates less degradation. All baselines (Diffusion Forcing with $\sigma_{\text{test}} \in 0, 0.2$, History Guidance) are trained on the same seed data and compute as ours. BAgger improves steadily across rounds, with Round 3 achieving the strongest frame-wise quality and consistency while maintaining comparable motion metrics.

Model	Model Params	Teacher Model?	NFE	Motion Quality ↑	Frame Quality ↑
MAGI-1 [3]	4.5B	No	20	73.11	57.06
SkyReels-V2 [12]	1.3B	No	20	71.87	56.72
Self Forcing [29]	1.3B	Yes (14B)	4	73.93	**64.81**
Ours	1.3B	No	20	**87.59**	59.38

Table 2. **Comparison to open-source AR models.** We report parameters, presence of a teacher, denoising steps (NFE), and motion/frame-quality scores. Despite using no teacher, BAgger attains substantially higher motion quality and competitive frame quality (second only to Self Forcing distilled from a 14B teacher).

are kept clean, and the other where a small amount of noise is added to the context frames during generation. For history guidance, we adopt "fractional history guidance" with a guidance scale of $w = 1.2$. We additionally compare our method against open-sourced AR video models. This includes two autoregressive models MAGI-1![3] and SkyReels-V2 [12], along with Self Forcing [29], a few-step model distilled from Wan2.1 14B. For all multi-step models, we perform inference with 20 denoising steps.

Evaluation Metrics. We compare our method against relevant baselines on the task of long text-to-video generation. For each method, we sample 50s long videos across a diverse set of text prompts covering human/animal motion, natural landscapes, and more [50]. Please refer to the supplementary materials for a detailed discussion on sampling configurations for each baseline. We adopt VBench [30, 31, 79] as our method of evaluation. Specifically, we evaluate each method along 6 distinct dimensions, which includes subject and background consistency, the smoothness and dynamic degree of motion, and the per-frame aesthetic and imaging quality. The per-frame quality metrics for each video are calculated as an average across all frames, therefore it struggles to capture the decay in quality caused by exposure bias. With this in mind, following prior work [76], we additionally report the change in imaging and

aesthetic quality over the course of each long video by taking the difference in scores between the first 20% of frames and the average across all frames in each generated video.

Qualitative results. Figure 5 presents side-by-side text-to-video comparisons. BAgger maintains subject identity, scene layout, and local contrast over substantially longer horizons than baseline AR methods. While exposure bias manifests differently across approaches, common failure modes include over-smoothing of foreground and background details, over-saturation, and reduced motion diversity. Diffusion Forcing exhibits pronounced over-smoothing and over-saturation, with SkyReels-V2 showing the same tendencies. History Guidance, by tightly anchoring to context frames, over-attends to already drifted states, further intensifying saturation and suppressing motion diversity in later segments. Self Forcing delivers strong initial frame quality (benefiting from a 14B teacher) but progressively loses dynamics, with error accumulation leading to near-static later frames alongside with saturation artifacts also settling in. In contrast, BAgger's trajectories remain visually stable and motion-rich over time.

Quantitative comparisons. We report quantitative comparisons with Diffusion Forcing and History Guidance, trained on the same dataset under a matched compute budget, in Tab. 1. After three BAgger rounds, our model attains the best frame-wise quality and subject/background consistency, while maintaining comparable performance on motion-related metrics. We further benchmark against open-source baselines in Tab. 2. For readability, we group metrics into *motion-related* and *frame-quality* categories. Relative to open-source models, our method yields higher motion quality and stronger per-frame scores than other non-distilled methods, trailing only Self Forcing, which leverages distillation from a 14B teacher to achieve higher frame-wise quality. While Self Forcing produces videos with great frame-wise quality, it achieves this by generating

relatively static video sequences. This lack in motion diversity is reflected in its relatively poor motion quality scores.

4.2. Applications.

Multi-Prompt Generation. A natural application of our method lies in multi-prompt video generation [74], where the textual conditioning can evolve dynamically throughout the generation process. Unlike conventional text-to-video models that rely on a single static prompt [8, 62, 71], our approach allows prompt changes on the fly during autoregressive inference. This enables the model to transition smoothly between distinct scenes or actions while maintaining temporal consistency. As illustrated in Fig. 3, this capability allows the creation of long, coherent videos with fine-grained and diverse control over content and motion through dynamic prompting. Such flexibility has broad implications for world modeling [4, 9, 47, 51, 59] and interactive video generation [29, 55], where user inputs or simulated environmental events can continuously steer the evolving visual narrative in real time.

Video Extension. As a result of diffusion forcing training, our model naturally supports video extension by initializing the key–value cache of the first chunk with embeddings from an input clip, enabling seamless continuation. While our block-causal attention pattern is designed for multi-frame chunks and does not directly support image-to-video, this can be achieved by modifying the causal pattern to include a single latent frame in the first chunk, as done in prior work [29, 55]. As shown in Fig. 3, our model produces smooth and consistent extensions, demonstrating strong generalization to continuation tasks.

4.3. Ablation

Effect over Multiple BAgger Rounds. We ablate across multiple rounds of the BAgger algorithm to study its effect on long-horizon generation. A single round is insufficient to capture the full distribution of drifted states and can even degrade performance relative to training on seed data alone. As shown in Tab. 1, performance improves steadily over multiple rounds. Fig. 6 visualizes examples from each round, sampling three 30-second generations per model and showing the final frame for each sample. The seed-only model exhibits severe over-saturation, while one round of BAgger over-corrects toward under-saturated outputs. By the second round, saturation and contrast stabilize, producing more natural color balance, and the third round yields further improvements in detail and temporal consistency.

5. Discussion

Limitations and Future Work. While our method effectively mitigates drift, drifting artifacts can still appear over very long horizons, as a few rounds of BAgger may not fully

Figure 6. **Ablation Over BAgger Rounds.** Final frames from three 30-second generations per model. Training on seed data alone leads to severe over-saturation, while one round of BAgger over-corrects toward under-saturated outputs. By the second round, color balance and contrast stabilize, and the third round further improves visual fidelity and temporal consistency.

cover the space of drifted states. For fair comparison, we retrain each round from the bidirectional teacher to isolate the effect of aggregation rather than accumulated compute. However, as shown in prior work [68], continuing training from the previous round could further improve performance and efficiency. The ratio between seed and corrective data also remains an open design choice that may influence stability and generalization. Finally, although inference at 20 steps makes KV-cache re-initialization overhead negligible, future work could explore training with rolling KV-cache updates [14, 14, 55] to improve runtime efficiency. While out model is not real-time, future work could explore layering few-step distillation on top of BAgger. Crucially, since BAgger does not rely on distribution-matching losses, distillation would not be constrained to mode-seeking objectives used in prior works [29, 74].

Conclusion. We present Backwards Aggregation (BAgger), a self-supervised approach to mitigating exposure bias in autoregressive video diffusion. By reversing a model's own drifting rollouts to form corrective trajectories, BAgger gradually trains on the model's own induced distribution, eliminating the train-test mismatch at the heart of autoregressive deterioration. Implemented with a causal diffusion transformer, it improves long-horizon generation without modifications to the original flow or score-matching objective. Crucially, BAgger avoids distribution-matching losses used in prior work [29, 74], helping preserve motion diversity without sacrificing visual fidelity, as supported by our qualitative and quantitative evaluations.

References

[1] Pexels: Free stock photos, royalty free images, and copyright-free pictures. https://www.pexels.com/. Accessed: 2025-11-11. 6, 1

[2] Nvidia Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Samuel Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Kl'ar, Grace Lam, Shiyi Lan, Laura Leal-Taixé, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum A. Reda, Xiao-Shuai Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne P. Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Rajan Varghese, Hao Wang, Haoxiang Wang, Hengyi Wang, Tingwei Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yuan Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical ai. *ArXiv*, abs/2501.03575, 2025. 1, 5

[3] Sand. ai, Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, W. Q. Zhang, Weifeng Luo, Xiaoyang Kang, Yuchen Sun, Yue Cao, Yunpeng Huang, Yutong Lin, Yuxin Fang, Zewei Tao, Zheng Zhang, Zhongshu Wang, Zixun Liu, Dai Shi, Guoli Su, Hanwen Sun, Hong-Fei Pan, Jie Wang, Jiexin Sheng, Mingyan Cui, Min Hu, Ming Yan, Shucheng Yin, Si-Hui Zhang, Tingting Liu, Xi Yin, Xiaoyu Yang, Xin Song, Xuan Hu, Yankai Zhang, and Yu-Qian Li. Magi-1: Autoregressive video generation at scale. *ArXiv*, abs/2505.13211, 2025. 1, 4, 5, 6, 7

[4] Philip J. Ball, Jakob Bauer, Frank Belletti, Bethanie Brownfield, Ariel Ephrat, Shlomi Fruchter, Agrim Gupta, Kristian Holsheimer, Aleksander Holynski, Jiri Hron, Christos Kaplanis, Marjorie Limont, Matt McGill, Yanko Oliveira, Jack Parker-Holder, Frank Perbet, Guy Scully, Jeremy Shar, Stephen Spencer, Omer Tov, Ruben Villegas, Emma Wang, Jessica Yung, Cip Baetu, Jordi Berbel, David Bridson, Jake Bruce, Gavin Buttimore, Sarah Chakera, Bilva Chandra, Paul Collins, Alex Cullum, Bogdan Damoc, Vibha Dasagi, Maxime Gazeau, Charles Gbadamosi, Woohyun Han, Ed Hirst, Ashyana Kachra, Lucie Kerley, Kristian Kjems, Eva Knoepfel, Vika Koriakin, Jessica Lo, Cong Lu, Zeb Mehring, Alex Moufarek, Henna Nandwani, Valeria Oliveira, Fabio Pardo, Jane Park, Andrew Pierson, Ben Poole, Helen Ran, Tim Salimans, Manuel Sanchez, Igor Saprykin, Amy Shen, Sailesh Sidhwani, Duncan Smith, Joe Stanton, Hamish Tomlinson, Dimple Vijaykumar, Luyu Wang, Piers Wingfield, Nat Wong, Keyang Xu, Christopher Yew, Nick Young, Vadim Zubov, Douglas Eck, Dumitru Erhan, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghara-

mani, Raia Hadsell, Aäron van den Oord, Inbar Mosseri, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 3: A new frontier for world models. 2025. 1, 3, 8

[5] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation. *SIGGRAPH Asia 2024 Conference Papers*, 2024. 2

[6] A. Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, and Dominik Lorenz. Stable video diffusion: Scaling latent video diffusion models to large datasets. *ArXiv*, abs/2311.15127, 2023.

[7] A. 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. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22563–22575, 2023.

[8] 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. 2, 8

[9] Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal M. P. Behbahani, Stephanie Chan, Nicolas Manfred Otto Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktaschel. Genie: Generative interactive environments. *ArXiv*, abs/2402.15391, 2024. 1, 3, 8

[10] Shengqu Cai, Ceyuan Yang, Lvmin Zhang, Yuwei Guo, Junfei Xiao, Ziyan Yang, Yinghao Xu, Zhenheng Yang, Alan L. Yuille, Leonidas J. Guibas, Maneesh Agrawala, Lu Jiang, and Gordon Wetzstein. Mixture of contexts for long video generation. *ArXiv*, abs/2508.21058, 2025. 2

[11] Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *ArXiv*, abs/2407.01392, 2024. 2, 3, 4, 5, 6, 7

[12] Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Juncheng Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zhenghao Chen, Chengchen Ma, Weiming Xiong, Wei Wang, Nuo Pang, Kang Kang, Zhi-Xin Xu, Yuzhe Jin, Yu Liang, Yu-Ning Song, Peng Zhao, Bo Xu, Di Qiu, Debang Li, Zhengcong Fei, Yang Li, and Yahui Zhou. Skyreels-v2: Infinite-length film generative model. *ArXiv*, abs/2504.13074, 2025. 7, 1

[13] Li Chen, Peng Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:10164–10183, 2023. 5

[14] Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-

forcing++: Towards minute-scale high-quality video generation. *ArXiv*, abs/2510.02283, 2025. 1, 3, 6, 8

[15] Juechu Dong, Boyuan Feng, Driss Guessous, Yanbo Liang, and Horace He. Flex attention: A programming model for generating optimized attention kernels. *ArXiv*, abs/2412.05496, 2024. 6, 1

[16] Kaifeng Gao, Jiaxin Shi, Hanwang Zhang, Chunping Wang, Jun Xiao, and Long Chen. Ca2-vdm: Efficient autoregressive video diffusion model with causal generation and cache sharing. *ArXiv*, abs/2411.16375, 2024. 3

[17] Yuan Gao, Mattia Piccinini, Yuchen Zhang, Dingrui Wang, Korbinian Moller, Roberto Brusnicki, Baha Zarrouki, Alessio Gambi, Jan Frederik Totz, Kai Storms, Steven Peters, Andrea Stocco, Bassam Alrifaee, Marco Pavone, and Johannes Betz. Foundation models in autonomous driving: A survey on scenario generation and scenario analysis. *ArXiv*, abs/2506.11526, 2025. 5

[18] Songwei Ge, Thomas Hayes, Harry Yang, Xiaoyue Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *European Conference on Computer Vision*, 2022. 1

[19] Yuchao Gu, Weijia Mao, and Mike Zheng Shou. Long-context autoregressive video modeling with next-frame prediction. *ArXiv*, abs/2503.19325, 2025. 3

[20] Yuwei Guo, Ceyuan Yang, Ziyan Yang, Zhibei Ma, Zhijie Lin, Zhenheng Yang, Dahua Lin, and Lu Jiang. Long context tuning for video generation. *ArXiv*, abs/2503.10589, 2025. 3

[21] 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 *European Conference on Computer Vision*, 2023. 2

[22] David R Ha and Jürgen Schmidhuber. World models. *ArXiv*, abs/1803.10122, 2018. 1

[23] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, David-Pur Moshe, Eitan Richardson, E. I. 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. *ArXiv*, abs/2501.00103, 2024. 2

[24] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey A. Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *ArXiv*, abs/2210.02303, 2022.

[25] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. *ArXiv*, abs/2204.03458, 2022.

[26] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *ArXiv*, abs/2205.15868, 2022. 1, 2

[27] Jinyi Hu, Shengding Hu, Yuxuan Song, Yufei Huang, Mingxuan Wang, Hao Zhou, Zhiyuan Liu, Wei-Ying Ma, and Maosong Sun. Acdit: Interpolating autoregressive conditional modeling and diffusion transformer. *ArXiv*, abs/2412.07720, 2024. 3

[28] Xun Huang. Towards video world models, 2025. 1

[29] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *ArXiv*, abs/2506.08009, 2025. 1, 3, 4, 5, 6, 7, 8

[30] 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. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21807–21818, 2023. 7, 2

[31] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *ArXiv*, abs/2411.13503, 2024. 7, 2

[32] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Zhuang Nan, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. *ArXiv*, abs/2410.05954, 2024. 3

[33] Xu Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *ArXiv*, abs/2407.06358, 2024. 6, 1

[34] D. Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vighnesh Birodkar, Yong Cheng, Ming-Chang Chiu, Josh Dillon, Irfan Essa, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David C. Minnen, David A. Ross, Grant Schindler, Mikhail Sirotenko, Kihyuk Sohn, Krishna Somandepalli, Huisheng Wang, Jimmy Yan, Ming Yang, Xuan Yang, Bryan Seybold, and Lu Jiang. Videopoet: A large language model for zero-shot video generation. *ArXiv*, abs/2312.14125, 2023. 1, 3

[35] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jia-Liang Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fan Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Peng-Yu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhen Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Z. Xu, Yang-Dan Tao, Qinglin Lu, Songtao Liu, Daquan Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models. *ArXiv*, abs/2412.03603, 2024. 2

[36] Jiaqi Li, Junshu Tang, Zhi-Ting Xu, Longhuang Wu, Yuan Zhou, Shuai Shao, Tianbao Yu, Zhiguo Cao, and Qinglin Lu. Hunyuan-gamecraft: High-dynamic interactive game video generation with hybrid history condition. *ArXiv*, abs/2506.17201, 2025. 1, 3

[37] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *ArXiv*, abs/2406.11838, 2024. 2

[38] Wuyang Li, Wentao Pan, Po-Chien Luan, Yang Gao, and Alexandre Alahi. Stable video infinity: Infinite-length video generation with error recycling. *ArXiv*, abs/2510.09212, 2025. 3

[39] Zongyi Li, Shujie Hu, Shujie Liu, Long Zhou, Jeongsoo Choi, Lingwei Meng, Xun Guo, Jinyu Li, Hefei Ling, and Furu Wei. Arlon: Boosting diffusion transformers with autoregressive models for long video generation. *ArXiv*, abs/2410.20502, 2024. 3

[40] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *ArXiv*, abs/2210.02747, 2022. 3, 1

[41] Haozhe Liu, Shikun Liu, Zijian Zhou, Mengmeng Xu, Yanping Xie, Xiao Han, Juan C. P'erez, Ding Liu, Kumara Kahatapitiya, Menglin Jia, Jui-Chieh Wu, Sen He, Tao Xiang, Jürgen Schmidhuber, and Juan-Manuel P'erez-R'ua. Mardini: Masked autoregressive diffusion for video generation at scale. *ArXiv*, abs/2410.20280, 2024. 3

[42] Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *ArXiv*, abs/2509.25161, 2025. 1, 3

[43] Yaofang Liu, Yumeng Ren, Xiaodong Cun, Aitor Artola, Yang Liu, Tieyong Zeng, Raymond H. Chan, and Jean-Michel Morel. Redefining temporal modeling in video diffusion: The vectorized timestep approach. *ArXiv*, abs/2410.03160, 2024. 3

[44] Romain Lopez, Pierre Boyeau, Nir Yosef, Michael I. Jordan, and Jeffrey Regier. Decision-making with auto-encoding variational bayes. *ArXiv*, abs/2002.07217, 2020. 6

[45] Mang Ning, Mingxiao Li, Jianlin Su, A. A. Salah, and Itir Onal Ertugrul. Elucidating the exposure bias in diffusion models. *ArXiv*, abs/2308.15321, 2023. 1

[46] NVIDIA, Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, Prithvijit Chattopadhyay, Mike Chen, Yongxin Chen, Yu Chen, Shuai Cheng, Yin Cui, Jenna Diamond, Yifan Ding, Jiaojiao Fan, Linxi Fan, Liang Feng, Francesco Ferroni, Sanja Fidler, Xiao Fu, Ruiyuan Gao, Yunhao Ge, Jinwei Gu, Aryaman Gupta, Siddharth Gururani, Imad El Hanafi, Ali Hassani, Zekun Hao, Jacob Huffman, Joel Jang, Pooya Jannaty, Jan Kautz, Grace Lam, Xuan Li, Zhaoshuo Li, Maosheng Liao, Chen-Hsuan Lin, Tsung-Yi Lin, Yen-Chen Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Yifan Lu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Seungjun Nah, Yashraj Narang, Abhijeet Panaskar, Lindsey Pavao, Trung Pham, Morteza Ramezanali, Fitsum Reda, Scott Reed, Xuanchi Ren, Haonan Shao, Yue Shen, Stella Shi, Shuran Song, Bartosz Stefaniak, Shangkun Sun, Shitao Tang, Sameena Tasmeen, Lyne Tchapmi, Wei-Cheng Tseng, Jibin Varghese, Andrew Z. Wang, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Jiashu Xu, Dinghao Yang, Xiaodong Yang, Haotian Ye, Seonghyeon Ye, Xiaohui Zeng, Jing Zhang, Qinsheng Zhang, Kaiwen Zheng, Andrew Zhu, and Yuke Zhu. World simulation with video foundation models for physical ai, 2025. 1, 5

[47] Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, Stephen Spencer, Jessica Yung, Michael Dennis, Sultan Kenjeyev, Shangbang Long, Vlad Mnih, Harris Chan, Maxime Gazeau, Bonnie Li, Fabio Pardo, Luyu Wang, Lei Zhang, Frederic Besse, Tim Harley, Anna Mitenkova, Jane Wang, Jeff Clune, Demis Hassabis, Raia Hadsell, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 2: A large-scale foundation world model. 2024. 1, 8

[48] William S. Peebles and Saining Xie. Scalable diffusion models with transformers. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4172–4182, 2022. 2, 4, 5

[49] Ryan Po, Yotam Nitzan, Richard Zhang, Berlin Chen, Tri Dao, Eli Shechtman, Gordon Wetzstein, and Xun Huang. Long-context state-space video world models. *ArXiv*, abs/2505.20171, 2025. 1, 4, 5

[50] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, Geet Sethi, Guan Pang, Haoyu Ma, Ishan Misra, Ji Hou, Jialiang Wang, Ki ran Jagadeesh, Kunpeng Li, Luxin Zhang, Mannat Singh, Mary Williamson, Matt Le, Matthew Yu, Mitesh Kumar Singh, Peizhao Zhang, Peter Vajda, Quentin Duval, Rohit Girdhar, Roshan Sumbaly, Sai Saketh Rambhatla, Sam S. Tsai, Samaneh Azadi, Samyak Datta, Sanyuan Chen, Sean Bell, Sharadh Ramaswamy, Shelly Sheynin, Siddharth Bhattacharya, Simran Motwani, Tao Xu, Tianhe Li, Tingbo Hou, Wei-Ning Hsu, Xi Yin, Xiaoliang Dai, Yaniv Taigman, Yaqiao Luo, Yen-Cheng Liu, Yi-Chiao Wu, Yue Zhao, Yuval Kirstain, Zecheng He, Zijian He, Albert Pumarola, Ali K. Thabet, Artsiom Sanakoyeu, Arun Mallya, Baishan Guo, Boris Araya, Breena Kerr, Carleigh Wood, Ce Liu, Cen Peng, Dimitry Vengertsev, Edgar Schonfeld, Elliot Blanchard, Felix Juefei-Xu, Fraylie Nord, Jeff Liang, John Hoffman, Jonas Kohler, Kaolin Fire, Karthik Sivakumar, Lawrence Chen, Licheng Yu, Luya Gao, Markos Georgopoulos, Rashel Moritz, Sara K. Sampson, Shikai Li, Simone Parmeggiani, Steve Fine, Tara Fowler, Vladan Petrovic, and Yuming Du. Movie gen: A cast of media foundation models. *ArXiv*, abs/2410.13720, 2024. 2, 7

[51] Julian Quevedo, Quinn McIntyre, Spruce Campbell, Xinlei Chen, and Robert Wachen. Oasis: A universe in a transformer. 2024. 2, 8

[52] Shuhai Ren, Shuming Ma, Xu Sun, and Furu Wei. Next block prediction: Video generation via semi-autoregressive modeling. *ArXiv*, abs/2502.07737, 2025. 3

[53] Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics*, 2010. 2, 4, 5

[54] Florian Schmidt. Generalization in generation: A closer look at exposure bias. In *Conference on Empirical Methods in Natural Language Processing*, 2019. 1

[55] Joonghyuk Shin, Zhengqi Li, Richard Zhang, Jun-Yan Zhu, Jaesik Park, Eli Schechtman, and Xun Huang. Motionstream: Real-time video generation with interactive motion controls. 2025. 1, 3, 8

[56] Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. *ArXiv*, abs/2502.06764, 2025. 2, 3, 6, 7, 1

[57] Faraz Torabi, Garrett Warnell, and Peter Stone. Behavioral cloning from observation. In *International Joint Conference on Artificial Intelligence*, 2018. 5

[58] Individualized Treat and Jinsung Yoon. Generative adversarial nets. 2018. 2, 3

[59] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. *ArXiv*, abs/2408.14837, 2024. 2, 3, 8

[60] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Neural Information Processing Systems*, 2017. 6

[61] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and D. Erhan. Phenaki: Variable length video generation from open domain textual description. *ArXiv*, abs/2210.02399, 2022. 2

[62] 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, Xiaofeng Meng, Ningying Zhang, Pandeng Li, Ping Wu, Ruihang Chu, Rui Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenchao Zhou, Wente Wang, Wen Shen, Wenyuan Yu, Xianzhong Shi, Xiaomin Huang, Xin Xu, Yan Kou, Yan-Mei 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, Zhengbin Han, Zhigang Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *ArXiv*, abs/2503.20314, 2025. 1, 2, 5, 8

[63] Jing Wang, Fengzhuo Zhang, Xiaoli Li, Vincent Y. F. Tan, Tianyu Pang, Chao Du, Aixin Sun, and Zhuoran Yang. Error analyses of auto-regressive video diffusion models: A unified framework. *ArXiv*, abs/2503.10704, 2025. 1

[64] Yuqing Wang, Tianwei Xiong, Daquan Zhou, Zhijie Lin, Yang Zhao, Bingyi Kang, Jiashi Feng, and Xihui Liu. Loong: Generating minute-level long videos with autoregressive language models. *ArXiv*, abs/2410.02757, 2024. 3

[65] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. *ArXiv*, abs/1906.02634, 2019. 3

[66] Wenming Weng, Ruoyu Feng, Yanhui Wang, Qi Dai, Chunyu Wang, Dacheng Yin, Zhiyuan Zhao, Kai Qiu, Jianmin Bao, Yuhui Yuan, Chong Luo, Yueyi Zhang, and Zhiwei Xiong. Art•v: Auto-regressive text-to-video generation with diffusion models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 7395–7405, 2023. 3

[67] Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. *ArXiv*, abs/2506.05284, 2025. 1, 3

[68] Xiaomeng Xu, Yifan Hou, Zeyi Liu, and Shuran Song. Compliant residual dagger: Improving real-world contact-rich manipulation with human corrections. *ArXiv*, abs/2506.16685, 2025. 8

[69] Wilson Yan, Yunzhi Zhang, P. Abbeel, and A. Srinivas. Videogpt: Video generation using vq-vae and transformers. *ArXiv*, abs/2104.10157, 2021. 1, 3

[70] Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Ying-Cong Chen, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation. *ArXiv*, abs/2509.22622, 2025. 1, 3

[71] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihan Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. *ArXiv*, abs/2408.06072, 2024. 1, 2, 8

[72] Tianwei Yin, Michael Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6613–6623, 2023. 2, 3

[73] Tianwei Yin, Michael Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. *ArXiv*, abs/2405.14867, 2024. 2, 3

[74] Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Frédo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22963–22974, 2024. 1, 3, 4, 5, 6, 8

[75] Lijun Yu, José Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David C. Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. Language model beats diffusion – tokenizer is key to visual generation. 2023. 1

[76] Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. 2025. 1, 2, 3, 7

[77] Tianyuan Zhang, Sai Bi, Yicong Hong, Kai Zhang, Fujun Luan, Songlin Yang, Kalyan Sunkavalli, William T. Freeman, and Hao Tan. Test-time training done right. *ArXiv*, abs/2505.23884, 2025. 3

[78] Yuan Zhang, Jiacheng Jiang, Guoqing Ma, Zhiying Lu, Haoyang Huang, Jian min Yuan, and Nan Duan. Generative pre-trained autoregressive diffusion transformer. *ArXiv*, abs/2505.07344, 2025. 3

[79] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *ArXiv*, abs/2503.21755, 2025. 7, 2

Mitigating Drift in Autoregressive Video Diffusion Models
BAgger: Backwards Aggregation for

Supplementary Material

A. Video Results

Supplementary Website. Because static figures in the main manuscript cannot fully convey our video results, we provide corresponding videos for each main result on the supplementary website. These can be accessed via the attached HTML file. Due to the 200MB submission size limit and the length of our videos, many clips have been compressed and may appear lower quality than the original generations.

B. Implementation Details

B.1. Training Configurations

Model. We primarily build our method upon the Wan2.1 [62] and Self Forcing [29] codebases. We primarily work with the Wan2.1 1.3B Text-to-Video model as our base model, fine-tuning it into a block-causal model using FlexAttention [15]. As Wan2.1 1.3B supports bidirectional generation of 21 latent frames, we follow prior works [29, 74] and use a chunk size of 3 frames per chunk, resulting in 7 total chunks. This configuration provides a reasonable balance between parallelism and causal granularity. We apply the same chunking and causal masking scheme to all autoregressive baselines to ensure a fair architectural comparison.

Data. We use a subset of the Pexels video dataset [1] for causal fine-tuning and generation of corrective trajectories. Specifically, we use the Pexels subset from the MiraData [33] dataset, and use the "dense prompt" field as the corresponding text for each corresponding video clip. Due to varying FPS and clip lengths in the dataset, we subsample videos that have higher than 16FPS to match the frame rate of our video model (16FPS) and randomly sample a 5 seconds clip from each data sample. For generation of corrective trajectories, we take the initial 9 frames = 3 latent frames = 1 chunk, from a chosen video clip and perform rollouts conditioned on an initial ground truth chunk and the corresponding text prompt.

Compute and Hyperparameters. We perform all of our experiments on 16 NVIDIA H100 GPUs (80GB VRAM per GPU). For each round of training, we perform 16,000 training iterations at a batch size of 96, using a constant learning rate of $4e^{-6}$. We follow the original Wan2.1 flow matching [40] objective, and adopt a time step shift in the form, $t'(k, t) = (kt/1000)/(1 + (k - 1)(t/1000)) \cdot 1000$ and a shift factor $k = 8$.

Model	Smooth.	Dynamic	Aesthetic	Imaging
MAGI-1 [3]	99.35	46.88	55.69	58.43
SkyReels-V2 [12]	99.20	44.53	55.60	57.84
Self Forcing [29]	98.64	49.22	58.33	71.29
Ours	98.61	76.56	55.35	63.41

Table 3. Extended results comparing our method against relevant open baseline models. Our model shows much better motion dynamics while maintaining simialr levels of motion smoothness. For per-frame quality metrics, on average, our method is only second to Self Forcing which is distilled from a 14B base model. Our method outperforms other AR diffusion models (MAGI-1 and SkyReels) in imaging quality, and has similar aesthetic performance, despite being fine-tuned on a much smaller dataset.

B.2. Inference Configurations

We perform inference for all multi-step AR diffusion models (ours and baselines) at 20 inference steps per chunk. For our model, we perform inference with a classifier-free guidance scale of $w = 6$, and a time step shift coefficient of $k = 8$. For corrective trajectory generations, the same inference configurations are used. For long video generations with sliding windows, we perform KV-cache reinitialization at every sliding window step to eliminate any potential issues caused by OOD KV-cache values as mentioned in Xun et al. [29]. Each sliding window instance, we keep 12 latent frames from the tail end of the previous generation as the initial context of the next set of generations. For multi-prompt generations, we simply change the text-prompt conditioning during the sliding window change.

C. Evaluations

C.1. Baseline Details

History Guidance. We implement a version of History Guidance resembling Song et al. [56], performing joint guidance between the text-conditioned prediction and fractional history guidance as defined in [56]. The resulting velocity prediction used takes the form,

$$v_{\text{hg}}(x_t^i, t, c) = v(x^{<i}, t, \emptyset)$$
$$+ w_{\text{text}}[v(x^{<i}, t, c) - v(x^{<i}, t, \emptyset)] \quad (4)$$
$$+ w_{\text{hg}}[v(x_p^{<i}, t, \emptyset) - v(\emptyset^{<i}, t, \emptyset)].$$

Where w_{text}, and w_{hg} are the guidance scale for the text conditioned and history conditioned terms, $p = 800$ is a partial noise level chosen for fractional history guidance, c is the text prompt, and $\emptyset$ represents the empty string, and $\emptyset^{<i}$ randomly sampled noise in the same shape as the context. In our evaluations we use $w_{\text{text}} = 6$ and $w_{\text{hg}} = 1.2$.

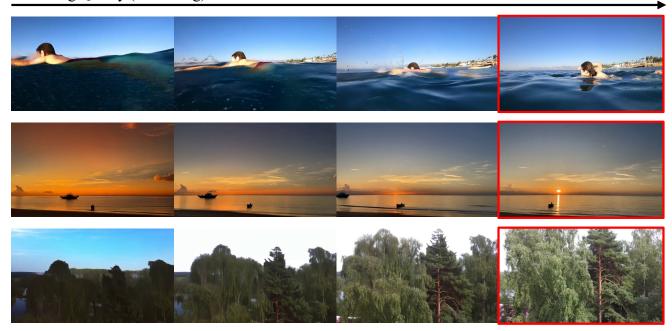Increasing Quality (Correcting)

Figure 7. More examples of corrective trajectories gathered from first round of BAgger. Ground truth frames are highlighted in red. Note that the above sequences show generated frames in reverse order, therefore the order above is the order in which the model is trained on.

C.2. Evaluations

Metric Details. We base our evaluations on the custom evaluation suite of VBench [30, 31, 79], which spans the 6 dimensions. As mentioned for Tab. 1, we also report drifting metrics which measures the drop in per-frame quality metrics between the first 20% of all frames and the average across all frames. We also report the full metrics for Tab. 2 in Tab. 3. Note that the results in Tab. 2 were calculated by taking the averages between the two motion metrics, and two frame-wise quality metrics respectively.

D. Additional Results

Rollout Examples. We provide additional examples of corrective/drifting trajectories generated during the BAgger algorithm. Our model takes in a set of ground truth frames and generated a set of (potentially) drifted extensions of these frames by rolling out the AR model trained from the previous round.

E. Ethical Considerations.

Video generative models inherently carry substantial potential for misuse. High-fidelity outputs can enable convincing deepfakes that are increasingly difficult to distinguish from real footage, posing risks for misinformation, manipulation, and privacy. Without careful governance and responsible deployment, these models may also amplify harmful societal biases present in their training data. At the same time,

they offer meaningful benefits, such as powering world models for robotics, enhancing accessibility, and supporting safer real-world decision-making. We acknowledge this duality and emphasize the importance of continued research, policy development, and safeguards that mitigate negative impacts while enabling socially beneficial uses.