MITIGATING COMPOUNDING ERROR VIA VIDEO REPRESENTATION REGULARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Video diffusion-based world models enable long autoregressive video generation for robotics, autonomous driving and simulation tasks, yet sliding-window autoregressive inference suffers from severe error accumulation that degrades frame quality over time. Although this phenomenon has been widely observed, the underlying mechanism of compounding error and how to achieve stable long-horizon generation remain largely unresolved. In this paper, we investigate the internal representation dynamics of video world models and discover that compounding error is tightly coupled with dimensional collapse of hidden representations. Specifically, the effective rank of model representations sharply decreases at the onset of generation drift, revealing a strong connection between representational degradation and long-term rollout instability. Furthermore, we find that pure training data scaling fails to boost model resistance to error drift, contradicting mainstream scaling paradigms. To address this problem, we propose video representation regularization, a lightweight training constraint that stabilizes latent representations and suppresses iterative error accumulation. Compared with Diffusion Forcing, our method achieves improvements from 38.65 to 53.78 and from 44.37 to 72.08 on the Aesthetic Quality and Imaging Quality metrics of VBench. Our work establishes the first connection between autoregressive video drifting and model internal representations, adopts erank as a quantitative metric for error accumulation, reveals counterintuitive scaling limitations for video world models, and presents a simple yet effective regularization strategy to improve long video generation robustness.

1 INTRODUCTION

World models have recently attracted significant attention from the research community and technology companies due to their advantages in areas such as data synthesis, model-based planning, and simulation. Among these, driven by the continuous advancement of video generation technology, video world models have demonstrated tremendous influence and promising prospects in domains including autonomous driving (Hu et al., 2023; Ren et al., 2025), navigation (Bar et al., 2024), robotic manipulation (Wu et al., 2024; Azzolini et al., 2025; Liao et al., 2025; Maes et al., 2026), and games (Valevski et al., 2024; Decart et al., 2024; Che et al., 2024; Guo et al., 2025; Yu et al., 2025a). Video diffusion models, in particular, have become an important implementation pathway for video world models owing to their superior capabilities in high-resolution settings. Some works (Chen et al., 2024a; Huang et al., 2025; Jin et al., 2025) have further extended them into autoregressive diffusion models, making the generation of infinitely long videos possible.

However, due to the high-dimensional nature of video data, video world models struggle to achieve context lengths on the order of millions of tokens as easily as text-based models do. In contrast, a context consisting of merely tens of frames is sufficient to exhaust the GPU memory of a video world model. To tackle this limitation, the sliding window strategy is widely adopted: previously generated frames act as conditioning inputs for the autoregressive generation of subsequent frames. Nevertheless, these generated frames are inherently imperfect, which leads to rapid deterioration in video quality after tens of iterative generation steps, a phenomenon termed error accumulation, also referred to as drift, exposure bias, or compounding error. While numerous existing studies have proposed diverse techniques to alleviate this issue, a unified and adequately effective solution

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

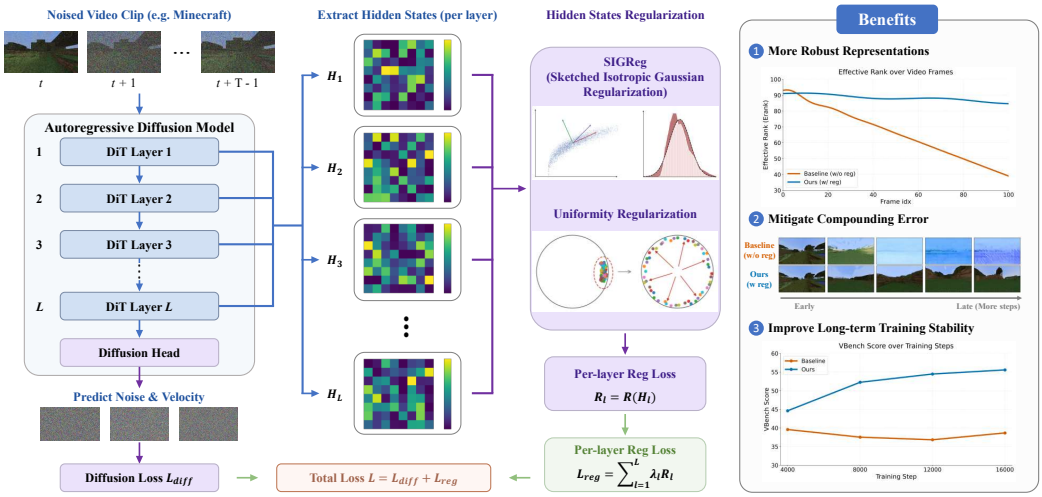


Figure 1: Overview of our proposed method. Existing autoregressive diffusion models suffer from compounding errors during long-range generation, and this issue cannot be effectively mitigated by simply scaling up training data. We introduce hidden states regularization into DiT layers to strengthen the model's representation capacity, alleviate compounding errors, and enable consistent performance gains as training proceeds.

is still lacking. Worse still, we find that merely expanding the training dataset fails to mitigate compounding errors; instead, it may aggravate the problem.

By examining videos exhibiting severe error accumulation, we observe that generation often deteriorates into random noise or overexposed frames, indicating a potential degradation of the model's internal representations. Motivated by this observation, we study how representations evolve throughout autoregressive video generation. We quantify the expressiveness of hidden states using effective rank, which captures the degree of dimensional collapse in learned representations. Interestingly, we find that the decline of effective rank closely coincides with the onset of visual collapse, revealing a strong connection between representation degradation and compounding error. Moreover, as the amount of training data increases, effective rank does not improve and even decreases, suggesting that simply scaling data is insufficient to enhance robustness against error accumulation.

Based on the observations above, we introduce **video representation regularization (VRR)**, which aims to improve the model's resistance to error accumulation by regularizing the model's representational capacity during training. We are the first to establish a connection between the error accumulation phenomenon in autoregressive video generation and model representations. We argue that as the amount of training data increases, the model tends to learn shortcuts, leading to the collapse of its representations. By incorporating regularization during training, the model can learn better representations, thereby improving its resistance to error accumulation. On VBench, our method achieves substantial improvements over Diffusion Forcing, increasing Aesthetic Quality from 38.65 to 53.78 and Imaging Quality from 44.37 to 72.08. In summary, the main contributions of this paper are as follows:

- We identify the relationship between video frame quality degradation and model representations. By introducing representation erank into video generation models, we obtain a sound explanation for frame collapse, which also serves as an effective quantitative metric for error accumulation.
- We discover that simply increasing training data fails to improve the model's resistance to error accumulation, which contradicts the conventional wisdom of scaling and provides a valuable reference for both the academic and industrial communities.
- We propose video representation regularization, which enhances the model's resistance to error accumulation through the addition of a simple regularization training term.

2 Related work

2.1 Video Diffusion Model

Diffusion models have become a dominant paradigm for image and video generation due to their strong sample quality and stable training dynamics (Blattmann et al., 2023b; Harvey et al., 2022; Esser et al., 2023; Blattmann et al., 2023a; Chen et al., 2024b; Ho et al., 2022; Singer et al., 2022; Hong et al., 2022; Yang et al., 2024; Wang et al., 2023; Ding et al., 2024; Zheng et al., 2024). Early diffusion generators typically adopt UNet-based denoising backbones (Ho et al., 2022; Blattmann et al., 2023a; Chen et al., 2024b), while recent models increasingly replace them with diffusion transformers (DiT) (Peebles & Xie, 2023) to improve scalability and capture long-range spatio-temporal dependencies. To reduce the computational cost of high-resolution generation, many systems perform denoising in the latent space of a variational autoencoder (VAE) (Kingma et al., 2013), rather than directly in pixel space (Rombach et al., 2022; Blattmann et al., 2023a). However, most existing video diffusion models are trained to generate clips with a fixed number of frames. To extend generation beyond the training horizon, several methods use sliding-window or overlapping-context inference, producing long videos by repeatedly denoising local temporal windows (Jin et al., 2025; Chen et al., 2024a; Yin et al., 2025; Cui et al., 2025; Xie et al., 2024).

2.2 Video World Model and Long Video Generation

World models aim to predict future states conditioned on past observations and, when available, actions (Ha & Schmidhuber, 2018). Existing approaches differ in the space where prediction is performed. Representation-based methods, such as JEPA-style models, first encode visual observations into latent representations and then learn predictive dynamics in that feature space (Bardes et al., 2024; Assran et al., 2025; Zhou et al., 2025; Maes et al., 2026). In contrast, generative world models use video generation models to directly predict future visual observations, often at the pixel or latent-pixel level (Wu et al., 2024; Chen et al., 2025; Parker-Holder et al., 2024; Ball et al., 2025; NVIDIA et al., 2025). This paper focuses on the latter setting, where a video diffusion model is rolled out auto-regressively for long-horizon generation.

Auto-regressive video generation naturally supports variable-length prediction, but it also suffers from compounding error: small mistakes in early generated frames are fed back as conditioning inputs and can gradually accumulate over time (Zhang & Agrawala, 2025; Song et al., 2025; Blattmann et al., 2023a). This phenomenon is closely related to exposure bias and temporal drifting in sequence generation. Prior work has attempted to mitigate this issue through modified sampling schedules (Qiu et al., 2024; Ruhe et al., 2024), frame anchoring (Khachatryan et al., 2023; Weng et al., 2024), and training-time noise injection or error recycle (Li et al., 2026; Po et al., 2026). Despite these improvements, maintaining stable long-horizon dynamics remains challenging, especially when the generated frames also serve as the model's future conditioning context.

2.3 Representation Learning for Generative Models

Recent studies have highlighted the importance of representations in improving generative models. One line of work introduces semantic representations to guide generative training, where pretrained encoders provide high-level supervision to improve generation quality and training efficiency (Edunov et al., 2019; Li et al., 2024). For instance, REPA aligns diffusion features with representations from pretrained visual encoders, encouraging diffusion models to learn more semantic representations (Yu et al., 2025b), while recent representation autoencoder (RAE)-based approaches further demonstrate the benefits of expressive latent representations for diffusion models (Zheng et al., 2025). Beyond representation alignment, representation regularization has also been explored to preserve feature quality and training stability. DispLoss (Wang & He, 2025) introduces representation-based objectives into the denoising process to accelerate convergence and improve sample fidelity. In latent world models, LeWM (Maes et al., 2026) adopts SIGReg (Balestriero & LeCun, 2025) as a regularizer to prevent latent representation collapse and maintain informative dynamics. However, existing studies mainly focus on improving representation quality during training, while the role of representation degradation in long-horizon autoregressive rollout remains unexplored.

3 Core Observations: Error Accumulation Coupled with Representation Collapse and Limits of Data Scaling

3.1 Preliminary: Auto-regressive Video Diffusion

Video Diffusion Model Video diffusion models operate either directly on pixel-level inputs $\boldsymbol{x} \in \mathbb{R}^{L\times H\times W\times 3}$, or on a latent representation $\boldsymbol{z} = \mathcal{E}(\boldsymbol{x})$, where $\mathcal{E}(\cdot)$ is a pretrained variational autoencoder (VAE) encoder. The forward process gradually adds Gaussian noise to the input according to a variance schedule $\{\beta_t\}_{t=1}^T$:

$$q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) = \mathcal{N}(\boldsymbol{z}_t; \sqrt{1-\beta_t}\boldsymbol{z}_{t-1}, \beta_t\mathbf{I}) \tag{1}$$

The model learns to reverse this process by predicting the noise $\boldsymbol{\epsilon}_\theta$ at each step:

$$\mathcal{L} = \mathbb{E}_{t,\boldsymbol{\epsilon},\boldsymbol{z}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t)\|_2^2\right] \tag{2}$$

where $\boldsymbol{z}_t = \sqrt{\bar{\alpha}_t}\,\boldsymbol{z}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

At inference time, new videos are generated by starting from random noise $\boldsymbol{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and iteratively denoising:

$$\boldsymbol{z}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{z}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t)\right) + \sigma_t\boldsymbol{\epsilon} \tag{3}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$.

Diffusion Forcing To enable long video generation, we apply the Diffusion Forcing (Chen et al., 2024a) technique. During training, we randomly add noise to each frame in the entire input video sequence according to the diffusion schedule: $z_t^i = \sqrt{\bar{\alpha}_t}z_0^i + \sqrt{1-\bar{\alpha}_t}\epsilon^i, \epsilon^i \sim \mathcal{N}(0, \mathbf{I})$, where z_t^i represents the noised latent of the i-th frame, and the training objective for action-conditioned autoregressive video models become:

$$\mathcal{L}_{\text{DF}} = \mathbb{E}_{[t],\boldsymbol{\epsilon},\mathbf{z},a}[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_{[t]}, [t], \boldsymbol{a})\|_2^2]$$
$$\boldsymbol{\epsilon} = \{\epsilon^i\}_{i=1}^L, \mathbf{z}_{[t]} = \{z_t^i\}_{i=1}^L$$

where $[t]$ is vector of L timesteps with different $t \in [T]$ for each frame and $\boldsymbol{a}$ is an action sequence $\boldsymbol{a} \in \mathbb{R}^{L\times A}$. The noise prediction model $\boldsymbol{\epsilon}_\theta$ conditioned on both the action sequence $\boldsymbol{a}$ and noised frames $\mathbf{z}_{[t]}$.

3.2 Error Accumulation Is Coupled with Representation Collapse

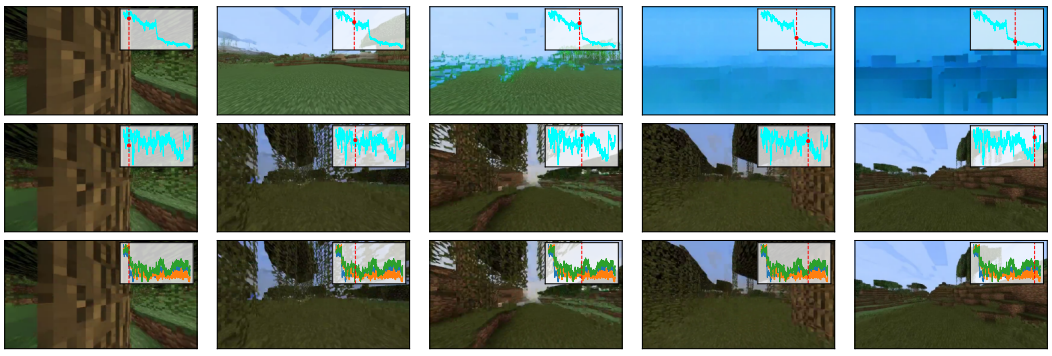


Figure 2: **Top row:** Video frames sampled from a 40k-step checkpoint trained with the vanilla diffusion forcing method. Hidden states from DiT's 7th layer are extracted to analyze the effective rank (Erank), which reveals a direct correlation between visual collapse and abrupt Erank reduction. **Middle row:** Samples generated by a 4k-step checkpoint under the same training scheme; stable Erank values are observed when video content remains coherent. **Bottom row:** Identical experimental settings as the middle row, with frame quality quantified via normalized SSIM, PSNR and LPIPS.

A central finding of our work is that the abrupt onset of error accumulation during autoregressive video generation is intrinsically coupled with representation collapse in the underlying diffusion model. To quantify the fidelity of internal representations, we measure the **effective rank** (Roy & Vetterli, 2007) of the hidden states extracted from intermediate layers of the DiT backbone, which captures how many linearly independent directions are actively used by the model at each generation step.

As illustrated in Figure 2, there is a correspondence between the frame at which generated video quality suddenly drifts—manifesting as semantic incoherence or visual artifacts—and the frame at which the effective rank undergoes a sharp collapse. This suggests that the model's internal representation loses its expressive capacity precisely when error accumulation becomes catastrophic.

To validate that effective rank is a uniquely informative indicator—and not merely one of many correlated signals—we also conduct a comparison against a suite of alternative diagnostics, including standard image quality metrics such as SSIM, PSNR and LPIPS. Crucially, none of these alternatives consistently pinpoints the collapse frame.

3.3 More Data Cannot Cure Error Accumulation

Another central finding of our investigation is that naïvely scaling up the training set does not alleviate error accumulation in long video generation. Following the experimental protocol of VRag (Chen et al., 2025), we curate a dataset of 18,000 Minecraft gameplay sequences, each spanning 1,200 frames. We train a single epoch on this dataset—deliberately avoiding multiple passes to rule out overfitting—and save checkpoints at regular intervals throughout training. For each checkpoint we perform long-video inference and measure the effective rank of the generated sequences as a proxy for temporal diversity and structural coherence.

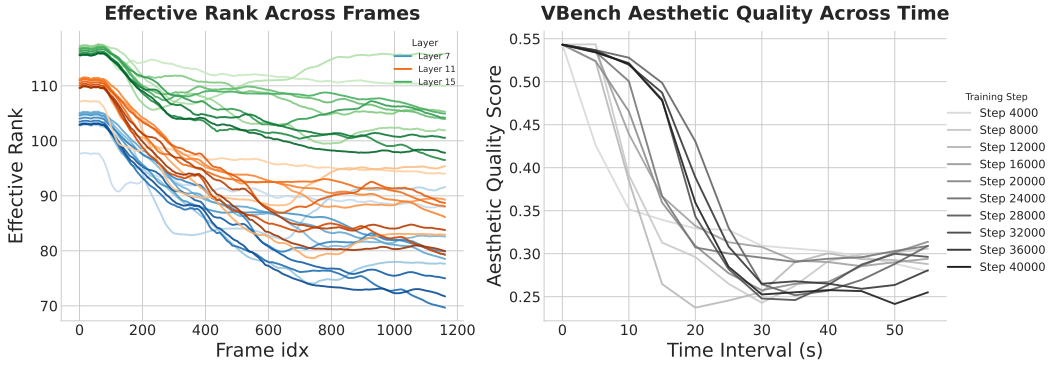


Figure 3: We trained for one epoch on the 18k training set using the vanilla diffusion forcing method and calculated the Erank corresponding to these checkpoints. We observed that the Erank values at step 4000 and step 8000 are generally higher and maintain stability over longer durations. In contrast, increasing the number of training steps not only reduces the overall Erank but also accelerates its degradation during long-range reasoning.

The results are shown in Figure 3. Strikingly, the effective rank peaks after exposure to only $\sim$10% of the training data; every subsequent checkpoint exhibits a monotonic decline. The ability to leverage large-scale data for improved generalization has long been regarded as a defining strength of video generation models. Our experiment challenges this assumption in the context of autoregressive generation: error accumulation is not a data-scarcity problem that can be remedied by collecting more trajectories. Instead, it reflects a structural limitation of the generation paradigm itself, motivating the architectural intervention we introduce in the following section.

4 Video Representation Regularization

4.1 Effective Rank Degradation in Long-term Training

Motivated by the observations in the previous section, our core objective is to improve the effective rank (erank) of the diffusion model over long-term training. Given a hidden state matrix $\mathbf{H} \in \mathbb{R}^{n \times d}$,

erank is defined as

$$\text{erank}(\mathbf{H}) = \exp\left(-\sum_{i=1}^{r} p_i \log p_i\right), \qquad p_i = \frac{\sigma_i(\mathbf{H})}{\sum_{j=1}^{r} \sigma_j(\mathbf{H})}, \tag{4}$$

where $\sigma_i(\mathbf{H})$ denotes the i-th singular value of $\mathbf{H}$ and r is its rank. To maintain a high erank throughout training, we impose a representation regularization loss on the hidden states, encouraging a more uniform singular value distribution.

4.2 Formulation of Video Representation Regularization

We argue that the anomalous decreasing trend of effective rank (erank) as training steps increase stems from shortcut learning in autoregressive diffusion models. When frame contents change mildly, directly copying adjacent frames enables smooth temporal consistency and favorable generation quality. Nevertheless, such frame-copying behavior erodes the amount of valid information contained within latent representations.

To preserve representation robustness and deter the model from over-relying on trivial shortcuts, a straightforward solution is to introduce regularization. To this end, we propose Video Representation Regularization (VRR). Given an n-layer DiT model D, let x denote a video clip input with frame-wise independent noise levels and y denote its corresponding ground-truth video clip. Let $H_1, ..., H_n$ represent the hidden states from each layer. The overall training objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{DF}}(D(x), y) + \sum_i \lambda_i \mathcal{L}_{\text{reg}}(H_i)$$

where $\mathcal{L}_{\text{DF}}$ stands for the vanilla Diffusion Forcing loss, λ_i denotes the regularization weight coefficient, and $\mathcal{L}_{\text{reg}}$ refers to the regularization function.

In this work, we primarily adopt Sigreg and Uniformity loss as our regularization functions. Sigreg regularizes by constraining vectors to an isotropic Gaussian, whereas Uniformity Loss regularizes by constraining vectors to follow a uniform distribution. We also experiment with classic representation regularization terms including Barlow Twins and VICReg. Additionally, we conduct ablation studies that directly utilize effective rank as the regularization function. Detailed experimental results are presented in Section 5.4.

5 Experiments

5.1 Experimental Setup

Datasets We mainly conduct experiments on two datasets. We perform large-scale experiments on the Minecraft dataset and validation experiments on the RealEstate10K (Google, 2018) dataset. For Minecraft experiments, we use MineRL (Guss et al., 2019) to generate 18,000 training sequences following the protocols in VRAG (Chen et al., 2025). Each video contains 1,200 frames. For evaluation, we reserve 20 sequences from the dataset, which are sufficiently long to assess error accumulation in the diffusion model. During inference, we feed the model 100 frames of ground-truth video as the prompt, and the model generates the subsequent 1100 frames. Across all experiments, the window size of the DiT model is set to 20 frames. For the RealEstate10K dataset, we train our UNet model for 1 epoch on its training split and randomly select 50 samples from the test set for evaluation. The window size is set to 4 in this experiment.

Evaluation We adopt VBench (Huang et al., 2024) as our evaluation metric. VBench covers multiple assessment dimensions including text-to-video alignment, human characters, object motion, and background consistency, none of which are relevant to the Minecraft setting. For this reason, we primarily refer to two dimensions: imaging quality and aesthetic quality. These two metrics mainly evaluate whether the frames contain noise or suffer from visual collapse. For the complete VBench scores, please refer to Appendix B.1.

Baseline In addition to Diffusion Forcing, we also implement two baselines, namely Frame Anchor and SVI (Li et al., 2026). Frame Anchor has been adopted in numerous studies as a strategy to mitigate compounding error (Khachatryan et al., 2023; Weng et al., 2024). Its core mechanism involves inserting one completely clean frame as an anchor into the sliding window. SVI improves the model's robustness against compounding error by recapturing errors generated during training and incorporating them into the training samples.

Model Architecture For the Minecraft experiment, we adopt the DiT architecture. As shown in Figure 4, following the designs by previous work (Decart et al., 2024; Zheng et al., 2024; Chen et al., 2025), we decompose the attention mechanism in our DiT into two distinct modules: Spatial Axis Attention and Temporal Axis Attention. The Rotary Position Embedding (RoPE) (Su et al., 2024) is applied in both spatial and temporal dimensions, to enhance the capacity of both attention modules to capture spatial positional dependencies and temporal correlations. Conditioning information, specifically the timestep and action condition, is incorporated into the DiT model via adaptive Layer Normalization (adaLN). For the RealEstate10K dataset, following (Song et al., 2025), we employ the UNet architecture.

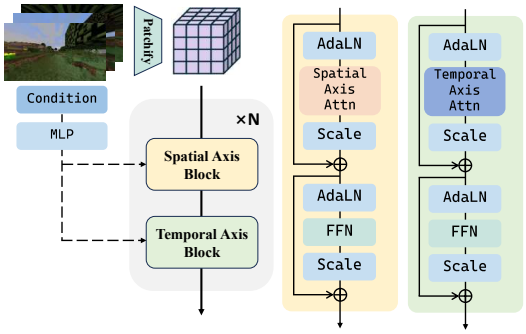


Figure 4: Architecture of DiT model.

5.2 Experimental Results

Table 1: Experimental results of all methods with different checkpoints after 16,000 training steps evaluated on VBench metrics. Our proposed VRR method achieves steady performance improvements throughout training, and consistently outperforms all baseline methods by a considerable margin.

VBench Metric	Aesthetic Quality↑				Imaging Quality↑			
Training Steps	4000	8000	12000	16000	4000	8000	12000	16000
Diffusion Forcing	39.57	37.52	36.82	38.65	46.20	33.78	39.46	44.37
Frame Archor	**45.95**	38.13	36.62	39.15	48.32	33.99	40.15	45.38
SVI	35.62	36.24	36.35	39.17	43.28	44.82	36.53	42.87
VRR-Unif	41.01	46.09	51.44	53.78	38.84	58.77	**70.54**	**72.08**
VRR-Sigreg	44.60	**52.24**	**54.46**	**55.56**	**60.92**	**67.87**	68.87	69.51

We train all compared methods for 16k training steps under identical experimental setups, and evaluate the holistic quality of several 1-minute generated videos via the VBench benchmark. Quantitative results are summarized in Table 1. For Vanilla Diffusion Forcing, the measured metrics corroborate our observations in Section 3.3. Its internal latent representations degrade progressively as training proceeds, which yields a matching declining trend across all VBench scores. Both Aesthetic Quality and Imaging Quality peak at merely 4,000 training steps; further prolonged training only degrades the final generation quality. Except for achieving a relatively high score at the 4000th step, the Frame Anchor method performs comparably to the diffusion forcing baseline with only marginal improvements. Meanwhile, SVI is not tailored for autoregressive diffusion architectures, making it ineffective at mitigating compounding prediction errors during sequential generation. In stark contrast, our representation regularization (VRR) framework achieves outstanding performance. On one hand, our approach avoids performance degradation with extended training iterations, demonstrating strong training robustness. On the other hand, VRR substantially outperforms all baseline methods across all VBench metrics, verifying its superior generation efficiency.

To further demonstrate the superiority of VRR, we present qualitative video visualizations in Figure 5. All methods take a 100-frame video prompt as the initial condition to perform autoregressive

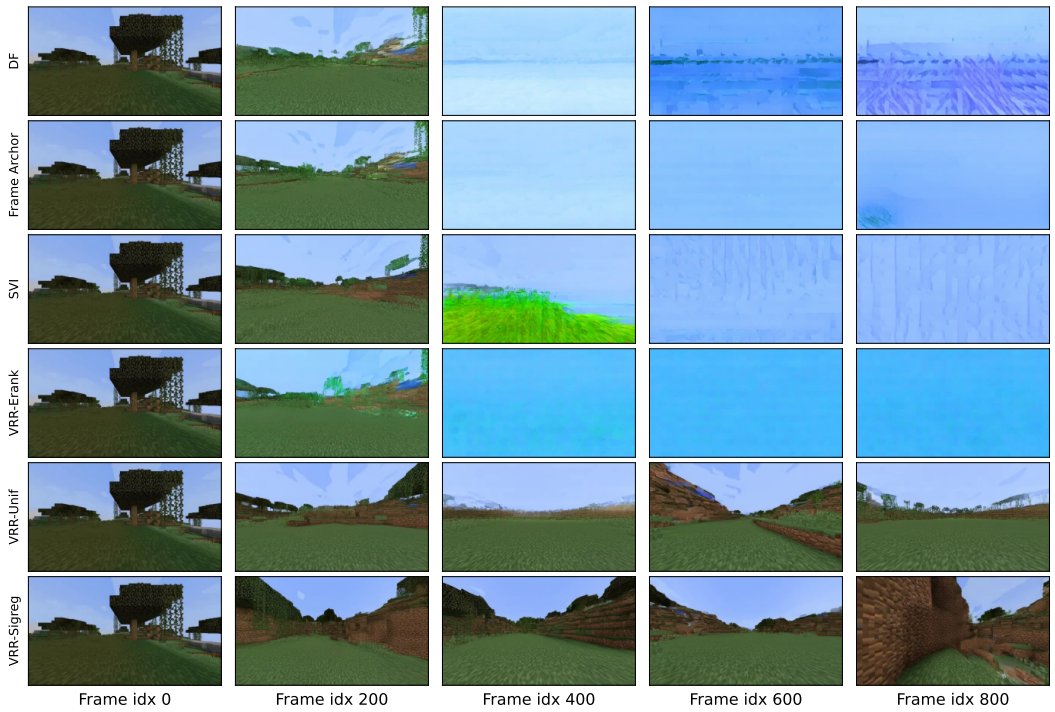


Figure 5: Frames of videos generated from the ckpts of all methods at the 16,000th training step

video extension. We observe that all approaches can generate coherent frames after the first 100 frames. Nevertheless, after an additional 200 frames, Diffusion Forcing, Frame Anchor, SVI and VRR-Erank suffer severely from compounding errors and fail to produce plausible video outputs, with subsequent frames collapsing completely. In contrast, our proposed VRR method, whether regularized via Sigreg or Uniformity loss, maintains stable generation over extremely long contexts and is largely immune to the adverse effects of compounding errors.

5.3 Mitigating Compounding Error

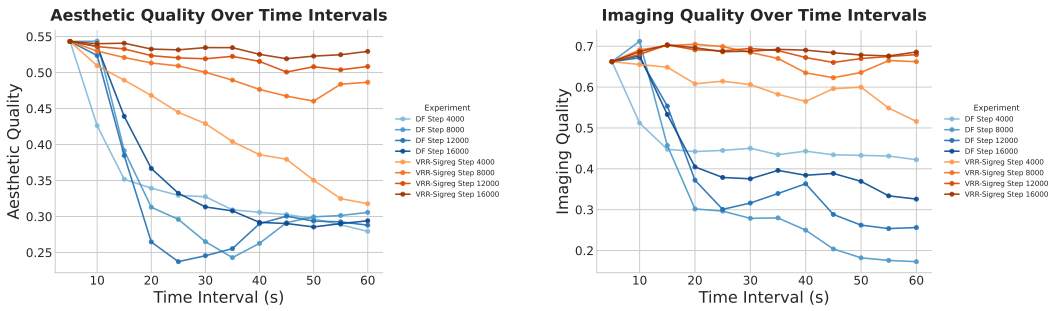


Figure 6: The video quality of vanilla Diffusion Forcing continuously degrades as the video progresses. Our method maintains high scores across VBench metrics.

To evaluate whether our method effectively mitigates the compounding error issue, we split each 1-minute long video into 12 non-overlapping 5-second clips, and sequentially evaluate the video quality of each clip using VBench. The results are presented in Figure 6. We observe that the video quality of vanilla Diffusion Forcing continuously degrades as the video progresses. In contrast, after incorporating our regularization technique and training for 4000 steps, the compounding error of the model is substantially alleviated, and the model maintains high scores across VBench metrics.

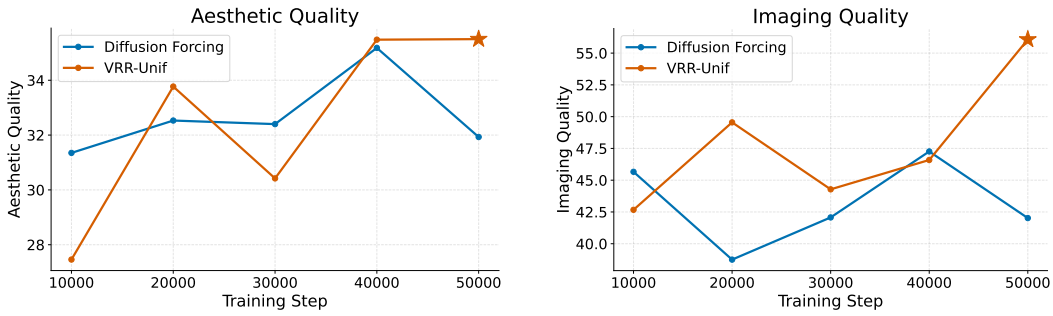
5.4 Ablation Study

Table 2: Experimental Results of Different Regularization Methods

VBench Metric	Aesthetic Quality↑				Imaging Quality↑			
Training Steps	4000	8000	12000	16000	4000	8000	12000	16000
Diffusion Forcing	39.57	37.52	36.82	38.65	46.20	33.78	39.46	44.37
Barlow twins	42.56	52.18	38.67	41.44	40.81	49.75	38.76	36.74
VICReg	34.55	29.37	32.00	41.44	47.93	67.52	46.64	36.74
Erank	**45.90**	35.20	39.88	37.05	53.70	36.24	39.30	37.49
Unif	41.01	46.09	51.44	53.78	38.84	58.77	**70.54**	**72.08**
Sigreg	44.60	**52.24**	**54.46**	**55.56**	**60.92**	**67.87**	68.87	69.51

Comparison of Different Representation Regularization Methods We apply different representation regularization strategies to the hidden states of the diffusion model and compare their effectiveness. Results are shown in Table 2. Both SigReg and Unif achieve comparable performance, which substantially outperforms the baseline. The remaining methods yield mediocre results. The experimental observations of Erank regularization are similar to those of the diffusion forcing baseline. Barlow Twins and VICReg suffer from training instability, which may lead to sharp increases in the diffusion loss.

5.5 Experiments in a Real-World Setting



(a) Aesthetic Quality.

(b) Imaging Quality.

Figure 7: VBench results on RealEstate10K after training a UNet for one epoch.

We train a UNet on the RealEstate10K dataset for one epoch, with the results shown in Figure 7. Our VRR method outperforms the baseline on both the Aesthetic Quality and Imaging Quality dimensions of VBench. These results in a real-world setting further demonstrate the generalizability of our method.

6 Conclusion and Discussion

In this work, we present Video Representation Regularization (VRR), a method that regularizes representations during the training of autoregressive diffusion models to boost long-horizon generation performance and mitigate compounding errors. We observe that error accumulation in autoregressive video generation is typically accompanied by dimensional collapse of model representations. Using effective rank as a metric, we verify a strong correlation between the two phenomena. We further find that simply increasing training data cannot effectively alleviate compounding errors. Building on these observations, our VRR method effectively enhances representation expressiveness and reduces compounding errors, achieving substantially better performance than the baselines across VBench metrics. Despite the demonstrated effectiveness of our approach, the mechanisms underlying representation degradation as the training dataset grows, as well as the specific shortcuts acquired by DiT during training, remain to be elucidated. A systematic investigation of these mechanisms constitutes an important direction for future research.

AI USE STATEMENT

In this work, we did not use generative AI tools for any tasks requiring disclosure. In particular, the research ideas and hypotheses, conceptual and methodological development, experimental design, implementation of the proposed method, data analysis, interpretation of results, and scientific conclusions were produced by the authors. The remaining tasks requiring disclosure, such as generating synthetic datasets, formulating or proving mathematical claims, translation, and qualitative or thematic data analysis, were not applicable to this work. We used generative AI tools only for tasks for which disclosure is recommended: editing the manuscript to improve its readability and assisting with minor edits and refactoring of author-written software code. All suggested textual changes were reviewed and revised by the authors. All AI-assisted code changes were manually inspected and tested by the authors to verify their correctness and ensure that they did not alter the intended methodology or experimental protocol. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human participants or the collection of personally identifiable or sensitive personal information. We use research datasets in accordance with their licenses and intended terms of use. Nevertheless, video world models may reproduce biases present in their training data and may generate inaccurate, misleading, or otherwise harmful visual content. Such models could also be misused to create deceptive media. Moreover, although world models have potential applications in autonomous driving and robotics, the generated predictions should not be treated as reliable descriptions of the physical world or used directly for decision making in safety-critical systems without extensive validation and appropriate safeguards. We encourage responsible use of this research, transparent reporting of model limitations, and the implementation of suitable access controls and safety measures when deploying generative video models.

REPRODUCIBILITY STATEMENT

We support reproducibility by conducting our experiments on publicly available datasets and established evaluation benchmarks. The main paper describes the proposed regularization objectives, model architecture, training setup, evaluation protocol, and ablation studies. Upon publication, we will release the experimental code, including the implementation of the proposed method and the training and evaluation scripts needed to reproduce the reported results. The released repository will also document the required data preparation, dependencies, configurations, and commands for reproducing the experiments.

REFERENCES

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba, Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning, 2025.

Alisson Azzolini, Hannah Brandon, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, Francesco Ferroni, Rama Govindaraju, et al. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025.

Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

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 Gharamani, Raia Hadsell, Aäron van den Oord, Inbar Mosseri, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 3: A new frontier for world models. 2025. URL https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. *arXiv preprint arXiv:2412.03572*, 2024.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video, 2024.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22563–22575, 2023b.

Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. Gamegen-x: Interactive open-world game video generation. *arXiv preprint arXiv:2411.00769*, 2024.

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, 2024a.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7310–7320, 2024b.

Taiye Chen, Xun Hu, Zihan Ding, and Chi Jin. Learning world models for interactive video generation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 154456–154483. Curran Associates, Inc., 2025. doi: 10.52202/085713-5166.

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 preprint arXiv:2510.02283*, 2025.

Decart, Etched, Julian Quevedo, Quinn McIntyre, Spruce Campbell, Xinlei Chen, and Robert Wachen. Oasis: A universe in a transformer. 2024. URL https://oasis-model.github.io/.

Zihan Ding, Chi Jin, Difan Liu, Haitian Zheng, Krishna Kumar Singh, Qiang Zhang, Yan Kang, Zhe Lin, and Yuchen Liu. Dollar: Few-step video generation via distillation and latent reward optimization. *arXiv preprint arXiv:2412.15689*, 2024.

Sergey Edunov, Alexei Baevski, and Michael Auli. Pre-trained language model representations for language generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4052–4059, 2019.

Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7312–7322, 2023. doi: 10.1109/ICCV51070.2023.00675.

Google. Realestate10k. `https://google.github.io/realestate10k/index.html`, 2018. Accessed: 2026-09-26.

Junliang Guo, Yang Ye, Tianyu He, Haoyu Wu, Yushu Jiang, Tim Pearce, and Jiang Bian. Mineworld: a real-time and open-source interactive world model on minecraft. *arXiv preprint arXiv:2504.08388*, 2025.

William H Guss, Brandon Houghton, Nicholay Topin, Phillip Wang, Cayden Codel, Manuela Veloso, and Ruslan Salakhutdinov. Minerl: A large-scale dataset of minecraft demonstrations. *arXiv preprint arXiv:1907.13440*, 2019.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

William Harvey, Søren Nørskov, Niklas Kölch, and George Vogiatzis. Flexible diffusion modeling of long videos. *arXiv preprint arXiv:2205.11495*, 2022.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022.

Yu Hong, Jing Wei, Xing Liu, Xiaodi Wang, Yutong Bai, Haitao Li, Ming Zhang, and Hao Xu. Cogvideo: Large-scale pretraining for text-to-video generation with transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 167283–167308. Curran Associates, Inc., 2025. doi: 10.52202/085713-5576.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818, 2024. doi: 10.1109/CVPR52733.2024.02060.

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 *International Conference on Learning Representations*, volume 2025, pp. 23378–23402, 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.

Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

Tianhong Li, Dina Katabi, and Kaiming He. Return of unconditional generation: A self-supervised representation generation method. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 125441–125468. Curran Associates, Inc., 2024. doi: 10.52202/079017-3985.

Wuyang Li, Wentao Pan, Po-Chien Luan, Yang Gao, and Alexandre Alahi. Stable video infinity: Infinite-length video generation with error recycling. In *International Conference on Learning Representations 2025 (ICLR 2025)*, 2026.

Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Yue Hu, Jingbin Cai, Si Liu, Jianlan Luo, Liliang Chen, Shuicheng Yan, Maoqing Yao, and Guanghui Ren. Genie envisioner: A unified world foundation platform for robotic manipulation. *arXiv preprint arXiv:2508.05635*, 2025.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. Leworld-model: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026.

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 Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, 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 Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical ai, 2025.

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. URL `https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/`.

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.

Ryan Po, Eric Ryan Chan, Changan Chen, and Gordon Wetzstein. Bagger: Backwards aggregation for mitigating drift in autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43727–43739, 2026.

Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. In *International Conference on Learning Representations*, volume 2024, pp. 5260–5274, 2024.

Xuanchi Ren, Yifan Lu, Tianshi Cao, Ruiyuan Gao, Shengyu Huang, Amirmojtaba Sabour, Tianchang Shen, Tobias Pfaff, Jay Zhangjie Wu, Runjian Chen, Seung Wook Kim, Jun Gao, Laura Leal-Taixe, Mike Chen, Sanja Fidler, and Huan Ling. Cosmos-drive-dreams: Scalable synthetic driving data generation with world foundation models, 2025.

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.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pp. 606–610. IEEE, 2007.

David Ruhe, Jonathan Heek, Tim Salimans, and Emiel Hoogeboom. Rolling diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Uriel Singer, Adam Polyak, Eliya Nachmani, Guy Dahan, Eli Shechtman, and Haggai Hacohen. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

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

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*, 2024.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.

Runqian Wang and Kaiming He. Diffuse and disperse: Image generation with representation regularization, 2025.

Wenming Weng, Ruoyu Feng, Yanhui Wang, Qi Dai, Chunyu Wang, Dacheng Yin, Zhiyuan Zhao, Kai Qiu, Jianmin Bao, Yuhui Yuan, et al. Art-v: Auto-regressive text-to-video generation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7395–7405, 2024.

Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. ivideogpt: Interactive videogpts are scalable world models. *Advances in Neural Information Processing Systems*, 37:68082–68119, 2024.

Desai Xie, Zhan Xu, Yicong Hong, Hao Tan, Difan Liu, Feng Liu, Arie Kaufman, and Yang Zhou. Progressive autoregressive video diffusion models. *arXiv preprint arXiv:2410.08151*, 2024.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974, June 2025.

Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactory: Creating new games with generative interactive videos. *arXiv preprint arXiv:2501.08325*, 2025a.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025b.

Lvmin Zhang and Maneesh Agrawala. Packing input frame context in next-frame prediction models for video generation, 2025.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders, 2025.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning, 2025.

A HYPERPARAMETERS

All experiments are conducted on eight NVIDIA H20 GPUs. For the Minecraft experiments, we use a per-GPU batch size of 64, resulting in a global batch size of 512. We set the learning rate to 8×10^{-5} and use a linear warmup over the first 1,000 training steps. The DiT model consists of 16 DiT blocks with a hidden dimension of 1,024 and contains approximately 500 million parameters. The regularization coefficients for SigReg and Unif are set to 5×10^{-4} and 1, respectively.

For the RealEstate10K experiments, we use a per-GPU batch size of 16 and a global batch size of 128. We use the same learning rate of 8×10^{-5} and a linear warmup over the first 1,000 training steps. The regularization coefficients for SigReg and Unif are set to 0.02 and 10, respectively.

B MORE EXPERIMENT RESULTS

B.1 MORE COMPLETE VBENCH RESULTS

Table 3: Complete VBench results for Diffusion Forcing and the two variants of our VRR method. Higher scores are better for all metrics.

Method	Background Consistency↑	Temporal Flickering↑	Motion Smoothness↑	Aesthetic Quality↑	Imaging Quality↑
Diffusion Forcing	87.10	**96.61**	95.06	38.65	44.37
VRR-SigReg	**96.50**	94.43	95.70	**55.56**	69.51
VRR-Unif	96.22	95.19	**96.14**	53.78	**72.08**

Table 3 reports the complete VBench results. The scores for Temporal Flickering and Motion Smoothness are uniformly high across the three methods and therefore provide limited discrimination, so we do not analyze these two metrics further. On the remaining metrics, both VRR variants substantially outperform the Diffusion Forcing baseline, demonstrating clear improvements in background consistency, aesthetic quality, and imaging quality.

B.2 EFFECT OF REGULARIZING HIDDEN STATES AT DIFFERENT LAYERS

Table 4: Experimental results of regularizing hidden states at different layers using the VRR-Sigreg method

VBench Metric	Aesthetic Quality↑				Imaging Quality↑			
Training Steps	4000	8000	12000	16000	4000	8000	12000	16000
Diffusion Forcing	39.57	37.52	36.82	38.65	46.20	33.78	39.46	44.37
Reg at layer 7	31.99	**54.17**	48.60	53.39	24.10	64.88	53.73	66.57
Reg at layer 15	36.56	32.55	40.79	38.78	44.79	46.54	53.98	51.23
Reg at all layers	44.00	48.21	52.06	51.42	39.19	47.05	65.82	62.92
Reg at layer 0, 7, 15	**44.60**	52.24	**54.46**	**55.56**	**60.92**	**67.87**	**68.87**	**69.51**

To investigate which layers of the DiT layer are more critical for mitigating compounding errors, we conducted experiments that apply regularization to the hidden states of different layers. We observed that VRR consistently improves the performance of DiT relative to vanilla diffusion forcing, regardless of the regularization coefficient selected. Nevertheless, applying regularization solely to layers 0, 7, and 15 yields particularly prominent gains. These correspond to the first, middle, and final layers of the DiT model, respectively.