Figure 1: We reimplemented Diffusion Forcing with 3D-Unet and Transformer. We found that diffusion forcing scales up well with modern techniques like image-pretraining and latent diffusion. Figure shows non-cherry-picked samples of diffusion forcing interpolating start and end frames on RE10K, a hard dataset with high resolution of 512.

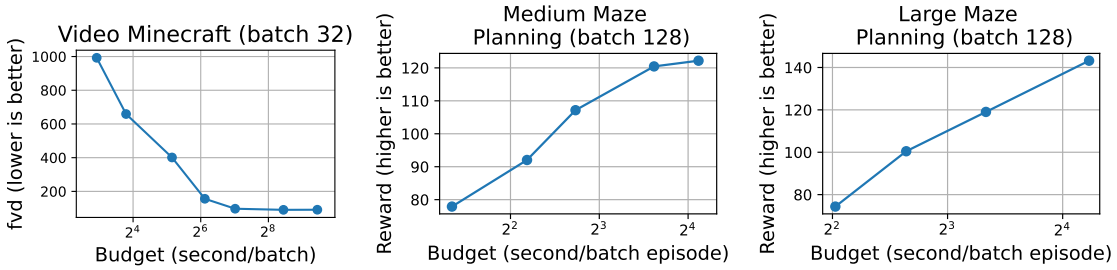
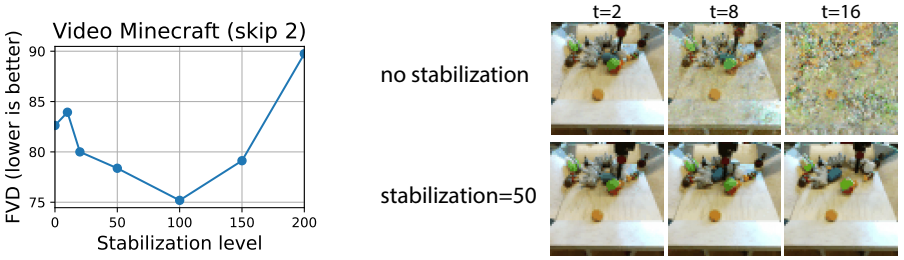


Figure 2: Performance of Diffusion Forcing with respect to compute budget (log scale). For Minecraft video generation, we vary the ddim sampling steps and report the standard video metric, Frechet Video Distance[2], under action conditional setting with 3D-Unet implementation. For maze planning, we vary replan frequency as it's a more dominant factor.

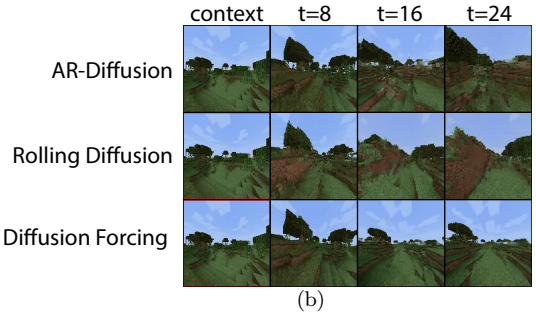


(a) (b)

Figure 3: **(a)** For minecraft, FVD improves as we go from no stabilization to 100, and then degregates. This trend is very obvious when the next token prediction is imperfect, such as when one uses small DDIM steps for faster speed, leading to higher compounding error. **(b)** We found that in highly stochastic datasets like BAIR robot pushing, stabilization is critical to not blowing up even under short horizon. It's also more helpful in latent diffusion since latents are more gaussian.

	FVD ↓
AR-Diffusion (pyramid)	288
Rolling Diffusion (pyramid)	351
Causal Full-sequence (uniform)	337.8
Full-sequence (uniform)	169.6
Teacher Forcing (autoregressive)	170.6
Causal Diffusion Forcing (autoregressive)	**94.8**
Diffusion Forcing (pyramid)	**91.9**

(a)



(b)

Figure 4: **(a)** Additional comparison with AR-Diffusion [3] and Rolling Diffusion [1], a recent, non-casual variant of it. Pyramid sampling across all methods follows the scheme used by AR-Diffusion and is using parameters tuned for AR-Diffusion. With our best effort to improve baselines, diffusion forcing has the best fvd even without the expensive reconstruction guidance. **(b)** Non-cherry-picked qualitative samples visualizing baselines. Diffusion Forcing (trained with independent noise) is best despite using the exact same sampling scheme of AR-Diffusion.

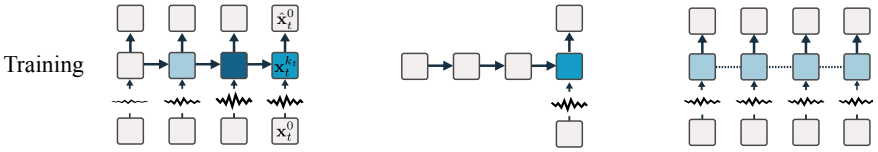


Figure 5: We propose to modify the training part of Figure 2 to this, removing latent circles.

References

[1] David Ruhe et al. "Rolling Diffusion Models". In: *arXiv preprint arXiv:2402.09470* (2024).

[2] Thomas Unterthiner et al. "FVD: A new metric for video generation". In: (2019).

[3] Tong Wu et al. "Ar-diffusion: Auto-regressive diffusion model for text generation". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 39957–39974.