History-Guided Video Diffusion

Motivation

How can we condition the video diffusion on a variable length history?

Guiding Video Diffusion with History

Let $x_{\mathcal{T}}$ be a T-frame video clip, with $\mathcal{T} = \{1, \cdots T\}$. Define $\mathcal{H} \subset \mathcal{T}$ as the indices of the history frames, and put $\mathcal{G} = \mathcal{T} \backslash \mathcal{H}$ as the indices of the frames to be generated. Then guiding generation of $x_{\mathcal{G}}$ with history $x_{\mathcal{H}}$ is modeling the conditional distribution $p(x_{\mathcal{G}} \,|\, x_{\mathcal{H}})$. In theory, this can be done by extending the classifier-free guidance (CFG) with

$$\nabla \log p_k(x^k_{\mathcal{G}}) + \omega \left[\nabla p_k(x^k \mathcal{G} \,|\, x_{\mathcal{H}}) - \nabla \log p_k(x^k_{\mathcal{G}}) \right].$$

However conventional diffusion models have two challenges modeling this distribution:

- Handling variable length conditioning
  - Fixed-length conditioning architectures
  - Independent conditioning architectures, which are inefficient and mostly limited to short history.
  - But technically, text conditioning is not fixed length...
- Framewise binary dropout performs poorly

Preliminary: Diffusion Forcing

Given a sequence of elements $x_{1:T}$, for instance patches of an image or frames of a video, the conventional diffusion model applies *identical* level of noise to all elements. However, Diffusion Forcing (DF) proposes a novel training framework where diffusion model are trained with independent *varied* noise level for each element.

Diffusion Forcing Transformer (DFoT)

If we allow varied noise level per frame, one can view the history frames as the *noise-free* frames. In other words, video frame with noise level k can be expressed as $x_1^{k_1}, \cdots, x_T^{k_T}$, where

$$k_t = \begin{cases} 0 & t \in \mathcal{H} \\ k & t \in \mathcal{G} \end{cases}.$$

Then during training, rather than fixed noise level of history frames to 0, authors follow *per-frame independent noise levels*:

$$\mathbb{E}_{k_\mathcal{T}, x_\mathcal{T}, \epsilon_\mathcal{T}} \left\{ \left\| \epsilon_\mathcal{T} - \epsilon_\theta(x_\mathcal{T}^{k_\mathcal{T}}, k_\mathcal{T}) \right\|^2 \right\},$$

for 1) efficient parallel training, 2) allowing non-causal conditioning on partially masked future frames.

History Guidance

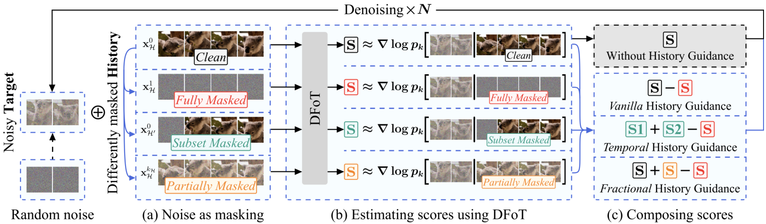

Figure: History guidances

- Vanilla History Guidance (HG-v) : Simple application of CFG
- Temporal History Guidance (HG-t)
  - As the length of history grows, the amount of data that is required to cover different combination of frames exponentially increases, which makes model prone to OOD
  - HG-t composes scores conditioned on different subsequences of history, for instance $\{\mathcal{H}_{\text{long}}, \mathcal{H}_{\text{short}}\}$.
- Fractional History Guidance (HG-f)
  - Major drawback of HG-v is that it generates static video.
  - Rather than setting noise level of the history to 0, HG-f introduce mild noise to history frames for variance in the high-frequency details.

Experiment

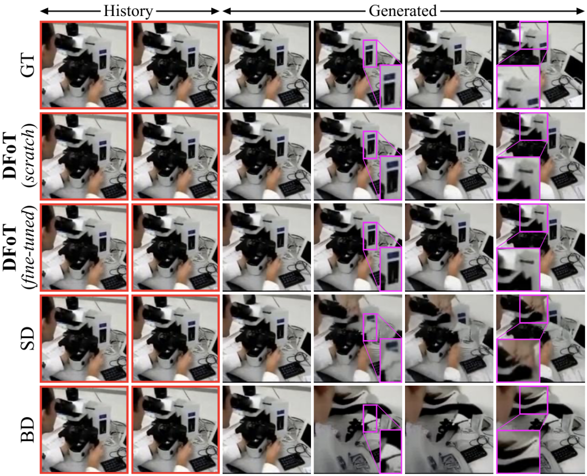

Figure: Consistency with history

Experiment

(a) **Vanilla history guidance significantly improves frame quality and consistency with an increasing guidance scale.** We sample with varying guidance scales $\omega = 1$ (top, *without history guidance*), 1.5 (middle), and 3 (bottom).

(b) **Fractional history guidance resolves the issue of static videos, improving dynamics by guiding with lower frequencies.** We sample with varying frequency scales, with $k_{\mathcal{H}} = 0$ (top, *vanilla guidance leading to static videos*), 0.3 (middle), and 0.6 (bottom).

Experiment

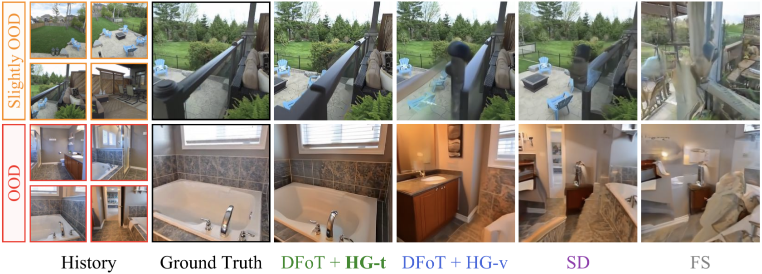

Figure: Robustness to OOD