Identifying and Solving Conditional Image Leakage in Image-to-Video Diffusion Model

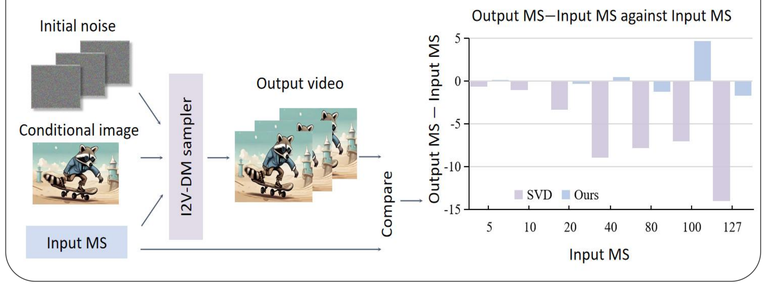
https://github.com/thu-ml/cond-image-leakage NeurIPS 2024

Tsinghua University, Renmin University of China

Min Zhao, Hongzhou Zhu, Chendong Xiang, Kaiwen Zheng,
Chongxuan Li and Jun Zhu

Motivation

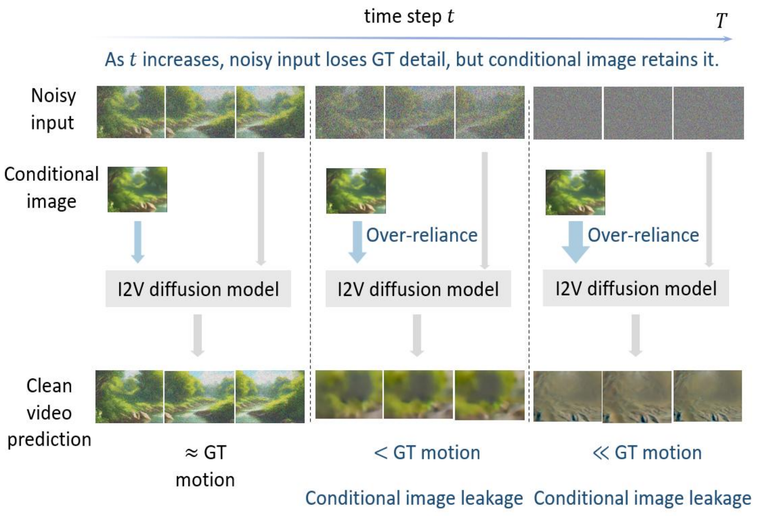
- Observation: In the image-to-video (I2V) generation task, the motion amplitude of the generated video is consistently less than expected, and even remains static.



Cause Analysis: Conditional Image Leakage

- When the noise level is high, the model faces challenges in predicting content from noisy videos, resulting in an over-reliance on the conditional images, which leads to a decrease in the dynamism of the generated videos.

Identifying the Conditional Image Leakage



Methods

- Grounded in the issue of Conditional Image Leakage, we propose both the inference strategy and the training strategy to address this problem.

- Inference strategy : Early sampling with Analytic-Init
  - Sample from an earlier timestep. And how to design the optimal initial distribution?
  - Analytic-Init : Estimate the optimal Gaussian distribution based on the training data to minimize the KL divergence between the initial distribution and the forward distribution.

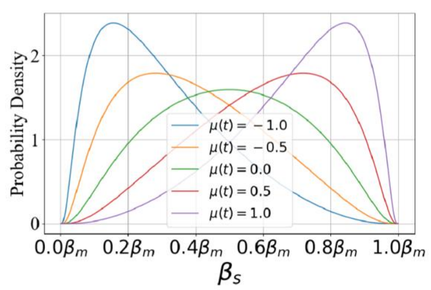
Proposition 1. *Given a normal distribution* $p_M(X_M) = \mathcal{N}(X_M; \boldsymbol{\mu}_p, \sigma_p^2 \boldsymbol{I})$ *and* $q_M(X_M)$ *is the margin distribution of diffusion forward process at time* M*, with the forward trainsition kernel* $q_{M|0}(X_M|X_0) = \mathcal{N}(X_M; \alpha_M X_0, \sigma_M^2 \boldsymbol{I})$*, the minimization problem* $\min_{\boldsymbol{\mu}_p, \sigma_p^2} D_{KL}(q_M(X_M)||p_M(X_M))$ *yields the following optimal solution:*

$$\boldsymbol{\mu}_p^* = \alpha_M \mathbb{E}_{q(X_0)}[X_0], \quad \sigma_p^{2*} = \alpha_M^2 \frac{\sum_{j=1}^{d}[Var(X_0^{(j)})]}{d} + \sigma_M^2, \tag{5}$$

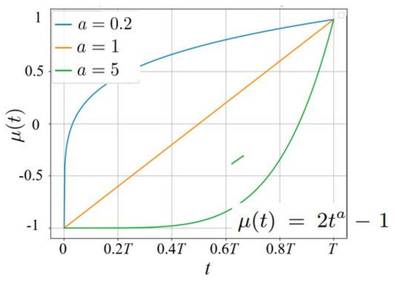
where $q(X_0)$ *denotes the data distribution ,* d *denotes the dimension of the data, and* $X_0^{(j)}$ *denotes the j-th component of* X_0*.*

- Training strategy : TimeNoise
  - To reduce the model's over-reliance on the conditional image, we add noise to the conditional image.
  - The extent of conditional image leakage becomes more pronounced as the timestep increases. → The added noise follows a logit-normal distribution parameterized by the timestep (TimeNoise).

$$p_t(\beta_s; \mu(t), \beta_m) = \frac{\beta_m}{\sqrt{2\pi}} \frac{1}{\beta_s(\beta_m - \beta_s)} e^{-\frac{(\text{logit}(\frac{\beta_s}{\beta_m}) - \mu(t))^2}{2}}$$



(a) TimeNoise $p_t(\beta_s)$ with hyperparameters β_m (maximum noise) and $\mu(t)$ (distribution center).



(b) Distribution center $\mu(t) = 2t^a - 1$ with a controlling monotonic behavior flexibly.

Experiments

Ablation studies of hyperparameters of TimeNoise.

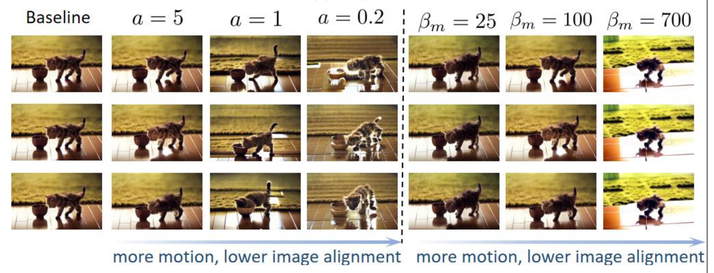


Table 1: Quantitative comparison

Model	FVD↓	IS↑	Motion Score ↑
DymiCrafter [13]	363.8	16.39	50.96
DymiCrafter + Analytic-Init	**316.3**	**17.66**	**71.04**
DymiCrafter-naive-tune	382.5	21.12	31.68
DymiCrafter-naive-tune + Analytic-Init	**342.9**	**22.71**	**50.08**
DymiCrafter-TimeNoise	**334.9**	**21.42**	**72.32**
DymiCrafter-CIL	**332.1**	**22.84**	**73.92**
VideoCrafter1 [12]	353.9	18.75	63.36
VideoCrafter1 + Analytic-Init	**341.6**	**19.86**	**139.04**
VideoCrafter1-naive-tune	460.3	23.98	62.72
VideoCrafter1-naive-tune + Analytic-Init	**450.1**	**24.50**	**65.12**
VideoCrafter1-TimeNoise	**452.2**	**24.62**	**64.80**
VideoCrafter1-CIL	**443.7**	**25.11**	**66.7**
SVD [9]	388.3	36.32	16.64
SVD + Analytic-Init	**382.0**	**36.81**	**19.68**
SVD-naive-tune	311.0	22.03	9.60
SVD-naive-tune + Analytic-Init	**277.1**	**22.18**	**20.64**
SVD-TimeNoise	**272.2**	**23.01**	**20.96**
SVD-CIL	**272.4**	**25.18**	**21.44**

Qualitative results of Analytic-Init and TimeNoise applied to various I2V-DMs