SenCache: Accelerating Diffusion Model Inference via Sensitivity-Aware Caching

Supplementary Material

Insight on CogVid and LTX high ε. We design the following diagnostic. For each of the three models, we generate 100 videos and compute the mean absolute error (MAE) between the denoiser outputs at two consecutive timesteps, i.e., $\|f(x_{t_k}, t_k) - f(x_{t_{k-1}}, t_{k-1})\|_1$. Smaller values indicate that reusing a cached output across nearby steps would introduce less error. The averages over 100 videos are reported in Figure 1. We observe that in the mid-range timesteps (approximately 800–200), where caching is most frequently applied, this consecutive-step MAE is consistently higher for CogVideoX and LTX-Video than for Wan. This suggests these models exhibit larger per-step variation (higher effective sensitivity), so achieving the same NFE reduction requires a larger caching tolerance ε, which inherently permits more approximation error and can lead to the observed quality drop.

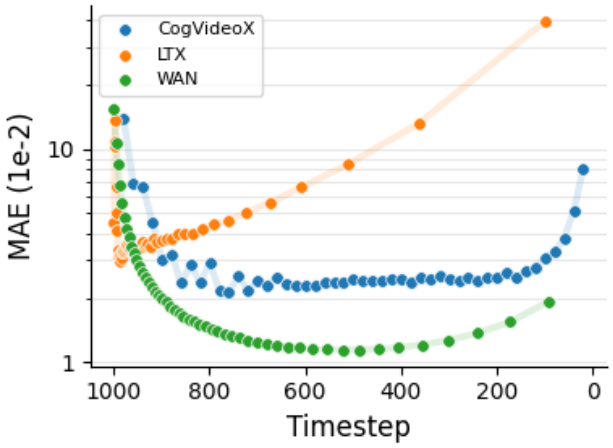

Figure 1. MAE between the denoiser outputs at two consecutive timesteps.

SenCache vs Global Timestep Optimization Methods. Local sensitivity in SenCache is a proxy for the marginal cost of skipping one more step. Global schedule methods can be seen as doing the same thing but with planning: they allocate the error budget across timesteps to avoid cases where many "small" local skips add up. In this view, SenCache is like using a fixed per-step budget (via ε), while global optimization is the more general version that chooses how that budget should vary over time. An interesting future direction is to combine the two: a global scheduler could provide dynamic $\varepsilon(t)$ values that SenCache uses for local decisions.

Additional Efficiency Metrics. On a GH200 GPU for Wan 2.1, our method reduces end-to-end wall-clock latency from $182.3\,\mathrm{s}$ (vanilla) to $107.3\,\mathrm{s}$ (41.1% speedup), compared to MagCache at $110.6\,\mathrm{s}$ (39.3% speedup); both reduce total compute from $8{,}244{,}043.09$ to $3{,}482{,}412.58$ GFLOPs (57.8% fewer).

Cross-Model Sensitivity Patterns. We provide a visualization of the estimated network sensitivity in **??**. Across all models, we first observe that variations in both the time step t and the noisy sample must be taken into account for effective caching, as the networks are sensitive to both. Second, a small batch of 8 diverse samples is already sufficient to obtain reliable sensitivity estimates; large batches are not necessary. Third, the sensitivity patterns differ markedly between models. For Wan 2.1 and LTX, at large timesteps, the model is highly sensitive to variations in t. However, this is not the case for CogVideoX. Moreover, while CogVideoX and LTX exhibit low sensitivity to input variations at small timesteps, wan 2.1 shows the opposite behavior and is highly sensitive in this regime. Finally, for LTX in particular, we observe that it is highly sensitive to both variation in t and the noisy latent at large timesteps, but is less sensitive at smaller time steps.