AdaCache

- □□□□□Adaptive Caching for Faster Video Generation with Diffusion Transformers
- □□□□□https://arxiv.org/abs/2411.02397
- ICCV 2025

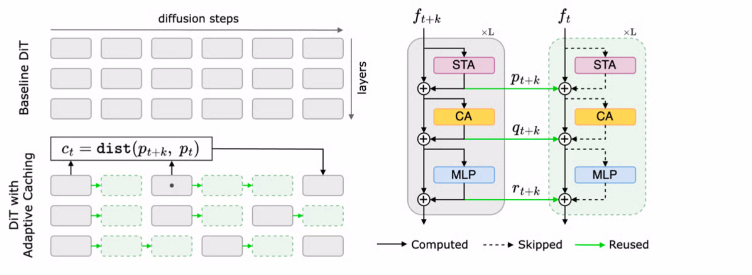

Figure 4 Overview of Adaptive Caching: (Left) During the diffusion process, we choose to cache residual computations within selected DiT blocks. The caching schedule is *content-dependent*, as we decide when to compute the next representation based on a distance metric (c_t). This metric measures the rate-of-change from previously-computed (and, stored) representation to the current one, and can be evaluated per-layer or the DiT as-a-whole. Each computed residual can be cached and reused across multiple steps. (Right) We only cache the residuals (*i.e.*, skip-connections) which amount to the actual computations (*e.g.* spatial-temporal/cross attention, MLP). The iteratively denoised representation (*i.e.*, f_{t+k}, f_t) always gets updated either with computed or cached residuals.

□□cache□□□

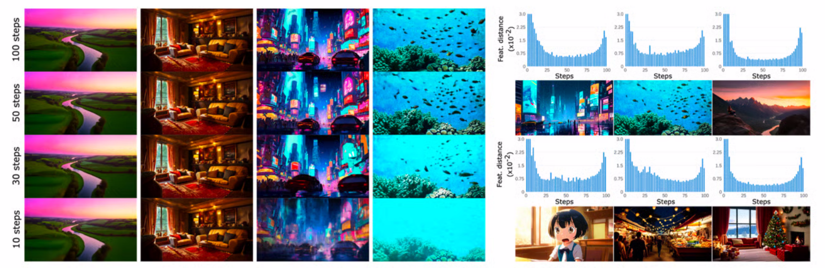

Figure 2 Not all videos are created equal: We show frames from 720p - 2s video generations based on Open-Sora (Zheng et al., 2024). (Left) We try to break each generation by reducing the number of diffusion steps. Interestingly, not all videos have the same break point. Some sequences are extremely robust (*e.g.* first-two columns), while others break easily. (Right) When we plot the difference between computed representations in subsequent diffusion steps, we see unique variations (Feature distance vs. #steps). If we are to reuse similar representations, it needs to be tailored to each video. Both these observations suggest the need for a content-dependent denoising process, which is the founding motivation of Adaptive Caching. Best-viewed with zoom-in. Prompts given in supplementary.

□□□□□□□AdaCache□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□codebook□□□□□□□□□□□□□□□□□□□□□□□□

□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□cache□□□□□□□□□□□□cache□□□

$$m_t^l = \|p_{t,\ i:N}^l - p_{t,\ 0:N-i}^l\| \ .$$

$$mg_t^l = (m_t^l - m_{t+k}^l)\ /\ k\ .$$

$$c_t^l = c_t^l \cdot (m_t^l + mg_t^l)\ .$$

- □□□https://github.com/AdaCache-DiT/AdaCache

Older

2025.11.04

2025-11-04

FasterCache