D2Cache: Second-Order Delta Caching for Higher Video Diffusion Acceleration

Supplementary Material

1. Detailed Proof of Theorem 1

Here we provide the complete proof of Theorem 1 in the main paper. Let $f(t) = \varepsilon_\theta(x_t, t)$. For brevity, we denote $f' = f'(t)$, $f'' = f''(t)$, $f''' = f'''(t)$, and omit the $O((\Delta t)^4)$ remainders where clear. Expanding around t:

$$f(t-1) = f(t) - f' + \tfrac{1}{2}f'' - \tfrac{1}{6}f''', \tag{1}$$

$$f(t+1) = f(t) + f' + \tfrac{1}{2}f'' + \tfrac{1}{6}f''', \tag{2}$$

$$f(t+2) = f(t) + 2f' + 2f'' + \tfrac{4}{3}f'''. \tag{3}$$

The first-order deltas are:

$$\begin{aligned}\delta_1(t) &= f(t) - f(t+1) \\ &= -f' - \tfrac{1}{2}f'' - \tfrac{1}{6}f''', \end{aligned} \tag{4}$$

$$\begin{aligned}\delta_1(t+1) &= f(t+1) - f(t+2) \\ &= -f' - \tfrac{3}{2}f'' - \tfrac{7}{6}f'''. \end{aligned} \tag{5}$$

The second-order delta is:

$$\delta_2(t) = \delta_1(t) - \delta_1(t+1) = f'' + f'''. \tag{6}$$

Substituting Eqs. (4) and (6) into the predictor $\hat{f}^{(2)}(t-1) = f(t) + \delta_1(t) + \delta_2(t)$:

$$\hat{f}^{(2)}(t-1) = f(t) - f' + \tfrac{1}{2}f'' + \tfrac{5}{6}f'''. \tag{7}$$

The error follows from Eqs. (1) and (7):

$$\begin{aligned}e_2(t) &= f(t-1) - \hat{f}^{(2)}(t-1) \\ &= -f''' + O((\Delta t)^4) = O((\Delta t)^3), \end{aligned} \tag{8}$$

improving over the first-order error $e_1(t) = f'' + O((\Delta t)^3) = O((\Delta t)^2)$ by one order. When $f'' \neq 0$, we have $\|e_2(t)\| \leq c\,\Delta t\,\|e_1(t)\|$ for some $c > 0$. Over N cached steps, the cumulative error reduces from $O(N(\Delta t)^2)$ to $O(N(\Delta t)^3)$, explaining D2Cache's growing advantage under higher acceleration.

2. More Results

Qualitative Results on Wan2.1-14B: To validate D2Cache's scalability, we provide qualitative comparisons on **Wan2.1-14B** (14 billion parameters, compared to 1.3B in the main paper). Figure 2 shows generated frames under superfast acceleration (3.61×). TeaCache exhibits visible degradation including blurred textures and motion artifacts, while D2Cache maintains sharp details close to the default generation. This confirms that D2Cache's second-order correction generalizes effectively to larger-scale models.

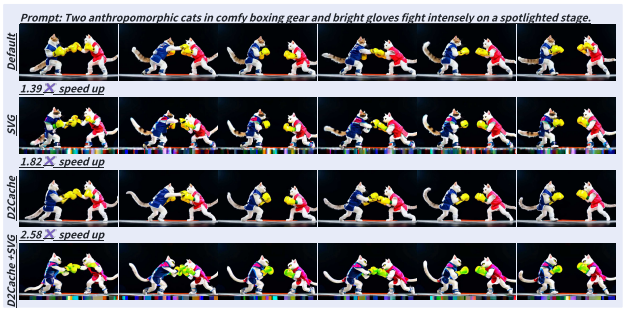

Figure 1. Qualitative results for D2Cache combined with Sparse VideoGen [1]. SVG (1.39×) + D2Cache-slow (1.82×) yields a combined 2.58× speedup.

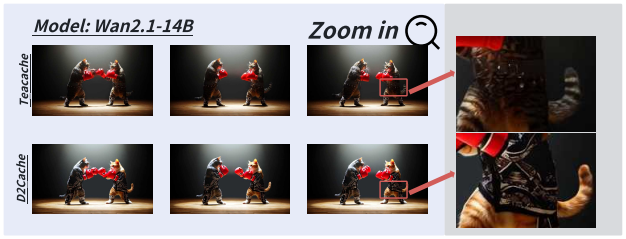

Figure 2. Qualitative comparisons using **Wan2.1-14B** under **superfast** acceleration (3.61×).

Compatibility with Sparse VideoGen: D2Cache is orthogonal to other training-free acceleration methods as they target different computational bottlenecks. We combine *Sparse VideoGen* [1] (SVG, 1.39×) with *D2Cache-slow* (1.82×) on Open-Sora 1.2 using an A6000 GPU, achieving a combined **2.58**× speedup. Figure 1 shows qualitative results, confirming that the combined approach preserves visual quality while achieving near-multiplicative acceleration.

References

[1] Haocheng Xi, Shuo Yang, Yilong Zhao, Chenfeng Xu, Muyang Li, Xiuyu Li, Yujun Lin, Han Cai, Jintao Zhang, Dacheng Li, et al. Sparse videogen: Accelerating video diffusion transformers with spatial-temporal sparsity. *arXiv preprint arXiv:2502.01776*, 2025. 1