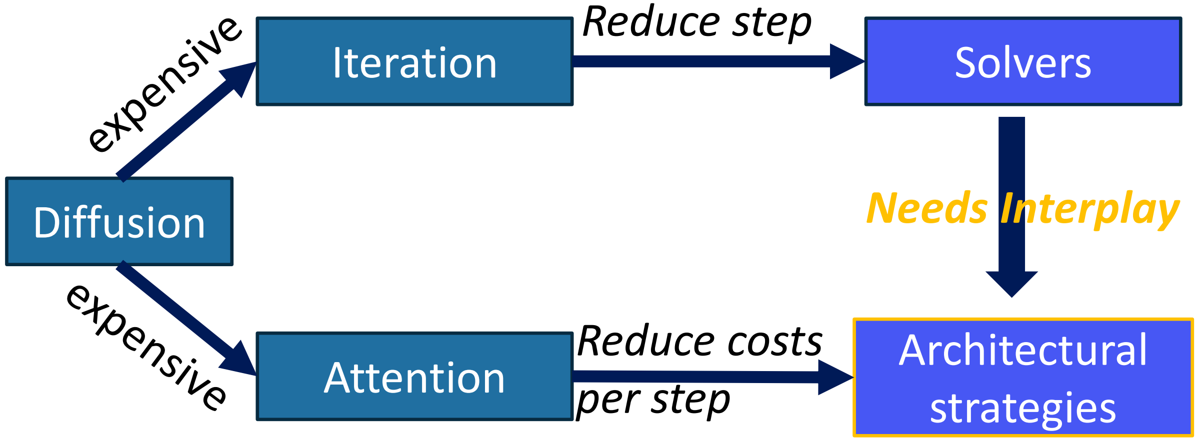
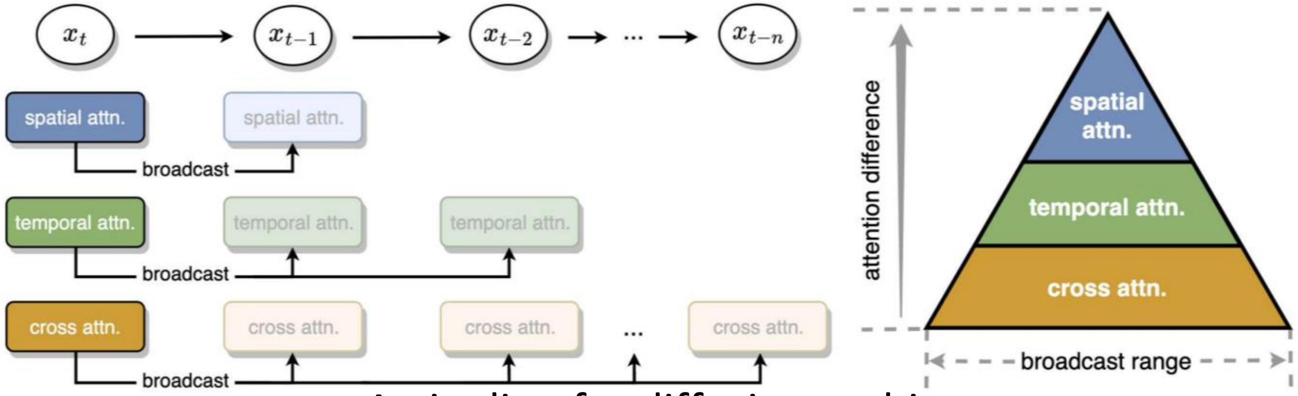
Training-free Inference Acceleration of Generative Models

Duke

Training-free acceleration of Diffusion Models

- ***While these architectural strategies effectively reduce sampling latency, they exhibit lower faithfulness to the original samples :***
  - (a) **Fixed acceleration patterns** cannot be effectively adapted to the **variability** of each prompt's denoising trajectory.
  - (b) Such methods do not explicitly leverage the **underlying ODE formulation** of the denoising process, nor interplay with the specific ODE-solver used

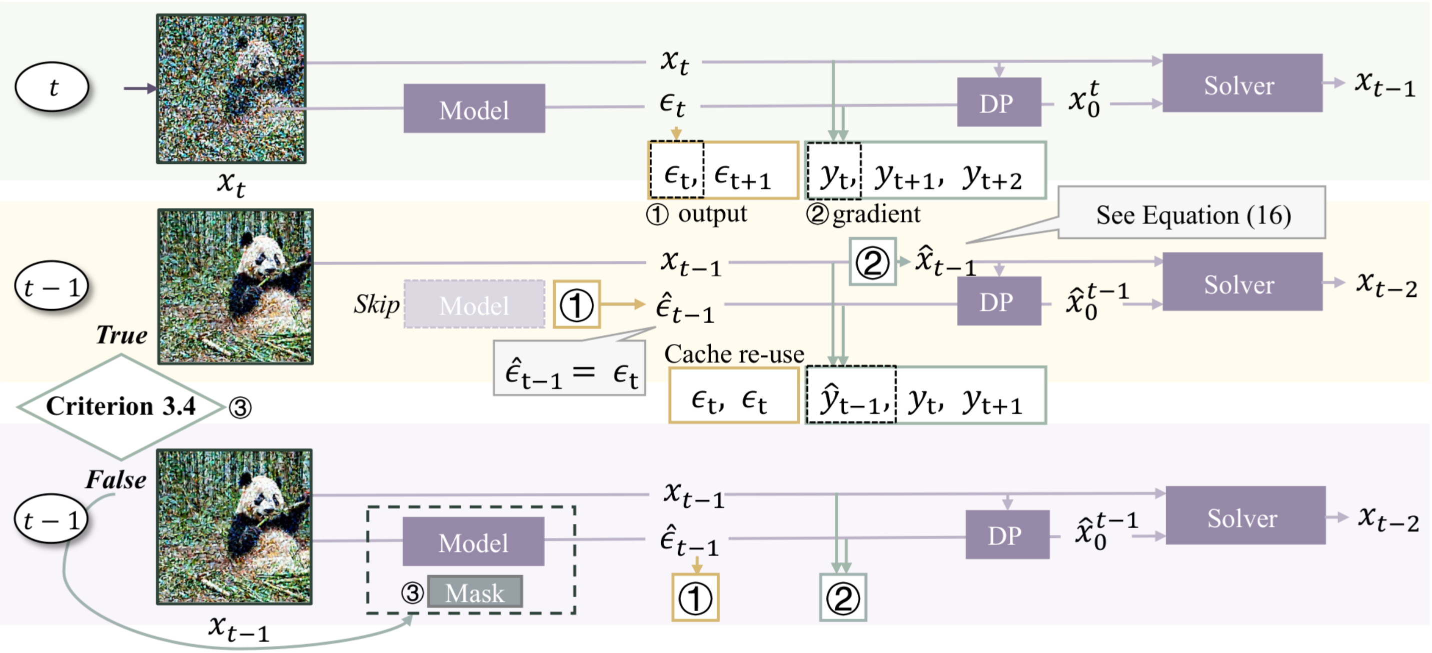


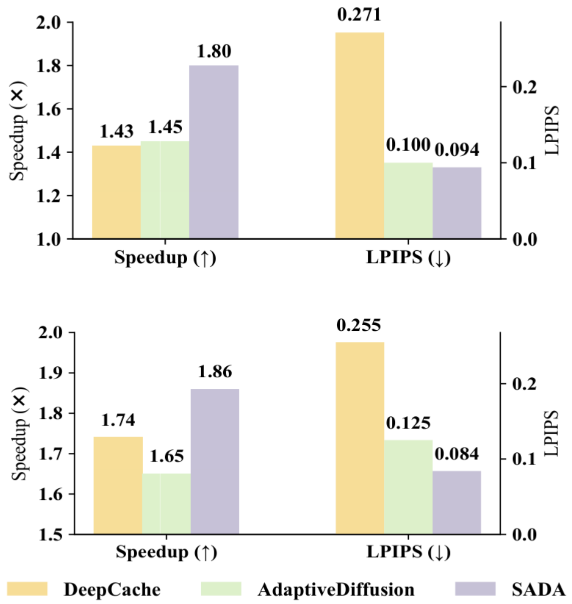


A pipeline for diffusion caching

Duke

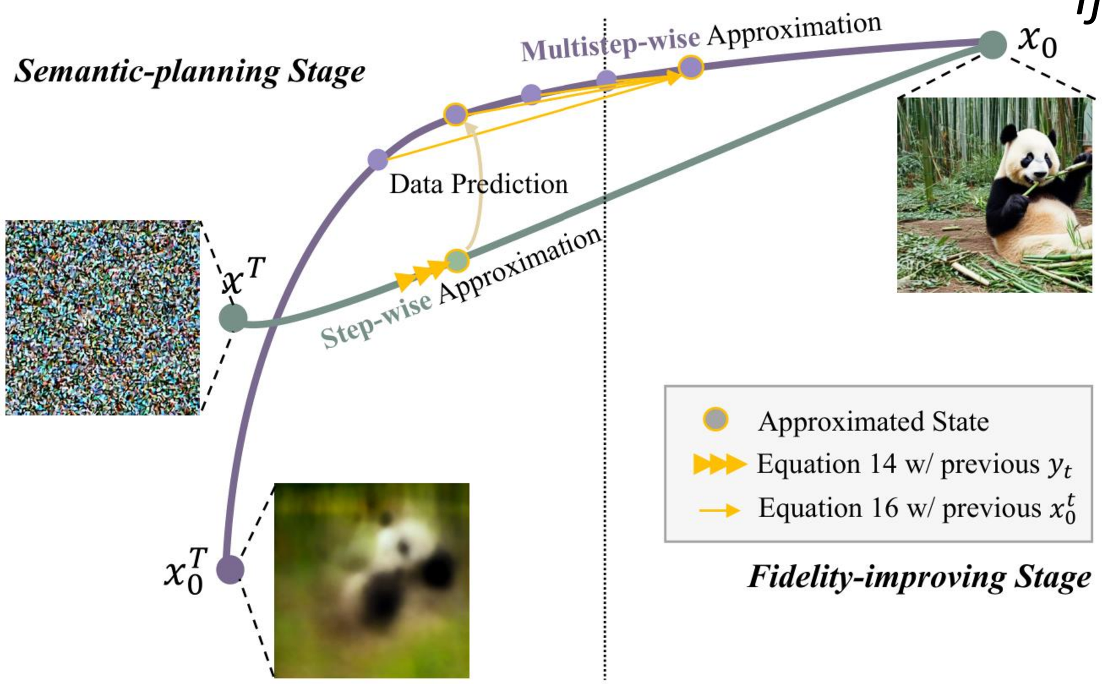
Adaptive Exploit Sparsity with Unified Stability Measure





- SADA: Adaptive acceleration based on a stability criterion measured via the velocity of sampling trajectory $y = \frac{dx_t}{dt}$. The **criterion** is derived as: $(x_{t-1} - \hat{x}_{t-1}) \cdot \Delta^{(2)} y_t < 0.$
  - (a) If < 0: returns **True**, conduct **Step-wise Cache-Assisted Pruning**
  - (b) If > 0: returns **False**, conduct **Token-wise Cache-Assisted Pruning**

Adaptive Sampling with Principled Approximation



*If the criterion returns **True***:

a. In the **first half** of diffusion process, **Step-wise**:

$$\hat{x}_{t-1} := x_t - \frac{5\Delta t}{6} y_t - \frac{5\Delta t}{6} y_{t+1} + \frac{2\Delta t}{3} y_{t+2}$$

We leverage the *precise gradients* calculated by the ODE Solvers

b. In the **second half** of diffusion process, **Multistep**:

$$\hat{x}_0^t := \sum_{i \in I} \left(\prod_{j \in I \setminus \{i\}} \frac{t - t_j}{t_i - t_j} \right) x_0^{t_i}.$$

*If returns **False***: c. the criterion becomes a binary mask that **prunes out "stable tokens"** and **compute the rest**

Note: **Both approximation schemes** eventually yields a **per-step clean image**, x_0^t , which error is bounded. See detailed proof in our original paper.

Duke

Empirical results

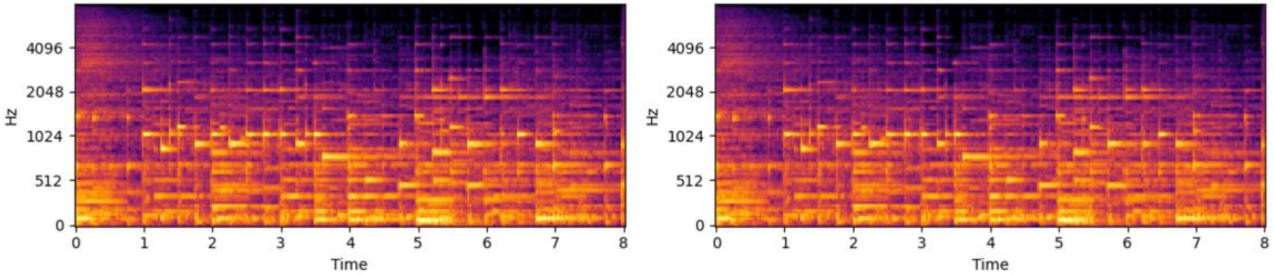
Table 1. Quantitative results on MS-COCO 2017 (Lin et al., 2014).

Model	Scheduler	Methods	PSNR ↑	LPIPS ↓	FID ↓	Speedup Ratio
SD-2	DPM++	DeepCache	17.70	0.271	7.83	1.43×
		AdaptiveDiffusion	24.30	0.100	4.35	1.45×
		SADA	**26.34**	**0.094**	**4.02**	**1.80**×
	Euler	DeepCache	18.90	0.239	7.40	1.45×
		AdaptiveDiffusion	21.90	0.173	7.58	**1.89**×
		SADA	**26.25**	**0.100**	**4.26**	1.81×
SDXL	DPM++	DeepCache	21.30	0.255	8.48	1.74×
		AdaptiveDiffusion	26.10	0.125	4.59	1.65×
		SADA	**29.36**	**0.084**	**3.51**	**1.86**×
	Euler	DeepCache	22.00	0.223	7.36	**2.16**×
		AdaptiveDiffusion	24.33	0.168	6.11	2.01×
		SADA	**28.97**	**0.093**	**3.76**	1.85×
Flux	Flow-matching	TeaCache	19.14	0.216	4.89	2.00×
		SADA	**29.44**	**0.060**	**1.95**	**2.02**×

Duke

Empirical results

"A fantastic piece of music with the deep sound of overlapping pianos" LPIPS = 0.017



Baseline: 2.57s **SADA**: 1.42s

"A deer." "Snowy forest."

Base Image **ControlNet Baseline** (4.76 s) **SADA** (3.36 s)