HarmoniCa: Harmonizing Training and Inference for Better Feature Caching in Diffusion Transformer Acceleration

Yushi Huang*, Zining Wang*, Ruihao Gong, Jing Liu, Xinjie Zhang, Jinyang Guo, Xianglong Liu, Jun Zhang

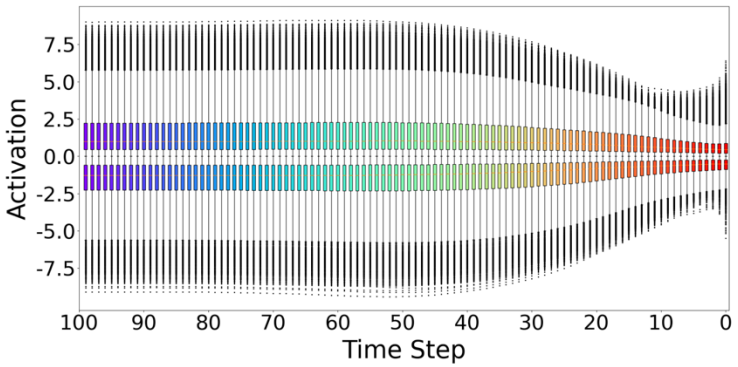
香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Background: Feature Caching for DiT

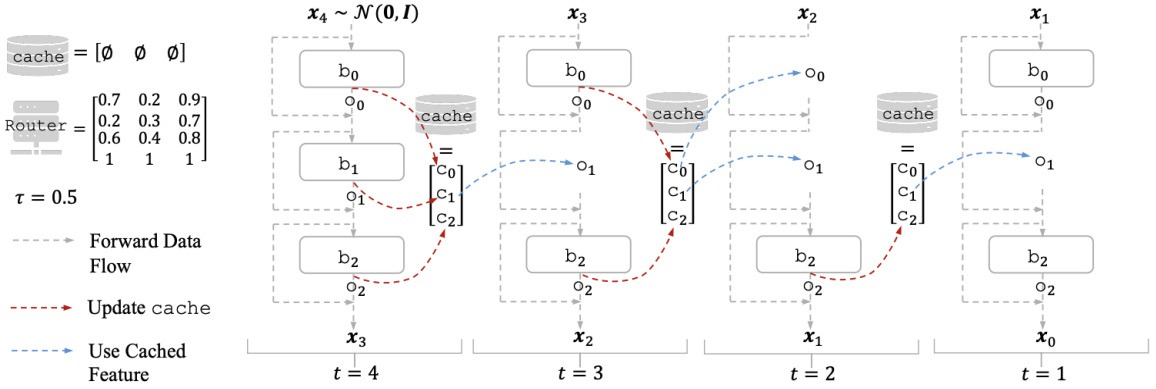
- High Inference Cost for Deployment:
  - Extensive parameter size
  - Multi-round denoising nature
- Feature Caching for Reduce Computation:
  - w/o impacting performance
  - Accelerate Inference

Pixel: 2K
Model: PixArt-Σ[1]
#Param: 0.6B
#Step: 20
Time (s): **14s**
Dev: H800 GPU

Minimal Visual Difference



High similarity between activation



Caching features at current timestep for reuse in subsequent timesteps

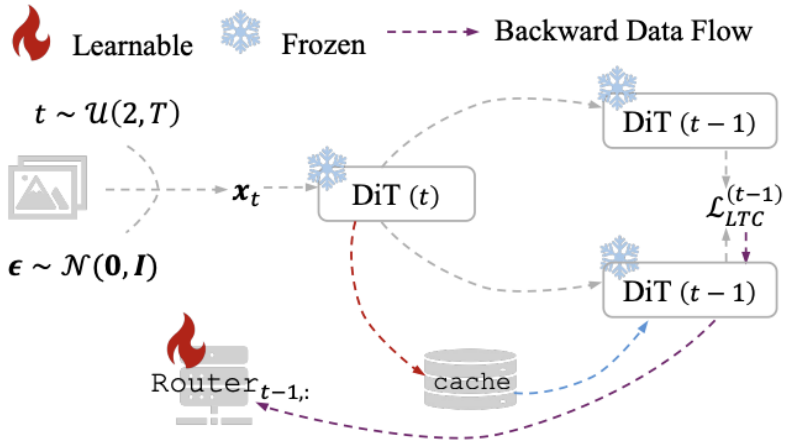
Time (s): **8s**
Real: **1.73×**
Theoretical: **> 2.2×**

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

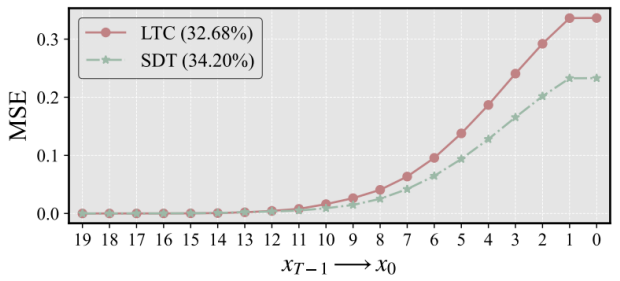
[1]Chen J, et al. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 74-91.

Motivation:

- The learning-based caching strategy is more adaptive. However, the existing work[2] faces the following discrepancies between training and inference:
  - Regard prior timestep:
    - Error on x_t;
    - Shaped context of cache.
  - Objective mismatch:
    - Align predicted noise *vs.* Generate high-quality image.



Uniform sample t and ignore the impact of caching in previous steps



Significant ***accumulated error*** due to regard prior steps

(a) DiT-XL/2

(b) SDT+$\mathcal{L}_{LTC}^{(t)}$ (1.40×)

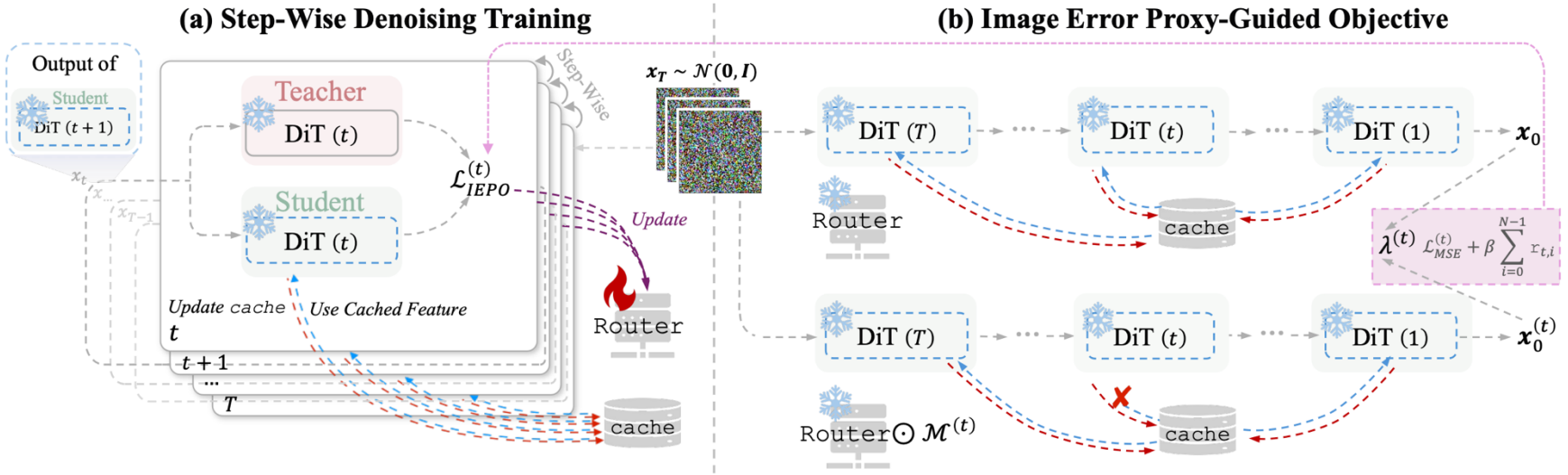
Optimization Shift and from misaligned objective

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

[2]Ma X, et al. Learning-to-Cache: Accelerating Diffusion Transformer via Layer Caching[J]. arXiv preprint arXiv:2406.01733, 2024.

Proposed Framework

- Step-Wise Training:
  - Denoising ***step-by-step*** during training to consider the prior impact.
- Image Error Proxy-Guided Objective:
  - Directly employing image error incurs significant cost (**5×Time, 10×Mem**)
  - Design ***an efficient proxy*** $\lambda^{(t)}$ to represent image error caused by reuse cached feature at t. Adds trade-off between the image quality and cache strategy.

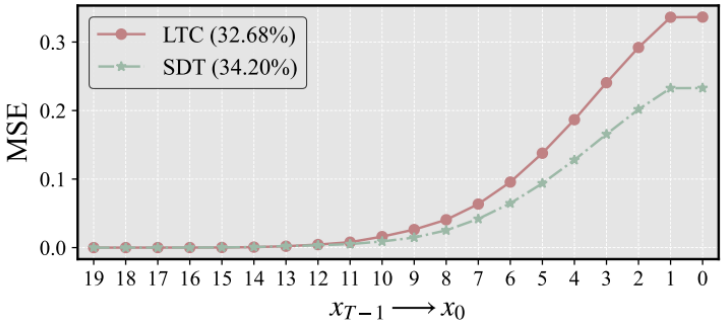


1. $\lambda^{(t)}$ is updated through ***gradient-free*** generation passes every C training iterations.

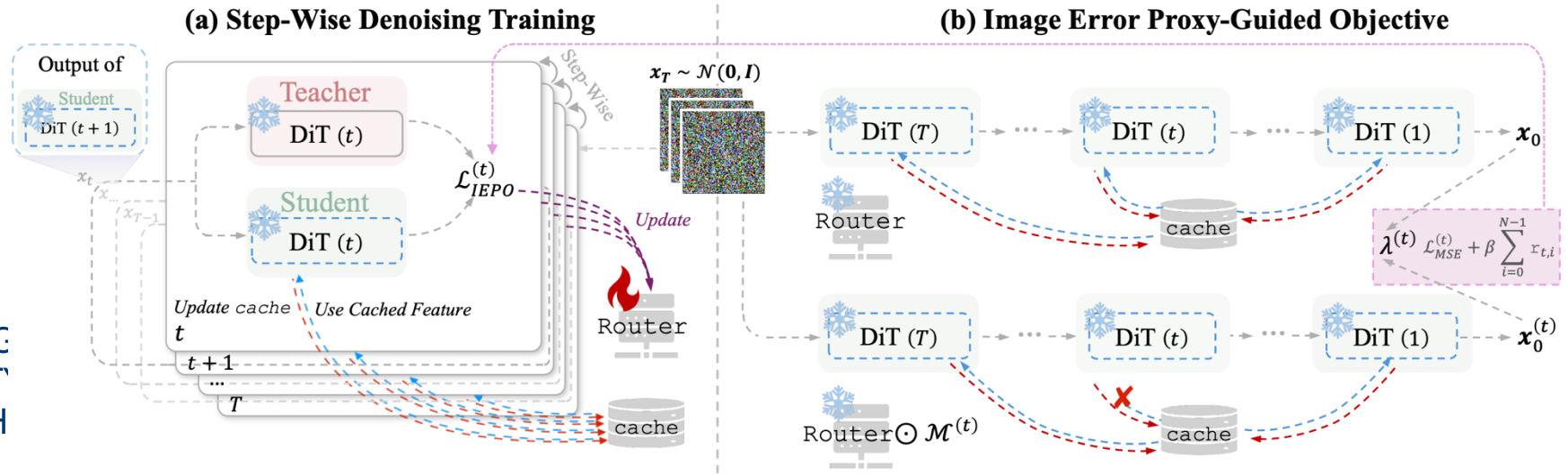
2. $\mathcal{M}^{(t)}$ is to disable the impact of the caching mechanism at t.

Proposed Framework

- Step-Wise Training:
  - Less ***accumulated error***;
  - More accurate x_0.
- Image Error Proxy-Guided Objective:
  - Accurate ***objective-level traits***.



(a) DiT-XL/2 (b) SDT+$\mathcal{L}_{LTC}^{(t)}$ (1.40×) (c) HarmoniCa (1.44×)



Efficiency Discussion

- Training Efficiency:
  - ***Image-free*** during training due to step-wise denoising from a random noise.
  - w/o pre-filling every training iteration (**1.25×speedup**).
- Inference Efficiency:
  - Less than 6% memory overhead by cache
  - **2.07× speedup** (theoretical) & **1.69× speedup** (real-time) with improved performance for non-accelerated PixArt-α[3]

Method	#Images	Time(h)	Memory(GB/GPU)
Learning-to-Cache	1.22M	2.15	33.33
SDT+$\mathcal{L}_{LTC}^{(t)}$	0	1.47	33.28
HarmoniCa	0	1.63	33.28

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

[3]Chen J, et al. Pixart-$\alpha $: Fast training of diffusion transformer for photorealistic text-to-image synthesis[J]. arXiv preprint arXiv:2310.00426, 2023.

Experiments

◆ Class-conditional Generation:

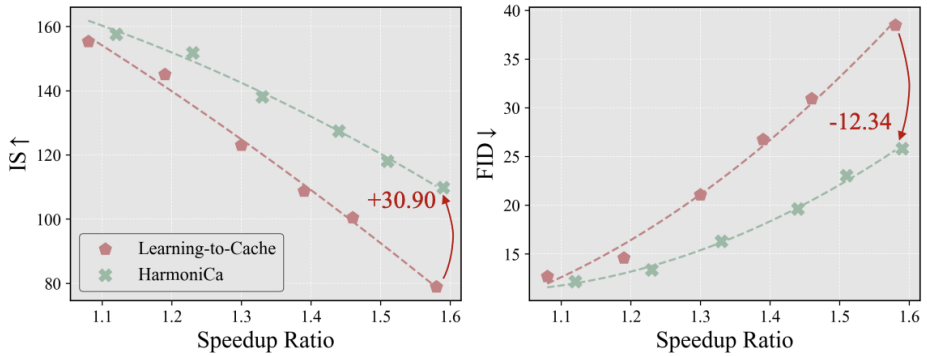
Method	T	IS↑	FID↓	sFID↓	Prec.↑	Recall↑	CUR(%)↑	Latency(s)↓
DiT-XL/2 256 × 256 (cfg = 1.5)								
DDIM (Song et al., 2020a)	50	240.37	2.27	4.25	80.25	59.77	-	1.767
DDIM (Song et al., 2020a)	39	237.84	2.37	4.32	80.22	59.31	-	$1.379_{(1.28\times)}$
Learning-to-Cache (Ma et al., 2024a)	50	233.26	2.62	4.50	79.40	59.15	23.39	$1.419_{(1.25\times)}$
HarmoniCa	50	**238.74**	**2.36**	**4.24**	**80.57**	**59.68**	**23.68**	**1.361**$_{(1.30\times)}$
DDIM (Song et al., 2020a)	20	224.37	3.52	4.96	78.47	58.33	-	0.658
DDIM (Song et al., 2020a)	14	201.83	5.77	6.61	75.14	55.08	-	$0.466_{(1.41\times)}$
Learning-to-Cache (Ma et al., 2024a)	20	201.37	5.34	6.36	75.04	56.09	35.60	$0.468_{(1.41\times)}$
HarmoniCa	20	**206.57**	**4.88**	**5.91**	**75.20**	**58.74**	**37.50**	**0.456**$_{(1.44\times)}$
DDIM (Song et al., 2020a)	10	159.93	12.16	11.31	67.10	52.27	-	0.332
DDIM (Song et al., 2020a)	9	140.37	16.54	14.44	62.63	50.08	-	$0.299_{(1.11\times)}$
Learning-to-Cache (Ma et al., 2024a)	10	145.09	14.59	11.58	64.03	52.06	19.11	$0.279_{(1.19\times)}$
HarmoniCa	10	**151.83**	**13.35**	**11.13**	**65.22**	**52.18**	**22.86**	**0.270**$_{(1.23\times)}$
DiT-XL/2 512 × 512 (cfg = 1.5)								
DDIM (Song et al., 2020a)	20	184.47	5.10	5.79	81.77	54.50	-	3.356
DDIM (Song et al., 2020a)	16	173.31	6.47	6.67	81.10	51.30	-	$2.688_{(1.25\times)}$
Learning-to-Cache (Ma et al., 2024a)	20	178.11	6.24	7.01	81.21	53.30	23.57	$2.633_{(1.28\times)}$
HarmoniCa	20	**179.84**	**5.72**	**6.61**	**81.33**	**55.80**	**25.98**	**2.574**$_{(1.30\times)}$

◆ Text-to-Image Generation:

Method	T	CLIP↑	FID↓	sFID↓	CUR(%)↑	Latency(s)↓
PixArt-α 256 × 256 (cfg = 4.5)						
DPM-Solver++ (Lu et al., 2022b)	20	30.96	27.68	36.39	-	0.553
DPM-Solver++ (Lu et al., 2022b)	15	30.77	31.68	38.92	-	$0.418_{(1.32\times)}$
FORA (Selvaraju et al., 2024)	20	31.10	27.42	37.98	50.00	$0.364_{(1.52\times)}$
HarmoniCa	20	**31.13**	**26.33**	**37.85**	**56.01**	**0.346**$_{(1.60\times)}$
IDDPM (Nichol & Dhariwal, 2021)	100	31.25	24.15	33.65	-	2.572
IDDPM (Nichol & Dhariwal, 2021)	75	31.25	24.17	33.73	-	$1.868_{(1.37\times)}$
FORA (Selvaraju et al., 2024)	100	31.25	25.16	33.62	50.00	$1.558_{(1.65\times)}$
HarmoniCa	100	**31.17**	**23.73**	**32.23**	**53.24**	**1.523**$_{(1.69\times)}$
SA-Solver (Xue et al., 2024)	25	31.31	26.78	38.35	-	0.891
SA-Solver (Xue et al., 2024)	20	31.23	27.45	39.01	-	$0.665_{(1.34\times)}$
HarmoniCa	25	**31.27**	**27.07**	**38.62**	**54.19**	**0.561**$_{(1.59\times)}$
PixArt-α 512 × 512 (cfg = 4.5)						
DPM-Solver++ (Lu et al., 2022b)	20	31.30	23.96	40.34	-	1.759
DPM-Solver++ (Lu et al., 2022b)	15	**31.29**	25.12	40.37	-	$1.291_{(1.36\times)}$
HarmoniCa	20	**31.29**	**24.81**	**40.18**	**54.64**	**1.072**$_{(1.64\times)}$
SA-Solver (Xue et al., 2024)	25	31.23	25.43	39.84	-	2.263
SA-Solver (Xue et al., 2024)	20	31.19	25.85	40.08	-	$1.738_{(1.30\times)}$
HarmoniCa	25	**31.20**	**25.74**	**39.99**	**54.24**	**1.406**$_{(1.61\times)}$
PixArt-α 1024 × 1024 (cfg = 4.5)						
DPM-Solver++ (Lu et al., 2022b)	20	31.10	25.01	37.80	-	9.470
DPM-Solver++ (Lu et al., 2022b)	15	31.07	25.77	42.50	-	$7.141_{(1.32\times)}$
HarmoniCa	20	**31.09**	**23.02**	**36.24**	**55.06**	**5.786**$_{(1.63\times)}$
SA-Solver (Xue et al., 2024)	25	31.05	23.65	38.12	-	11.931
SA-Solver (Xue et al., 2024)	20	31.02	23.88	39.41	-	$9.209_{(1.30\times)}$
Harmonica	25	**31.07**	**23.77**	**38.93**	**53.98**	**7.551**$_{(1.58\times)}$

Experiments

- Comparison w/ speedup increase:



- Comparison w/ additional feature caching:

Constrained by ***U-shape*** structure

Method	T	FID↓	Latency(s)↓
DPM-Solver (Lu et al., 2022a)	20	2.57	7.60
Faster Diffusion (Li et al., 2023a)	20	2.82	$5.95_{(1.28\times)}$
DeepCache (Ma et al., 2024b)	20	2.70	$4.68_{(1.62\times)}$
HarmoniCa	20	**2.61**	$\mathbf{4.60}_{(\mathbf{1.65\times})}$

- Compare w/ pruning & quantization:

Method	T	IS↑	FID↓	sFID↓	Latency(s)↓	Latency(s)↓*
DDIM (Zhang et al., 2022)	20	224.37	3.52	4.96	0.658	1.217
EfficientDM (He et al., 2024)	20	172.70	6.10	**4.55**	$0.591_{(1.11\times)}$	$0.842_{(1.45\times)}$
PTQ4DiT (Wu et al., 2024)	20	17.06	71.82	23.16	$0.577_{(1.14\times)}$	$0.839_{(1.45\times)}$
Diff-Pruning (Fang et al., 2023)	20	168.10	8.22	6.20	$0.458_{(1.44\times)}$	$\mathbf{0.813}_{(\mathbf{1.50\times})}$
HarmoniCa	20	**206.57**	**4.88**	5.91	$\mathbf{0.456}_{(\mathbf{1.44\times})}$	$0.815_{(1.49\times)}$

* denotes the latency was tested on A100; others were on H800.

- Combination with quantization:

Method	IS↑/CLIP↑	FID↓	sFID↓	CUR(%)↑	Latency(s)↓	#Size(GB)↓
DiT-XL/2 256 × 256 (cfg = 1.5)						
EfficientDM (He et al., 2024)	172.70	6.10	4.55	-	$0.591_{(1.11\times)}$	$0.64_{(3.93\times)}$
w/ HarmoniCa ($\beta = 4e^{-8}$)	168.16	6.48	4.32	26.25	$0.473_{(1.40\times)}$	$0.64_{(3.93\times)}$
PixArt-α 256 × 256 (cfg = 4.5)						
EfficientDM (He et al., 2024)	30.09	34.84	30.34	-	$0.469_{(1.18\times)}$	$0.59_{(1.98\times)}$
w/ HarmoniCa	30.15	34.96	30.55	53.34	$0.299_{(1.85\times)}$	$0.59_{(1.98\times)}$
PixArt-α 512 × 512 (cfg = 4.5)						
EfficientDM (He et al., 2024)	30.71	25.82	41.64	-	$0.461_{(1.20\times)}$	$0.59_{(1.98\times)}$
w/ HarmoniCa	30.75	26.15	41.99	53.11	$0.281_{(1.97\times)}$	$0.59_{(1.98\times)}$

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

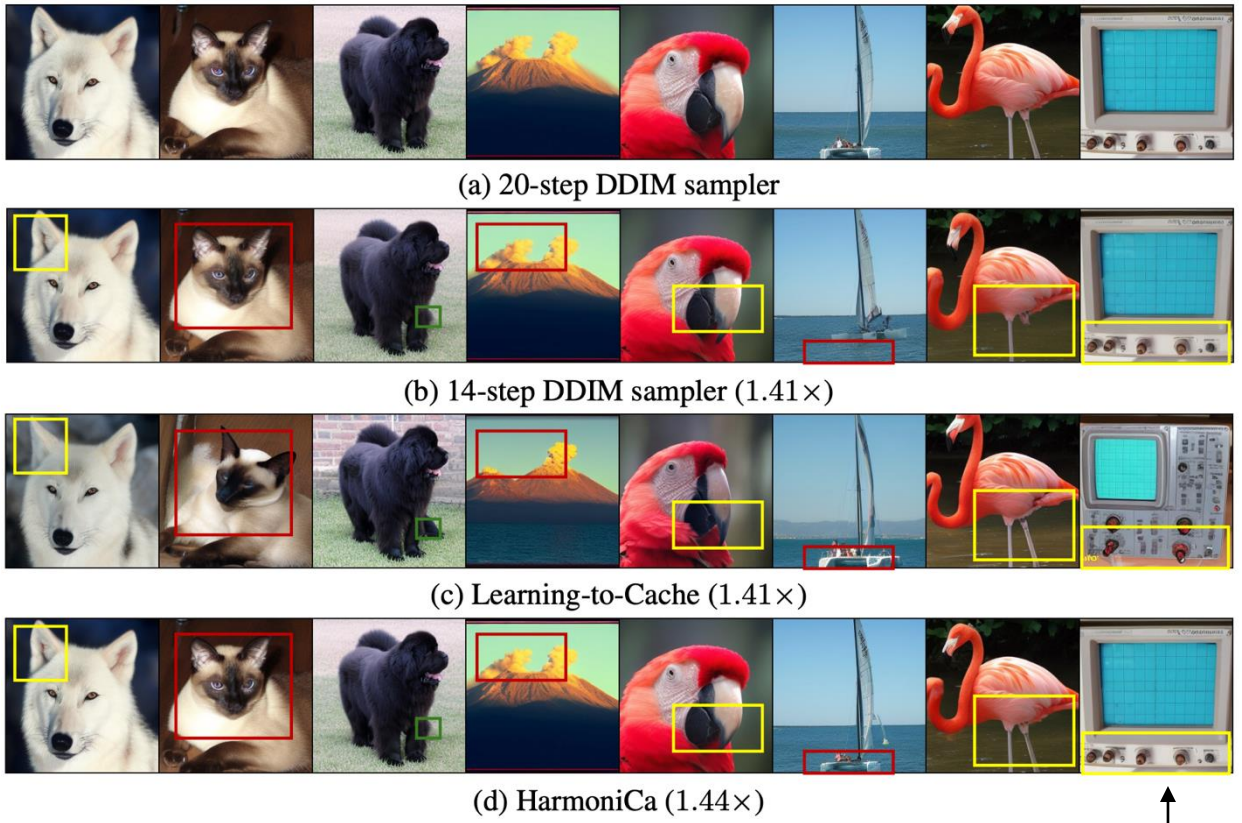
Visualization

(a) 20-step DPM-Solver

(b) 15-step DPM-Solver (1.36×)

(c) FORA (1.53×)

(d) HarmoniCa (1.64×)



(a) 20-step DDIM sampler

(b) 14-step DDIM sampler (1.41×)

(c) Learning-to-Cache (1.41×)

(d) HarmoniCa (1.44×)

All visulization results validate our **superior performance** with **much higher speedup ratio**.

PixArt-α 512×512

DiT-XL/2 256×256

Thank you!

香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY