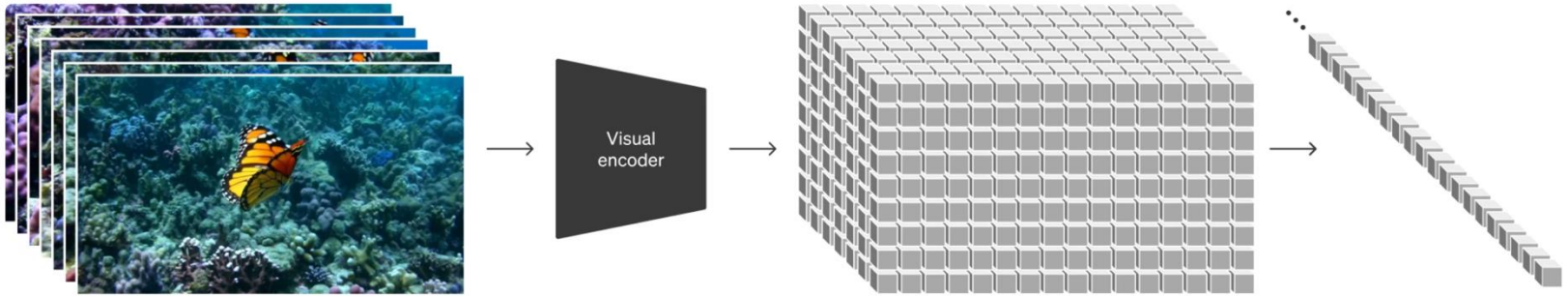
QVGen: Pushing the Limit of Quantized Video Generative Models

Yushi Huang, Ruihao Gong, Jing Liu, Yifu Ding, Chengtao Lv,
Haotong Qin, Jun Zhang

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

Background

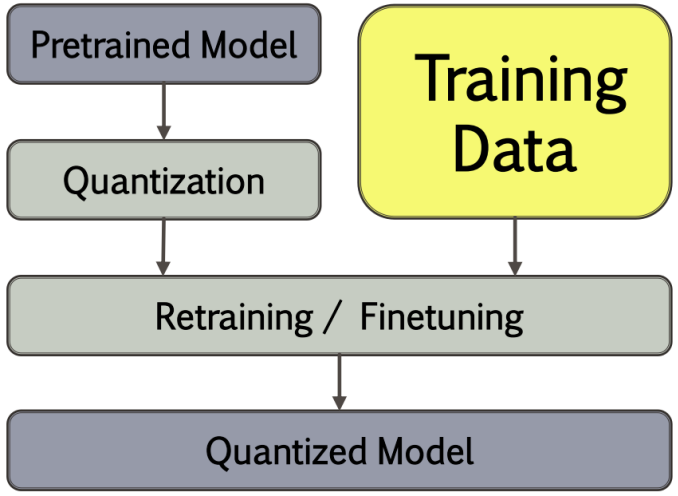
- **Video Diffusion Model** operates on spacetime patches of video and image latent codes. Visual input is represented as a sequence of spacetime patches which act as Transformer input tokens.



Given these noisy inputs (and conditioning information like text prompts), it's trained to predict the original "clean" patches.

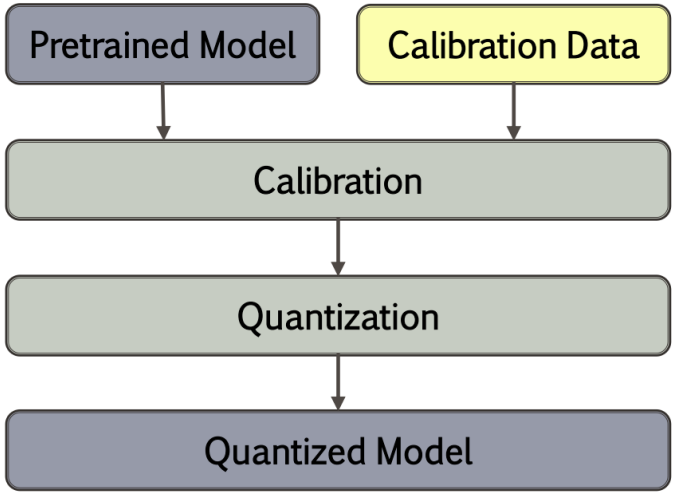
Background

- **Quantization** maps high-bit floating-point number to low-bit integer format. This technique can reduce memory costs and speedup computation on various devices.
  It can be categorized into:



Quantization-Aware Training (QAT)

Much Higher Accuracy!



Post-Training Quantization (PTQ)

Much Better Efficiency!

Quantization for Video Diffusion

full precision *vs.* 4-bit QAT (**4× mem. reduction and 3× speedup [1]**)

LSQ

Q-DM

EfficientDM

Ours

`Wan` 1.3B

Same Visual Quality

(~18 GB peak mem. and ~1 min for 480px generation)

More advanced QAT for video DMs is urgently needed!

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

[1] Li M, Lin Y, Zhang Z, et al. Svdqunat: Absorbing outliers by low-rank components for 4-bit diffusion models[J]. arXiv preprint arXiv:2411.05007, 2024

Improving Convergence with Φ

- **Motivation:** We theoretically find that minimizing $\|g_t\|_2$ is critical for improving convergence behavior of QAT:

$$\frac{R(T)}{T} \leq \frac{dD_\infty^2}{2T\eta_T} + \frac{1}{T}\sum_{t=1}^{T}\frac{\eta_t}{2}\|\boldsymbol{g}_t\|_2^2$$

- **Method:** A low $\|g_t\|_2$ presences in a stable training process (*e.g.*, few aggressive losses) [2]. We propose a Φ to achieve this:

$$\hat{\mathbf{Y}} = \mathcal{Q}_b(\mathbf{W})\mathcal{Q}_b(\mathbf{X}) + \Phi(\mathcal{Q}_b(\mathbf{X}))$$

Quantized weight

Quantized input

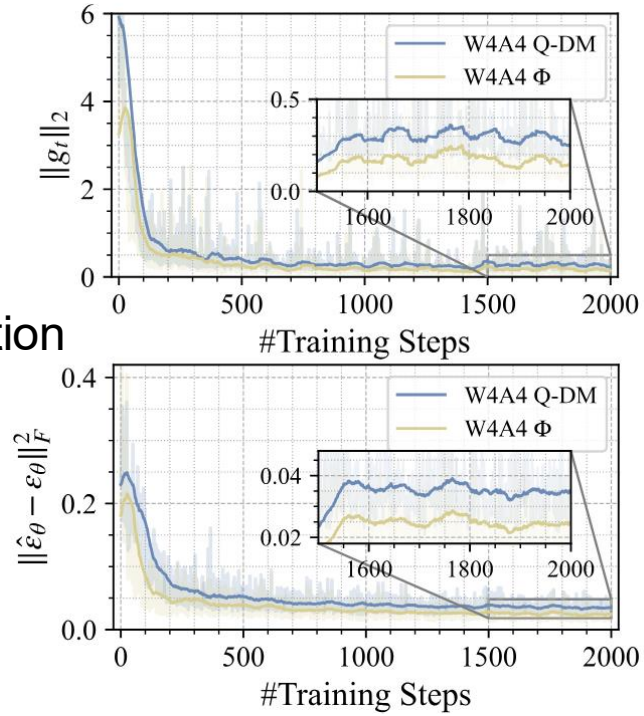
$$\mathbf{W}_\Phi \mathcal{Q}_b(\mathbf{X})$$

Learn to mitigate severe errors

Output w/ and w/o error

$$\mathbf{Y} = \mathbf{W}\mathbf{X}$$

Validation



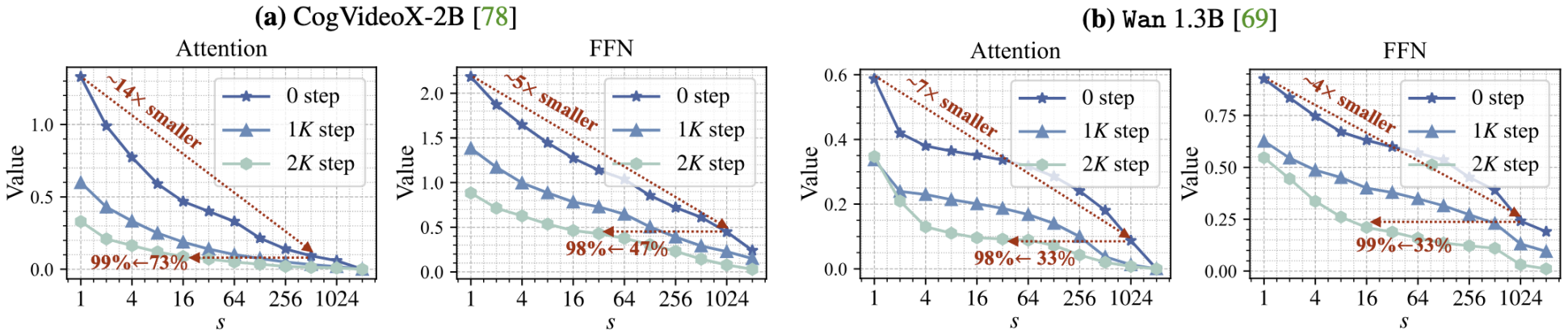
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

[2] Takase S, Kiyono S, Kobayashi S, et al. Spike no more: Stabilizing the pre-training of large language models[J]. arXiv preprint arXiv:2312.16903, 2023.

Progressively Shrinking Φ via Rank-Decay

- **Motivation**: Φ incurs significant inference overhead (severalfold than full precision). We apply SVD to the weight of Φ:

$$\mathbf{W}_\Phi = \sum_{s=1}^{d} \sigma_s \boldsymbol{u}_s \boldsymbol{v}_s^T$$



► $\mathbf{W}_\Phi$ contains a substantial number of small singular values. For example, approximately 73% of the average σ_s are $\sim 14\times$ smaller than the largest one σ_1;
► The presence of these small σ_s becomes increasingly pronounced as QAT progresses, with the proportion rising from 73% (at the 0-th step) to 99% (at the $2K$-th step).

Progressively Shrinking Φ via Rank-Decay

- **Method**: Iteratively applying the following two strategies:
  - SVD to identify the low-impact components in Φ;

Cosine decay factor

$$\gamma = \text{concat}([1]_{n\times(1-\lambda)r}, \boxed{[u]}_{n\times\lambda r})$$

$$\Phi(\mathcal{Q}_b(\mathbf{X})) = \mathbf{L}\mathbf{R}\mathcal{Q}_b(\mathbf{X}) \xrightarrow{\text{Decay}} \hat{\mathbf{Y}} = \mathcal{Q}_b(\mathbf{W})\mathcal{Q}_b(\mathbf{X}) + (\gamma \odot \mathbf{L})\mathbf{R}\mathcal{Q}_b(\mathbf{X})$$

Identify

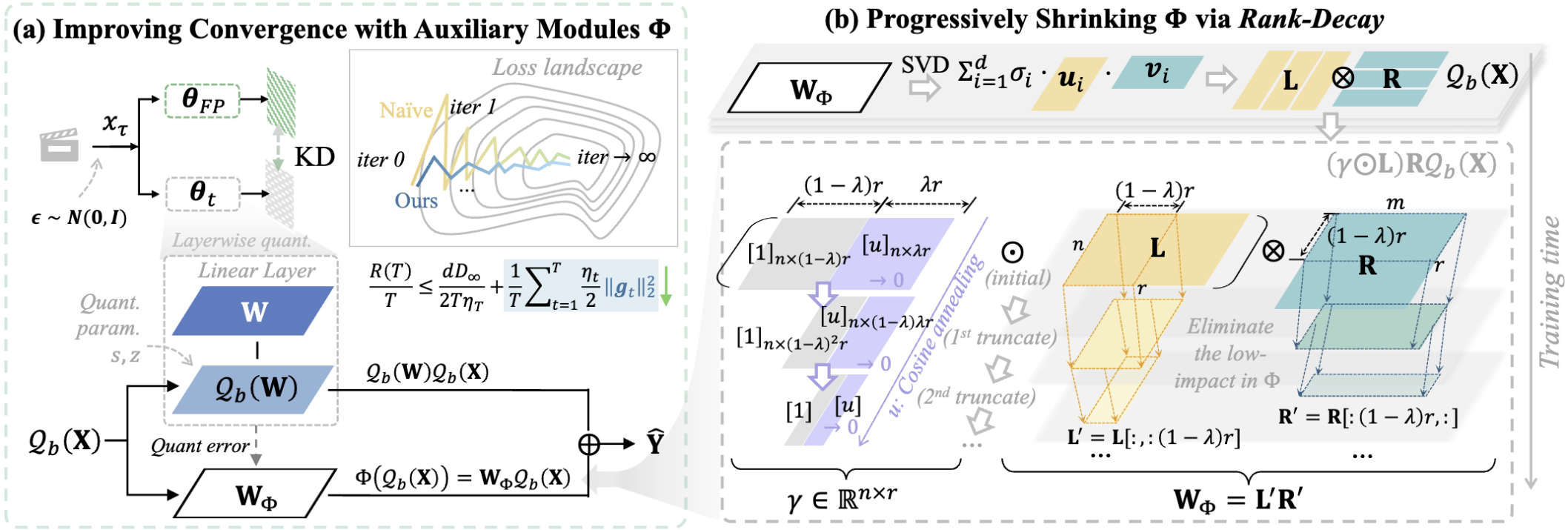
$$\mathbf{L} = [\sqrt{\sigma_1}\boldsymbol{u}_1, \ldots, \sqrt{\sigma_r}\boldsymbol{u}_r]$$
$$\mathbf{R} = [\sqrt{\sigma_1}\boldsymbol{v}_1, \ldots, \sqrt{\sigma_r}\boldsymbol{v}_r]^T$$

$$\begin{cases} \mathbf{L}' = \mathbf{L}[:, : (1-\lambda)r] \\ \mathbf{R}' = \mathbf{R}[: (1-\lambda)r, :] \end{cases} \Rightarrow \mathbf{W}_\Phi = \mathbf{L}'\mathbf{R}'$$

Shrink

  - A rank-based regularization γ to decay the identified components to ∅.

Overall Pipeline

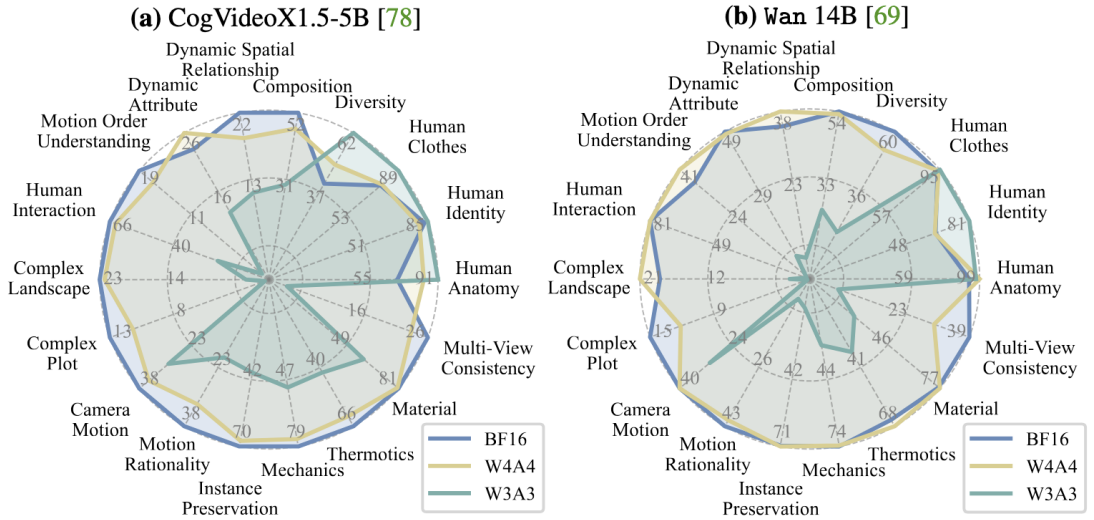


Overview of the proposed QVGen. (a) This framework integrates auxiliary modules Φ to improve training convergence (b) To maintain performance while eliminating inference overhead induced by Φ, we design a rank-decay schedule that progressively shrinks the entire Φ to ∅.

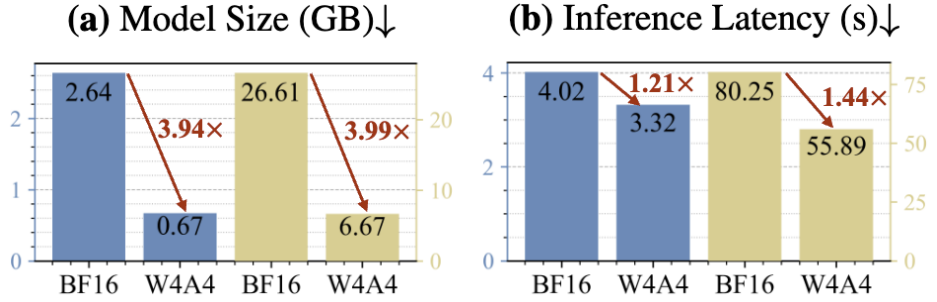
Experiments

Comparison of difference methods on Vbench.

Method	#Bits (W/A)	Imaging Quality ↑	Aesthetic Quality ↑	Motion Smoothness ↑	Dynamic Degree ↑	Background Consistency ↑	Subject Consistency ↑	Scene Consistency ↑	Overall Consistency ↑
CogVideoX-2B (CFG = 6.0, 480p, fps = 8)									
Full Prec.	16/16	59.15	54.49	97.43	67.78	94.79	92.82	36.24	25.06
ViDiT-Q [82]†	4/6	54.72	43.01	92.18	43.22	90.76	81.02	26.25	20.41
SVDQuant [34]†	4/6	58.27	47.06	95.28	40.83	92.41	87.45	27.69	21.34
SVDQuant [34]†♣	4/4	51.60	49.40	97.69	42.22	94.03	91.78	25.67	22.89
LSQ [10]*	4/4	58.73	54.20	97.57	45.00	92.97	92.41	24.06	23.17
Q-DM [37]*	4/4	54.96	52.71	98.00	48.61	93.82	91.86	28.02	23.87
EfficientDM [19]*	4/4	55.96	51.97	98.03	46.67	94.10	91.70	27.76	24.28
QVGen (Ours)*	4/4	**60.16**$_{+1.43}$	**54.61**$_{+0.41}$	**98.06**$_{+0.03}$	**67.22**$_{+18.61}$	**94.38**$_{+0.28}$	**93.01**$_{+0.60}$	**31.42**$_{+3.40}$	**24.61**$_{+0.33}$
LSQ [10]*	3/3	56.46	40.35	97.98	0.56	94.08	89.18	4.80	13.80
Q-DM [37]*	3/3	50.88	40.41	98.03	5.56	93.93	87.75	7.33	15.98
EfficientDM [19]*	3/3	52.86	44.58	97.13	28.61	93.15	88.26	15.42	20.42
QVGen (Ours)*	3/3	**58.36**$_{+1.90}$	**50.54**$_{+5.96}$	**98.37**$_{+0.34}$	**53.89**$_{+25.28}$	**94.55**$_{+0.47}$	**90.50**$_{+1.32}$	**23.85**$_{+8.43}$	**22.92**$_{+2.50}$
Wan 1.3B (CFG = 5.0, 480p, fps = 16)									
Full Prec.	16/16	64.30	58.21	97.37	70.28	95.94	93.84	28.05	24.67
ViDiT-Q [82]†	4/6	56.24	50.18	94.81	52.43	89.67	82.53	13.45	19.58
SVDQuant [34]†	4/6	58.16	51.27	97.05	49.44	93.74	91.71	14.18	23.26
SVDQuant [34]†♣	4/4	57.57	46.30	94.21	72.22	93.16	77.96	12.73	21.91
LSQ [10]*	4/4	59.11	49.09	**98.35**	71.11	92.66	91.67	10.38	18.83
Q-DM [37]*	4/4	60.40	52.50	97.22	76.67	93.37	89.26	13.28	21.63
EfficientDM [19]*	4/4	60.70	53.57	96.18	56.39	93.74	91.70	11.77	21.19
QVGen (Ours)*	4/4	**63.08**$_{+2.38}$	**54.67**$_{+1.10}$	98.25$_{-0.10}$	**77.78**$_{+1.11}$	**94.08**$_{+0.34}$	**92.57**$_{+0.87}$	**15.32**$_{+2.04}$	**23.01**$_{+1.38}$
LSQ [10]*	3/3	58.80	46.86	98.22	23.61	91.86	89.42	0.89	15.51
Q-DM [37]*	3/3	56.19	44.95	95.13	76.94	92.09	83.82	1.79	16.89
EfficientDM [19]*	3/3	42.32	33.52	96.50	70.28	92.10	74.79	0.04	11.38
QVGen (Ours)*	3/3	**67.35**$_{+8.55}$	**49.71**$_{+2.85}$	**98.93**$_{+0.71}$	**84.14**$_{+7.20}$	**93.62**$_{+1.52}$	**92.25**$_{+2.83}$	**5.71**$_{+3.92}$	**20.11**$_{+3.22}$



Results on huge models (≥5B). Our 4-bit models exhibit a minimal drop of ~1% in total score.



Evaluation of mem. and speedup (**faster kernel implementation is coming soon**).

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Qualitative Results

CogVideoX -2B

Wan 1.3B

Qualitative Results

Wan 14B

(a) BF16

(b) W4A4 QVGen (Ours)

Text prompt: *"A person is enjoying a delicious ice cream cone on a sunny day. They are standing in a bustling city park, surrounded by blooming flowers and green grass. The person has a friendly smile on their face, taking small bites of the ice cream as they savor each lick. Their casual attire includes a light jacket and jeans, with a pair of sunglasses perched on their nose. The background shows a vibrant cityscape with tall buildings and colorful street signs. The camera pans slightly from the person to capture the lively atmosphere of the park. Close-up medium shot, showing the person's joyful expression and the melting ice cream."*

(c) BF16

(d) W4A4 QVGen (Ours)

Text prompt: *"A smooth, sweeping camera circle around a lush garden. The garden is filled with vibrant flowers, tall green bushes, and neatly trimmed hedges. Sunlight filters through the leaves, casting dappled shadows on the ground. A small fountain sits in the center, gently spraying water into the air. Birds chirp and flutter among the branches. The camera gradually moves from a wide shot of the entire garden to closer views of individual plants and the intricate details of the landscape. The overall atmosphere is serene and inviting, with a soft, natural lighting style. Wide to medium shot, pans smoothly around the garden."*

Thank you!

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