Ca2-VDM: Efficient Autoregressive Video Diffusion Model with Causal Generation and Cache Sharing

Kaifeng Gao*, Jiaxin Shi*, Hanwang Zhang[4], Chunping Wang, Jun Xiao, Long Chen+

*Equal Contribution +Corresponding Author

ZHEJIANG UNIVERSITY · Xmax AI · MANYCORE · NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE · FinVolution · THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

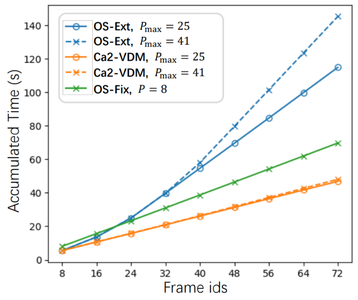
ICML International Conference On Machine Learning

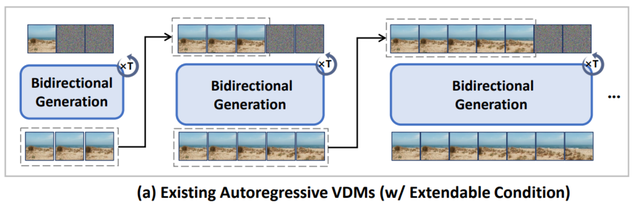
Paper Code

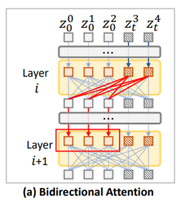
Challenges for Exisiting Autoregressive VDMs

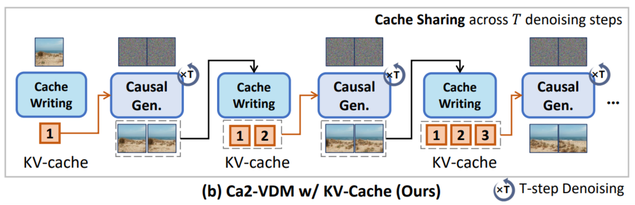
➢ **Redundant computation:**
- The conditional frames at each AR step introduce much redundant computation

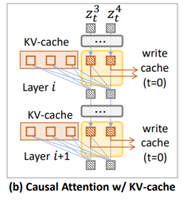
➢ **Quadratic computational complexity**
- The extended conditional frames result in an $O(n^2)$ computational demand w.r.t to the AR step









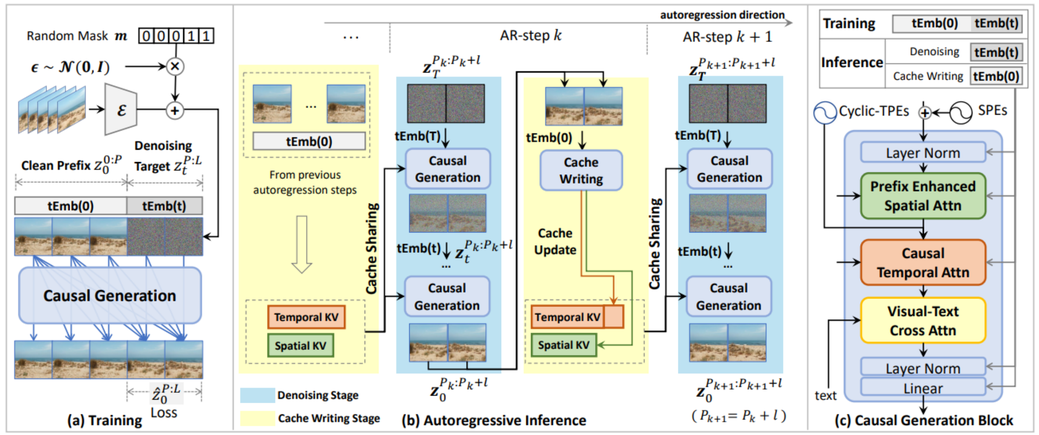


Our Contributions: CA2-VDM

➢ **KV-Cache Queue**
The key/value features of every temporal attention layer are written into the cache queue, and reused for every AR step to avoid redundant computation.

➢ **Causal Generation**
Causal temporal attention & prefix-enhanced spatial attention. They 1) ensures each generated frame only depends on preceding ones, and 2) enhances the guidance from prefix frames.

➢ **Cache Sharing** The KV-cache is shared across all the denoising steps.



Training with Clean Prefix

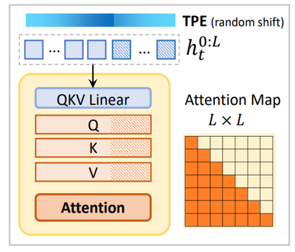
$$\widetilde{\mathcal{L}}_{\text{simple}}(\theta) = \mathop{\mathbb{E}}_{\boldsymbol{z},\boldsymbol{\epsilon},t}\left[\left\|(\boldsymbol{\epsilon}_\theta([\boldsymbol{z}_0^{0:P}, \boldsymbol{z}_t^{P:L}], t) - \boldsymbol{\epsilon}) \odot \boldsymbol{m}\right\|_2^2\right],$$

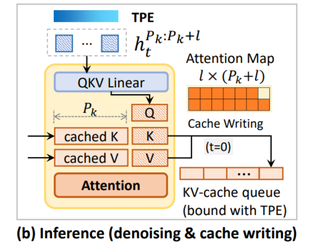
Autoregressive Inference with Cache Sharing

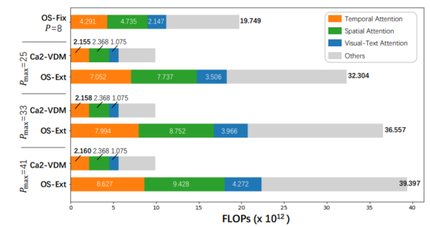
➢ **Denoising stage**

$$\boldsymbol{z}_{t-1}^{P_k:P_k+l} \sim p_\theta(\boldsymbol{z}_{t-1}^{P_k:P_k+l} | \boldsymbol{z}_t^{P_k:P_k+l}, \boldsymbol{z}_0^{0:P_k}). \quad \text{CausalAttn}(\boldsymbol{Q}_t^{P_k:P_k+l}, [\boldsymbol{K}_0^{0:P_k}, \boldsymbol{K}_t^{P_k:P_k+l}], [\boldsymbol{V}_0^{0:P_k}, \boldsymbol{V}_t^{P_k:P_k+l}]),$$

➢ **Cache writing stage**
The denoised $\boldsymbol{z}_0^{k:k+l}$ is input to the model again to compute its spatial and temporal KV-cache, which will be used in the next AR step







Quantitative Results

Zero-shot FVD on MSR-VTT

Method	C	MSR-VTT	UCF101
ModelScope (Wang et al., 2023b)	T	550	410
VideoComposer (Wang et al., 2023c)	T	580	-
Video-LDM (Blattmann et al., 2023b)	T	-	550.6
PYoCo (Ge et al., 2023)	T	-	355.2
Make-A-Video (Singer et al., 2023)	T	-	367.2
AnimateAnything (Dai et al., 2023)	T+I	443	-
PixelDance (Zeng et al., 2024)	T+I	381	**242.8**
SEINE (Chen et al., 2024b)	T+I	**181**	-
Ca2-VDM	T+I	**181**	277.7

Finetuned FVD on UCF-101

Method	Res.	FVD
MCVD (Voleti et al., 2022)	64^2	1143
VDT (Lu et al., 2024)	64^2	225.7
DIGAN* (Yu et al., 2022)	128^2	577
TATS (Ge et al., 2022)	128^2	420
VideoFusion (Luo et al., 2023)	128^2	220
LVDM* (He et al., 2022)	256^2	372
PVDM (Yu et al., 2023)	256^2	343.6
Latte (Ma et al., 2025)	256^2	333.6
Ca2-VDM	256^2	**184.5**

VBench Evaluation

Method	Aesthetic Quality	Imaging Quality	Motion Smoothness	Temporal Flickering
OS-Ext	44.39	50.74	98.93	98.57
Ca2-VDM	44.30	50.55	97.59	97.14

Generation Time Cost (80 frames)

Method	Ext.C.	Time (s)
StreamT2V		150
OS-Ext	✓	130.1
OS-Fix		77.5
Ca2-VDM	✓	**52.1**

GPU Memory Usage w/ KV-Cache

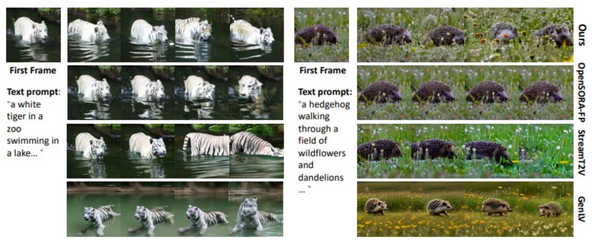
Method	Denoising Steps (T)	Model Forward Shape (B, C, L, H, W)	KV-cache Shape $(T, L_{\text{cond}}, hw, C')$	KV-cache Memory Cost	Total Memory Cost
Live2diff	4	(4 , 4, 1, 32, 32)	(4, 16, $h'w'$, C')	1.42 GB	10.90 GB
Live2diff	50	(50, 4, 1, 32, 32)	(50, 16, $h'w'$, C')	17.70 GB	29.46 GB
Ca2-VDM w/ PE	50	(1 , 4, 8, 32, 32)	(1, 25, hw, C)	0.86 GB	4.79 GB
Ca2-VDM w/o PE	50	(1 , 4, 8, 32, 32)	(1, 25, hw, C)	0.77 GB	3.95 GB

Qualitative Results

➢ **Better transition consistency**
E.g., the purple flower in the result of OpenSORA-PC starts changing at the 85th frame.

➢ **Better long-term consistency**
Our results have no severe content mutations compared to theirs (e.g., 24~25th frames of GenLV)

➢ **More Efficient Generation with Comparable Visual Quality.**
- Faster generation speed
- Less GPU memory cost w/ KV-cache



Long-term generation results vs. Open-SORA Extendable Condition (OS-Ext)