Hierarchical Patch Diffusion Models for High-Resolution Video Generation

Ivan Skorokhodov[1,2], Willi Menapace[1,3], Aliaksandr Siarohin[1], Sergey Tulyakov[1]
Snap Inc.[1], KAUST [2], University of Trento[2]

Snap Inc. King Abdullah University of Science and Technology UNIVERSITÀ DI TRENTO

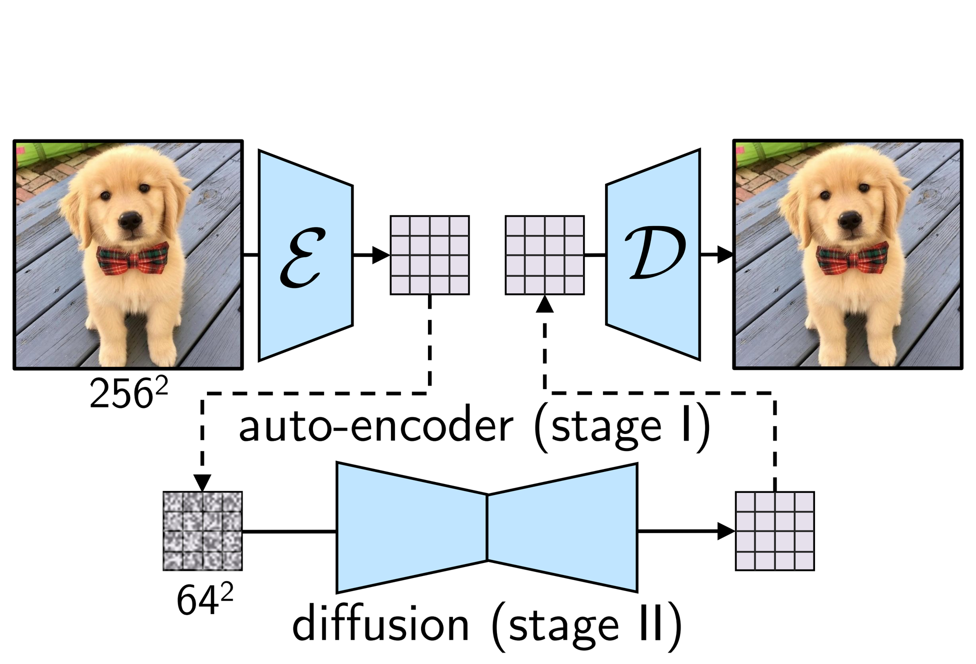
CVPR SEATTLE, WA JUNE 17-21, 2024

Overview

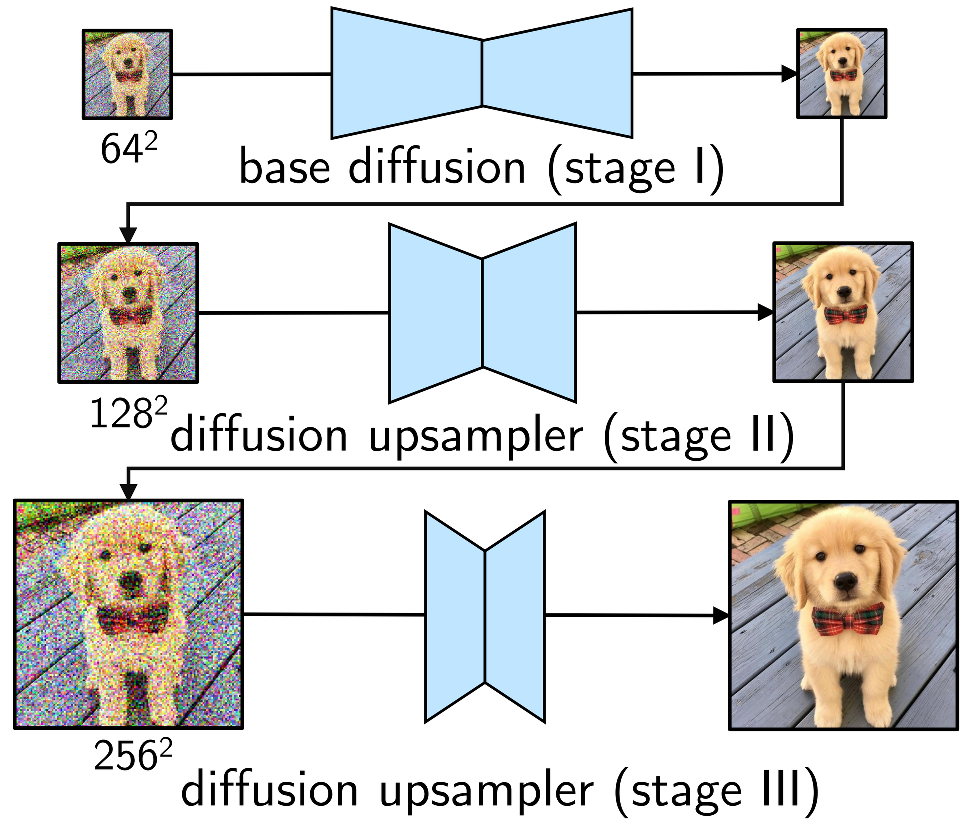
Latent DMs and Cascaded DMs are not end-to-end:

- They consist of multiple models and optimization stages
- This complicates training, inference and downstream applications

Latent Diffusion Model (LDM) [1]

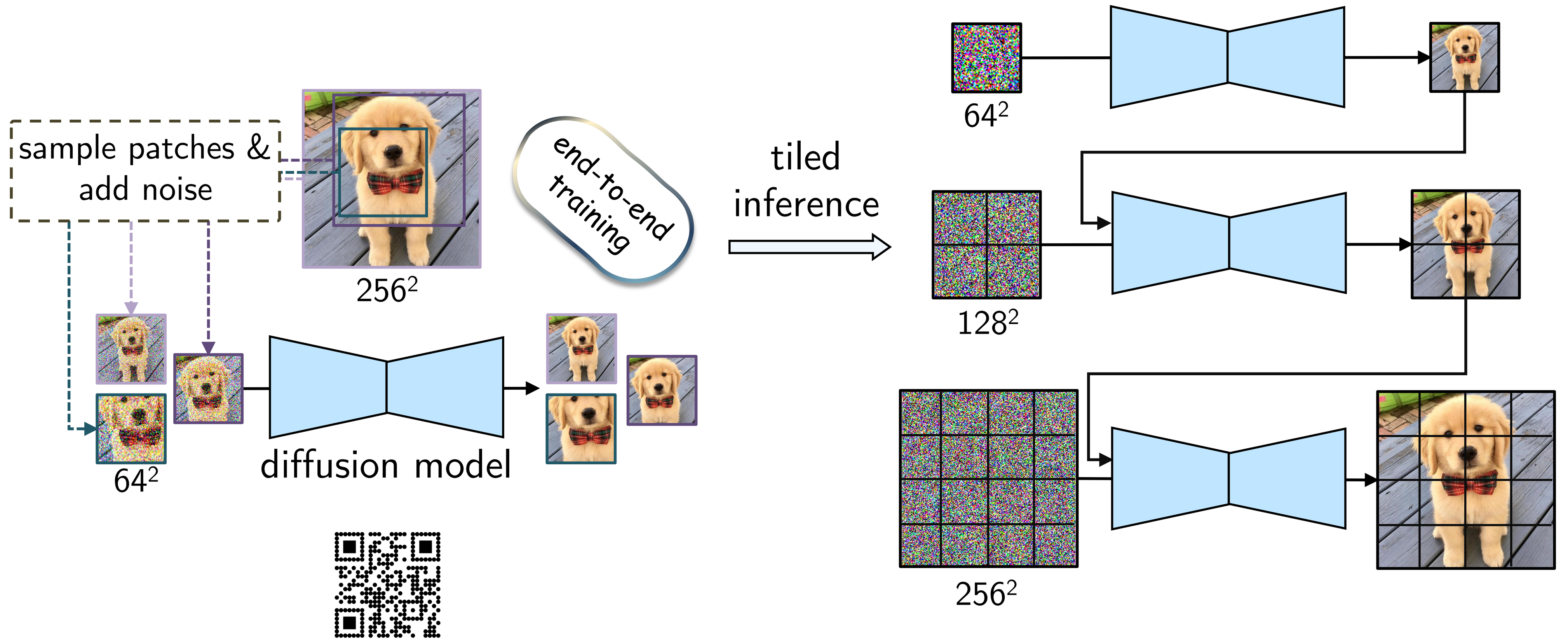


Cascaded Diffusion Model (CDM) [2]



We design a **Hierarchical Patch Diffusion Model (HPDM)**:

- End-to-end high-resolution video diffusion model;
- Obtains SotA results on UCF and comparable results on text2video;
- Can be quickly fine-tuned from a low-res video generator.



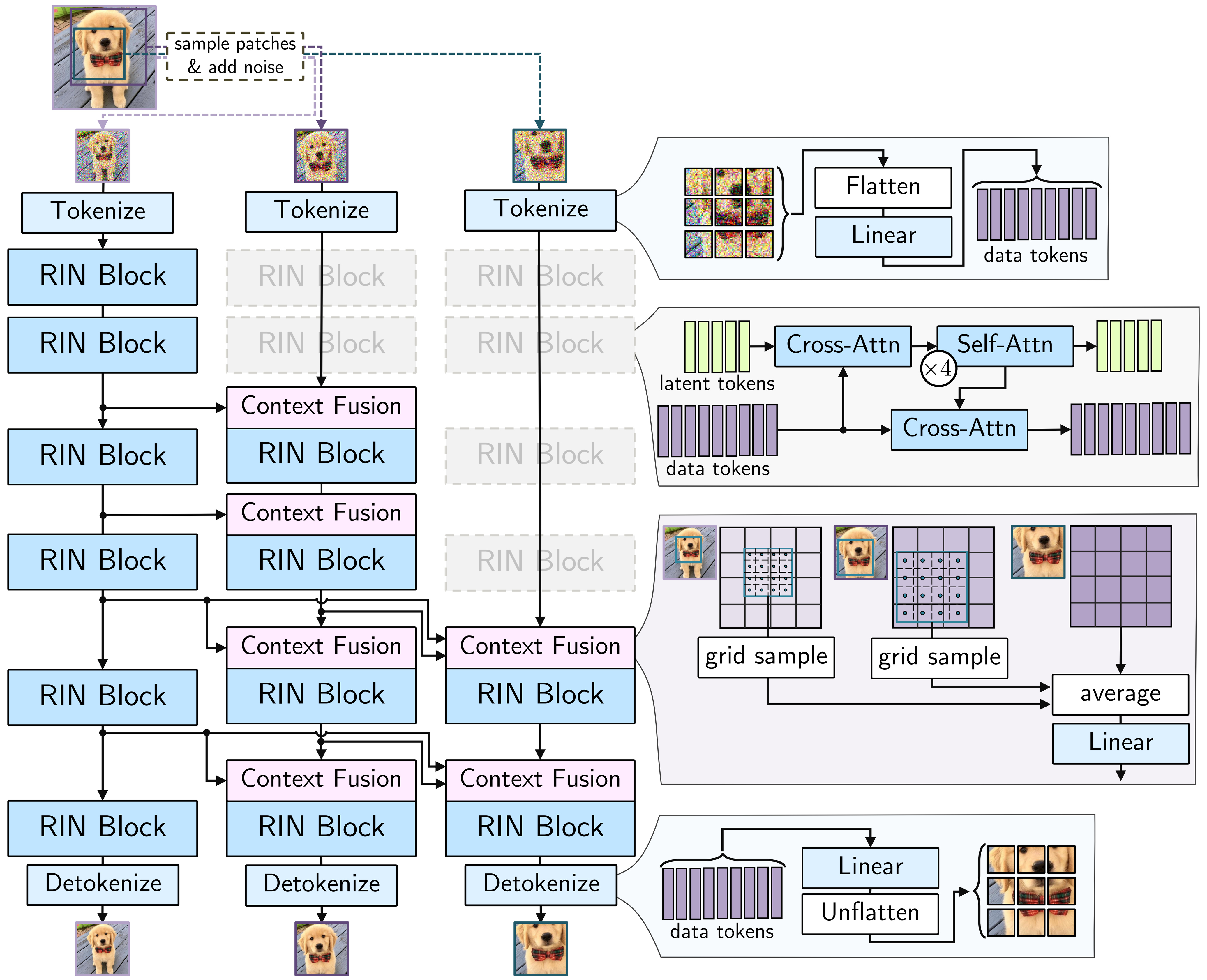
https://snap-research.github.io/hpdm/

Hierarchical Patch Diffusion Model

HPDM is latent transformer-based [3] joint patch diffusion model, and is based on three key ideas:

- *Hierarchical patch structure:* it trains jointly on a hierarchy of patches, "nested" into each other;
- *Context fusion:* input features from lower stages to higher ones;
- *Adaptive computation:* using fewer blocks in higher stages to reduce computational and memory costs.

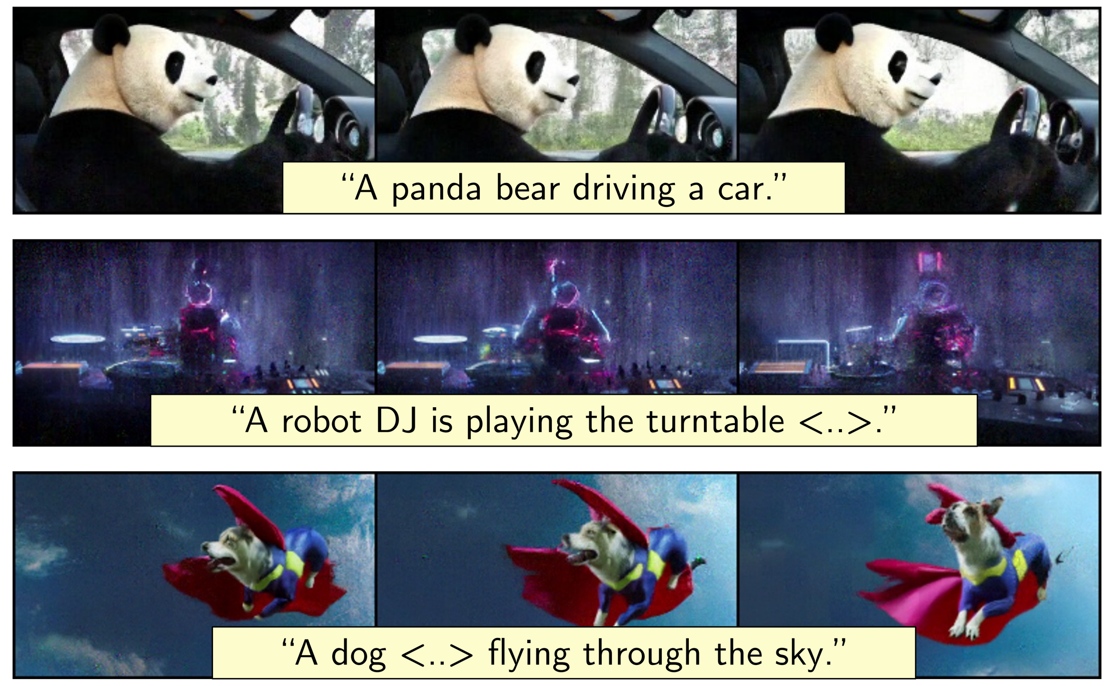
Surprisingly, a SotA video generator can be trained using just **up to ≈1.5% of the pixels from each video!**



Results

Method	FVD↓	IS↑
MoCoGAN-HD	700	33.95
TATS	635	57.63
VIDM	294.7	-
PVDM	343.6	74.4
Make-A-Video	81.25	82.55
HDPM-S	344.5	73.73
HPDM-M	143.1	84.29
HPDM-L	66.32	87.68

SotA results on UCF 256^2



64x288x512 text-to-video results after 15k fine-tuning steps of 16x36x64 SnapVideo [4]

Ablations

Method	FVD_{512}↓	FVD_{512}↓	FVD_{512}↓	Training speed ↑
Shallow context fusion	298.9	411.9	467.0	4.91
Detaching context from the graph	290.6	375.0	397.3	4.4
Non-adaptive computation	319.3	391.5	373.9	2.73
No coordinates conditioning	305.3	400.7	389.5	4.47
HPDM (full model)	287.6	376.6	378.2	4.4

Limitations

- *Stitching artifacts* due to tiled inference (though overlapping helps)
- *Slow inference*: NFEs grow exponentially with the number of stages
- *Error propagation*: errors in lower stages propagate to higher ones
- *"Dead" pixels*: transformer-based DMs are prone to spatial inconsistency

References

[1] Ho et al., "Cascaded Diffusion Models for High Fidelity Image Generation", JMLR 23 (2022)
[2] Rombach et al., "High-Resolution Image Synthesis with Latent Diffusion Models", CVPR 2022
[3] Jabri et al., "Scalable Adaptive Computation for Iterative Generation", ICML 2023
[4] Menapace et al., "Snap Video: Scaled Spatiotemporal Transformers for Text-to-Video Synthesis", CVPR 2024