NeRV-Diffusion: Diffuse Implicit Neural Representation for Video Synthesis

Yixuan Ren, Hanyu Wang, Bo He, Hao Chen, Abhinav Shrivastava

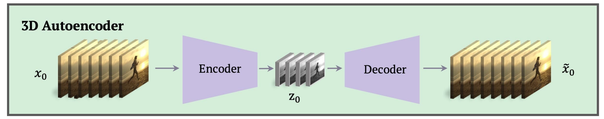
UNIVERSITY OF MARYLAND

ICLR

TL;DR: NeRV-Diffusion synthesizes videos by generating self-decoding INR weights from noise via diffusion, without temporal attention.

Motivation

Inefficiency in Video VAE and Diffusion



- Fixed Downsampling Factor
  - Quadratically (r^2) latent size w.r.t. RGB resolution
- Frame-wise Representation
- Cross-Frame Temporal Attention
  - Extra computation for temporal consistency
- Holistic 1D Tokenization
  - Usually discrete, compromise temporal granularity

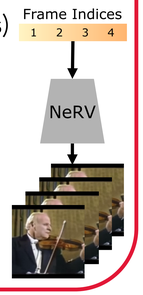
Preliminary

NeRV: Neural Representation for Videos

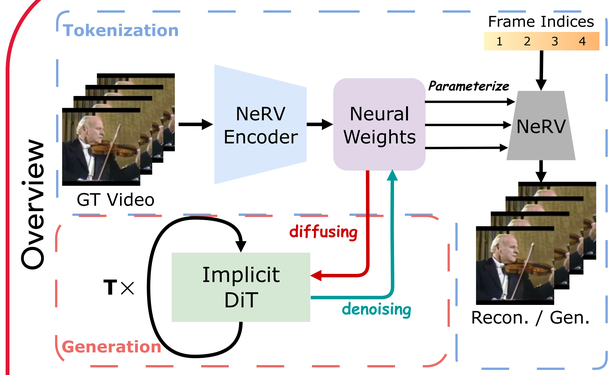
- Implicit Neural Representations (INRs)

$$(x,y,z,\theta,\phi) \rightarrow F_\Theta \rightarrow (RGB\sigma)$$

- NeRV: a convolutional INR for video
  - Input: frame index t
  - Output: one whole frame
- Video compression
  - Decode all frames with the same parameters for temporal coherency



Models



Stage 1: Tokenization by NeRV-VAE

1.1 NeRV Encoder compresses RGB video into Neural Weights;

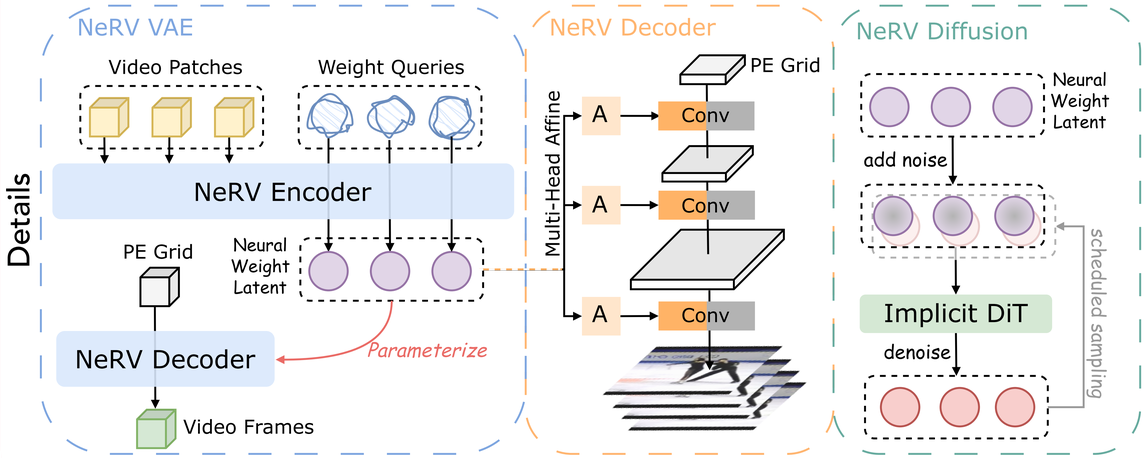
1.2 Neural Weights parameterize NeRV Decoder;

1.3 NeRV Decoder takes in Frame Indices and reconstruct RGB frames.

Stage 2: Generation by Implicit DiT

2.1 Noise is added on clean Neural Weights;

2.2 Implicit DiT denoises on Neural Weights.



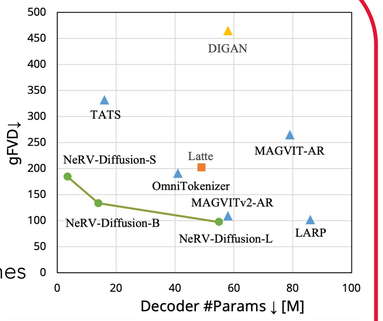
Neural Weight Latent is reused for all NeRV Decoder layers after Multi-Head Affine transforms. Transformed latent is reshaped and assigned as convolution kernels along channels of NeRV Decoder.

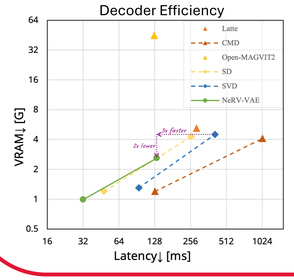
Experiments

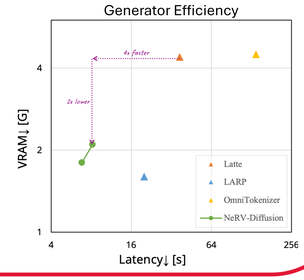
UCF/Kinetics
16/128 frames
128/256 resolution

Compression Ratio:
r128: $^{16\times128\times128}/_{128} = 2K$
r256: $^{16\times256\times256}/_{160} = 6.5K$

Efficient long-video training with temporally spaced frames + temporal interpolation







Highlights

- *Instance-specific decoding* for efficient, high-fidelity video synthesis
- *Holistic continuous* video representation that *bypasses temporal attention* while preserving temporal interpolability
- *Sublinear scaling overhead* w.r.t. video resolution and length