Householder Motion Embedding

Noise Modulation for Motion-Controllable Video Generation

Tao Jun Lin[1,2*] **Jinhong Ni**[1] **Ruyi Zha**[1] **Yifu Wang**[2] **Xibin Song**[2] **Pan Ji**[2] **Hongdong Li**[1,3]

[1]Australian National University [2]Vertex Lab [3]Amazon Australia

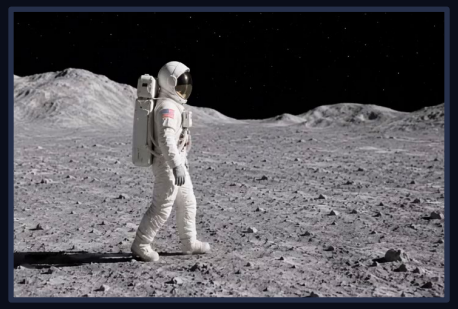
* Project lead



NEURAL INFORMATION
PROCESSING SYSTEMS

THE TASK

Motion-controlled video generation



first frame

+



motion input

→



starts from random noise

→



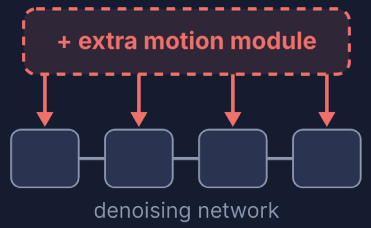
output video

How to embed motion input? Two existing routes:

① Change the network

ControlNet · camera embeddings

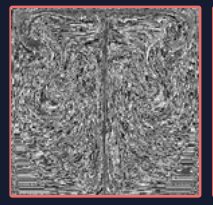
- ✗ Modifies the architecture
- ✗ Model-specific
- ✗ Extra modules to train



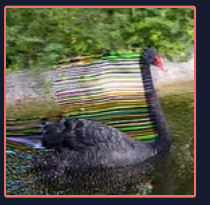
② Warp the noise

HIWYN · InfRes · GWTF

- ✓ Architecture unchanged
- ✗ Breaks Gaussian noise → artifacts
- ✗ Expensive to fix



warped noise

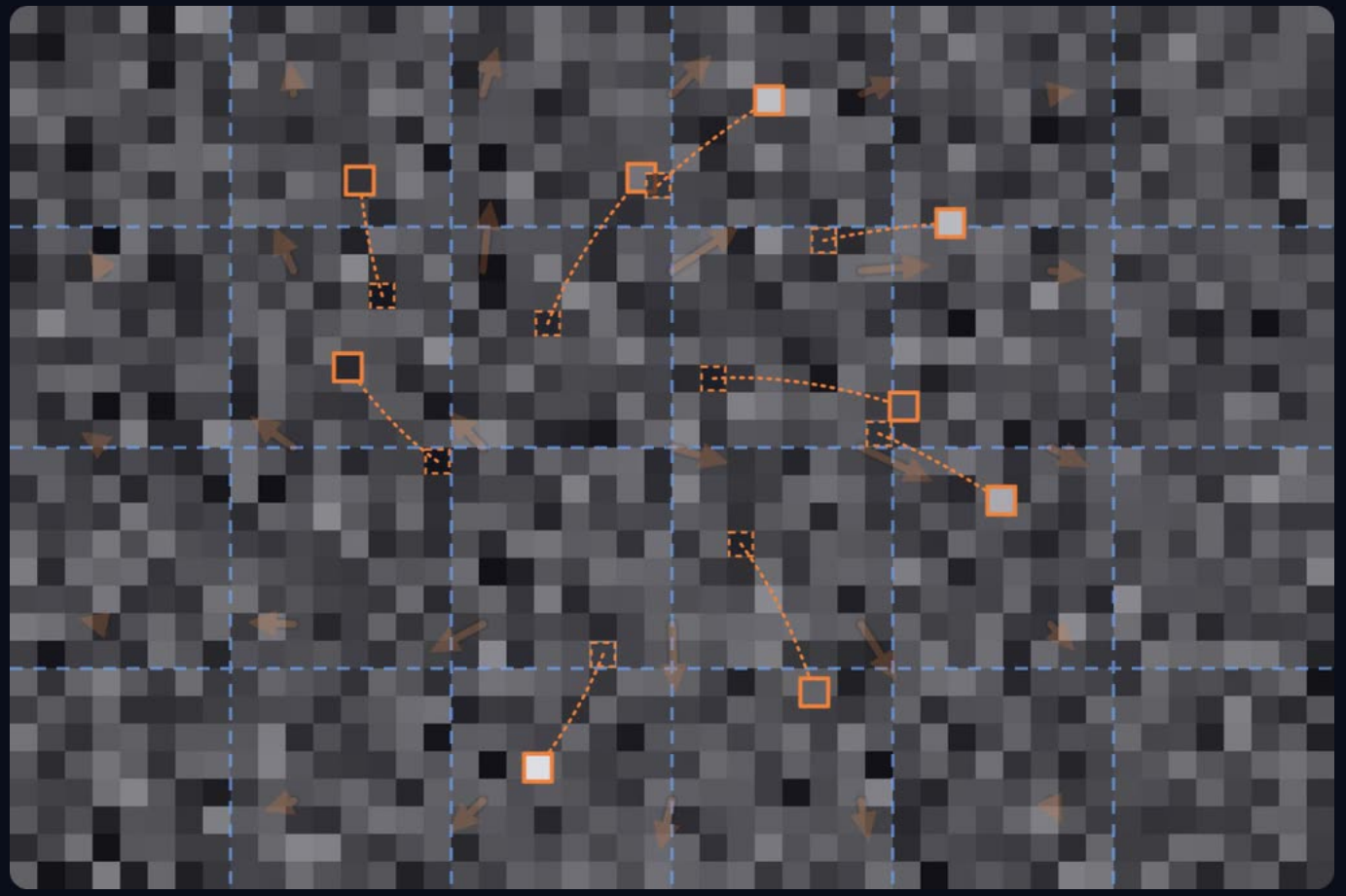


texture sticking

Can we embed motion in the noise without warping it?

PRIOR NOISE WARPING BASED METHODS

Spatial noise warping with a diffeomorphic deformation field



latent noise → 8×8 sub-pixels per latent pixel → warped and transported back to the latent grid

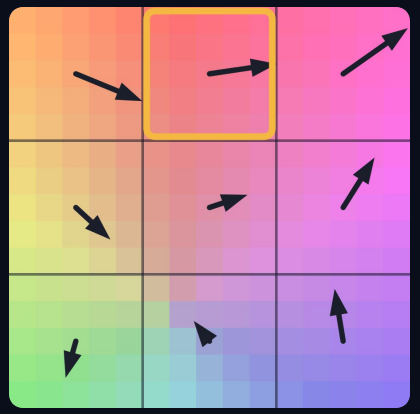
1. Up-sample
2. Warp
3. Transport

64× more samples to warp & rasterise
cheaper interpolation **breaks Gaussianity**

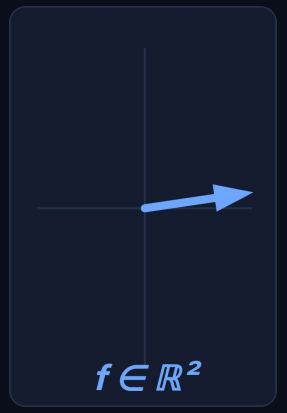
OUR IDEA

Don't move the noise — transform it in channel space

Exact per-frame Gaussianity | No correction step | No learnable parameters | Drop-in at noise initialisation

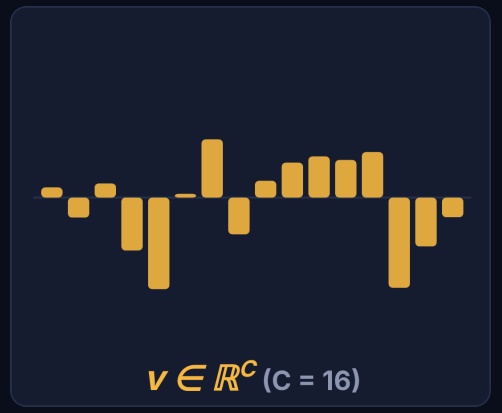


optical flow at each latent location



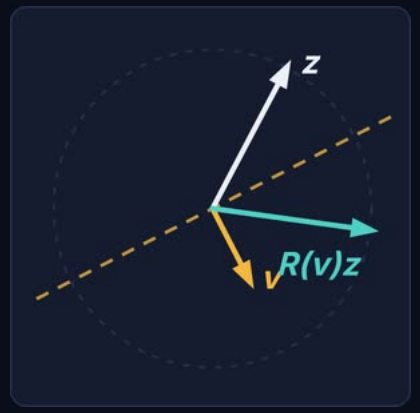
one location's flow

A f + b



lifted to channel space

$v_k = A_k f + b_k$



Householder reflection

$R(v) = I - 2\bar{v}\bar{v}^T$

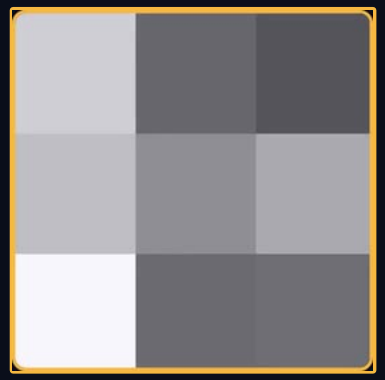
Applied **at every latent location**, with K reflections composed:

$Q = R(v_1) \cdots R(v_K)$



channel vectors per location

per-location orthogonal transform



correlated noise · positions never move

OURS VS NOISE WARPING · SAME INPUT

Same flow, two ways to build the noise



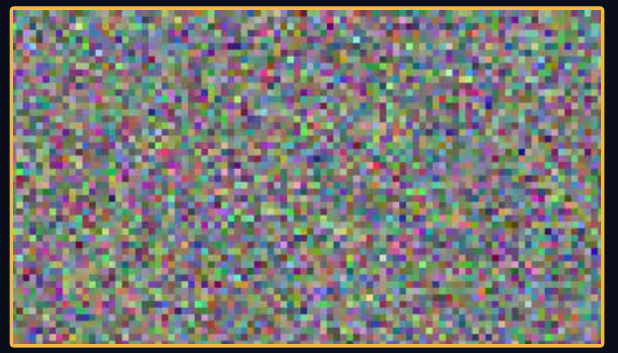
reference video



optical flow (RAFT)

Ours

noise stays in place; channels are transformed



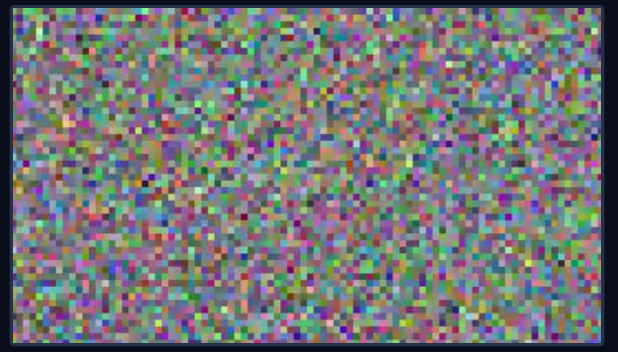
Householder noise



generated video

GWTF

noise pixels are moved along the flow



warped noise



generated video

Motion transfer



reference video (motion source)



Ours (HME)

Local object motion control



drag-defined motion signal



Ours (HME)

REAL-WORLD APPLICATIONS · 3 / 3

Camera motion control



camera-trajectory motion signal



Ours (HME)

QUALITATIVE COMPARISONS · DAVIS · MC-BENCH · DL3DV

Follows the motion, keeps the scene

Reference / signal	Ours (HME)	GWTF (noise warping)	TTM
			
			
			

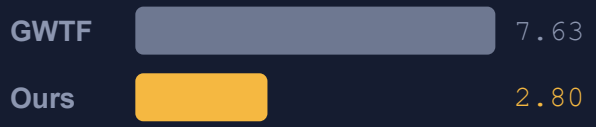
VERSUS THE STRONGEST NOISE-WARPING BASELINE

Cheaper noise, better motion

2.7× faster

noise construction

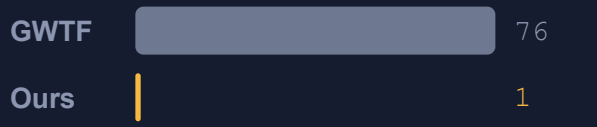
ms / latent · lower is better



1 MB vs 76 MB

extra GPU memory

MB · lower is better



21% lower

flow error on DAVIS

EPE · lower is better



Time & memory: 1024×1024 latent noise, C = 16, K = 2 Householder reflections. Flow EPE: DAVIS motion transfer, ours γ = 0.3 vs GWTF γ = 0.5, CogVideoX-5B-I2V.

Thank You!

Householder Motion Embedding

Noise Modulation for Motion-Controllable Video Generation

Tao Jun Lin[1,2*] **Jinhong Ni**[1] **Ruyi Zha**[1] **Yifu Wang**[2] **Xibin Song**[2]
Pan Ji[2] **Hongdong Li**[1,3]

[1]Australian National University [2]Vertex Lab [3]Amazon Australia

* Project lead

Australian National University 极顶数创 VERTEXLAB amazon

NEURAL INFORMATION PROCESSING SYSTEMS

Project page

coldsilence507.github.io/HouseholderMotionEmbedding