How I Warped Your Noise: A Temporally-Correlated Noise Prior For Diffusion Models

Pascal Chang   Jingwei Tang   Markus Gross   Vinicius C. Azevedo
ICLR 2024 (Oral)

Weekly Meeting - 2024-04-12
KAIST Geometric AI Lab - Jaihoon Kim

Motivation

Image Diffusion Model

DALL-E 3

Stable Diffusion

Video Generation

Sora

Lumiere

Video Editing

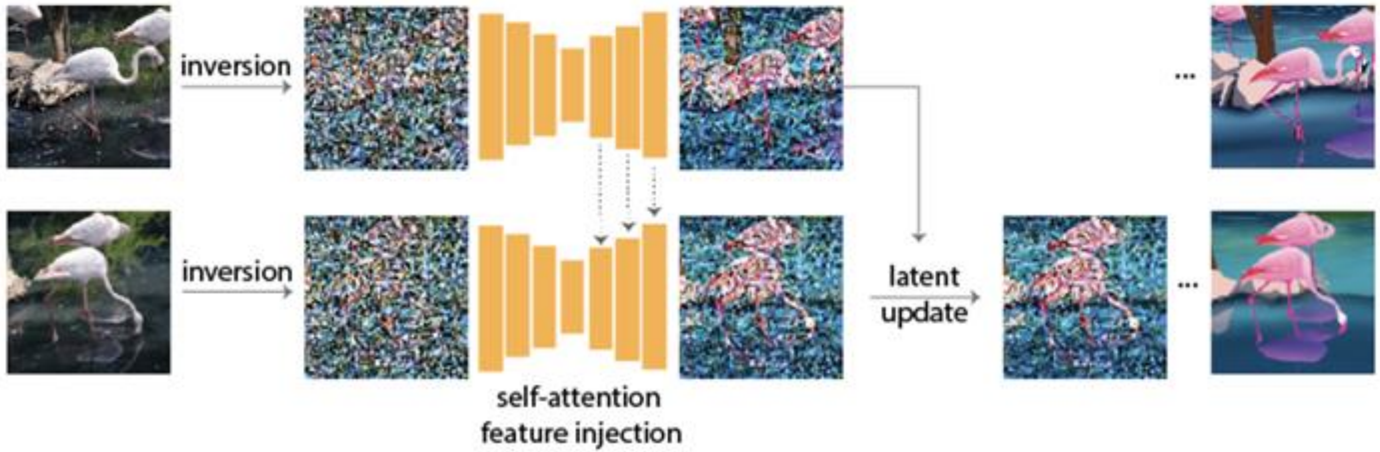
“giraffe with space suit standing on the moon”

“a kite-surfer in the magical starry ocean with an aurora borealis in the background”

Pix2Video

Video Editing

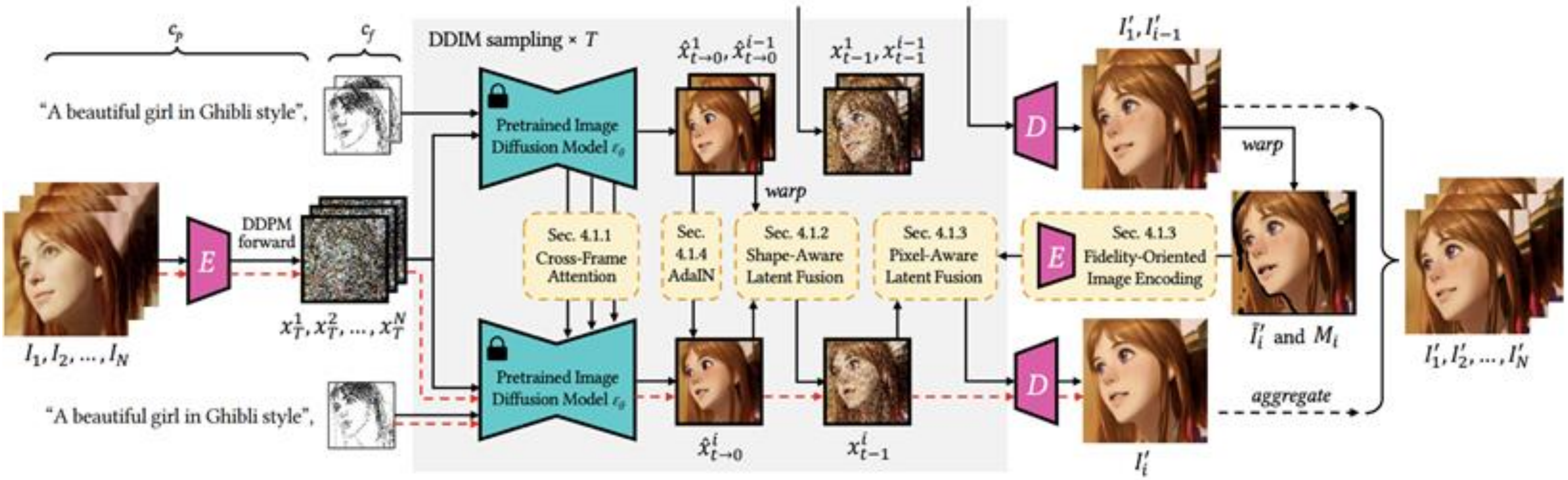
Inversion-based method: Entangled spatial & temporal correlations



Pix2Video

Video Editing

Fine Tuning: Fails to model high frequency details



Render a Video

Video Editing

How can we create a noise prior that preserves correlations present in an image sequence?

Video Editing - Limitations

Input

"Oil painting style"

Random noise

Temporal inconsistency

Fixed noise

Texture-Sticking artifacts

Video Editing - Limitations

Input

Random noise

Fixed noise

“Oil painting style”

Noise warping

- Temporal consistency
- Artifact-free

Method

i) Continuous Noise Representation

Discrete 2D Gaussian (DxD)

$$G : (i,j) \in \{1,\ldots,D\}^2 \rightarrow X_{i,j}$$

Infinite-resolution White Gaussian

$$W : A \in \mathcal{E} \rightarrow W(A) \sim \mathcal{N}(0, \nu(A))$$

$E = [0,D] \times [0,D]$

$\mathcal{E} = \mathcal{B}(E)$

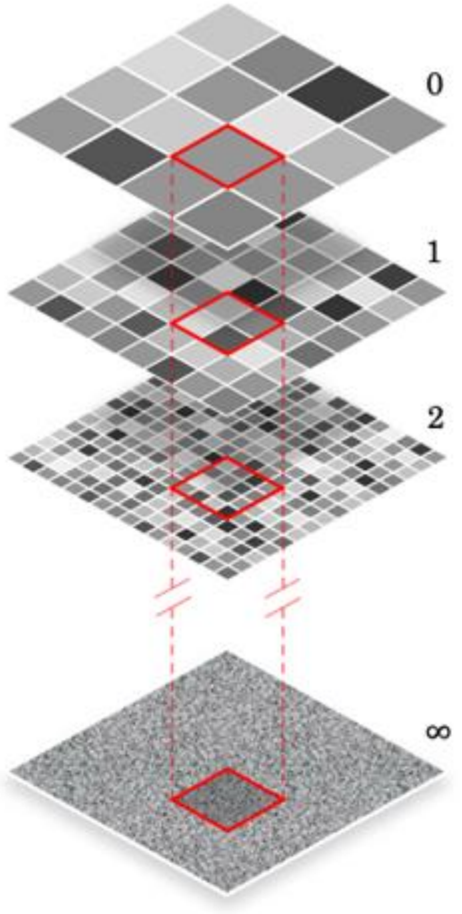
Level 0: Partitions the domain E into D x D.

Level k: Subdivide each pixel in level 0 into $N_k = 2^k \times 2^k$ sub pixels.

$\mathbb{A}^0 \subseteq \mathcal{E}$

$\mathbb{A}^k \subseteq \mathcal{E}$

$$\sum_{i=1}^{N_k} W(A_i^k) = W(\bigcup_{i=1}^{N_k} A_i^k) = W(A^0)$$



i) Continuous Noise Representation

White noise and Brownian motion. An alternative definition of white noise $\{W(\mathbf{x})\}_{\mathbf{x}\in E}$ is through the *distributional derivative* of a Brownian motion $\{B(\mathbf{x})\}_{\mathbf{x}\in E}$ (also called *Brownian sheet* for dimension ≥ 2) as $\boxed{W(\mathbf{x})d\mathbf{x} = dB(\mathbf{x}).}$

Itô integral. As the Itô integral $\int_A \phi(\mathbf{x})dB(\mathbf{x})$ of a deterministic function ϕ in L^2 is always Gaussian. The variance is given by:

$$\boxed{\mathbb{E}\left(\int \phi(\mathbf{x})dB(\mathbf{x})\right)^2 = \int \phi(\mathbf{x})^2 d\mathbf{x} = \|\phi\|_2^2.} \tag{16}$$

From Equation (16) we can relate back to the first definition of white noise by setting $\phi = \mathbf{1}$. Indeed,

$$W(A) = \int_{\mathbf{x}\in A} \mathbf{1} dB(\mathbf{x}) \tag{17}$$

is a Gaussian variable of variance $\int_{\mathbf{x}\in A} \phi(\mathbf{x})^2 d\mathbf{x} = \int_{\mathbf{x}\in A} 1 d\mathbf{x} = \nu(A)$.

ii) Conditional White Noise Sampling

Partitioning **X** = x into *N* x *N* sub-patches

$$\{B_{k,l} = [\tfrac{k-1}{N}, \tfrac{k}{N}] \times [\tfrac{l-1}{N}, \tfrac{l}{N}]\}_{(k,l)\in[1,N]^2} \qquad \mathbf{Y} = (W(B_{1,1}), ..., W(B_{N,N}))^\top$$

Given **Z** = (**Y**, **X**), the covariance between two individual Gaussian random variables is their intersected area $\nu(B_{k,l} \cap A) = \nu(B_{k,l}) = \frac{1}{N^2}$.

$$\mathbf{Z} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \text{with } \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_\mathbf{Y} \\ \boldsymbol{\mu}_\mathbf{X} \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{C}_{(\mathbf{Y},\mathbf{Y})} & \mathbf{C}_{(\mathbf{Y},\mathbf{X})} \\ \mathbf{C}_{(\mathbf{X},\mathbf{Y})} & \mathbf{C}_{(\mathbf{X},\mathbf{X})} \end{bmatrix}$$

$$\mathbf{C}_{(\mathbf{Y},\mathbf{Y})} = \frac{1}{N^2}\mathbf{I}_{N^2}, \quad \mathbf{C}_{(\mathbf{Y},\mathbf{X})} = \frac{1}{N^2}\mathbf{u}, \quad \mathbf{C}_{(\mathbf{X},\mathbf{Y})} = \mathbf{C}_{(\mathbf{Y},\mathbf{X})}^\top, \quad \mathbf{C}_{(\mathbf{X},\mathbf{X})} = \mathbf{I}_1,$$

$$\boldsymbol{\mu}_{(\mathbf{Y}|\mathbf{X})} = \boldsymbol{\mu}_\mathbf{Y} + \mathbf{C}_{(\mathbf{Y},\mathbf{X})} \left(\mathbf{C}_{(\mathbf{X},\mathbf{X})}\right)^{-1} (\mathbf{X} - \boldsymbol{\mu}_\mathbf{X}) = \mathbf{0}_{N^2} + \frac{1}{N^2}\mathbf{u}(\mathbf{X} - \mathbf{0}_1) = \frac{x}{N^2}\mathbf{u},$$

$$\boldsymbol{\Sigma}_{(\mathbf{Y}|\mathbf{X})} = \mathbf{C}_{(\mathbf{Y},\mathbf{Y})} - \mathbf{C}_{(\mathbf{Y},\mathbf{X})}\,\mathbf{C}_{(\mathbf{X},\mathbf{Y})} = \frac{1}{N^2}\mathbf{I}_{N^2} - \frac{1}{N^2}\mathbf{u}(\frac{1}{N^2}\mathbf{u})^\top = \frac{1}{N^2}I_{N^2} - \frac{1}{N^4}\mathbf{u}\mathbf{u}^\top.$$

$$(W(\mathbb{A}^k)|W(A^0) = x) \sim \mathcal{N}\left(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}}\right), \quad \text{with } \boxed{\bar{\boldsymbol{\mu}} = \frac{x}{N_k}\mathbf{u}, \bar{\boldsymbol{\Sigma}} = \frac{1}{N_k}\left(\boldsymbol{I}_{N_k} - \frac{1}{N_k}\mathbf{u}\mathbf{u}^\top\right)}$$

ii) Conditional White Noise Sampling

In practice, they generate higher- resolution discrete noise with unit variance

$$\boldsymbol{U} = \sqrt{N_k}\bar{\boldsymbol{\Sigma}}$$

$$\begin{aligned}(\mathbb{A}^k)|W(A^0) = x) &= \bar{\boldsymbol{\mu}} + \boldsymbol{U}Z \\ &= \frac{x}{N_k}\mathbf{u} + \frac{1}{\sqrt{N_k}}(Z - \langle Z\rangle\mathbf{u}), \quad \text{with } Z \sim (\mathbf{0}, \boldsymbol{I})\end{aligned}$$

1) Sample a discrete $N \times N$ Gaussian sample
2) remove its mean from it, and
3) add the pixel value x.

iii) Noise Transport Equation (Theoretical)

Noise is transported given a deformation field $\mathcal{T} : E \to E$

Differential form of the pixel ρ(**x**) transport and warped field (point):

$$\frac{\partial \rho(\mathbf{x})}{\partial t} = -\nabla \cdot (\rho(\mathbf{x})\, \mathbf{v}(\mathbf{x})) \qquad \tilde{\rho}(\mathbf{x}) = \rho(\mathcal{T}^{-1}(\mathbf{x}))$$

When transporting area, the cell-centered area-averaged integration is used.

$$\rho(\Delta\mathbf{x}) = \frac{1}{\nu(\Delta\mathbf{x})} \int_{\mathbf{x}\in\Delta\mathbf{x}} \rho(\mathbf{x})\, d\mathbf{x}.$$

$$\tilde{\rho}(\Delta\mathbf{x}) = \rho(\mathcal{T}^{-1}(\Delta\mathbf{x})) = \frac{1}{\nu(\mathcal{T}^{-1}(\Delta\mathbf{x}))} \int_{\mathbf{x}\in\mathcal{T}^{-1}(\Delta\mathbf{x})} \rho(\mathbf{x})\, d\mathbf{x}.$$

iii) Noise Transport Equation (Theoretical)

The transportation for white Gaussian noise becomes:

$$\widetilde{W}(\Delta\mathbf{x}) = \frac{1}{\nu(\mathcal{T}^{-1}(\Delta\mathbf{x}))} \int_{\mathbf{x}\in\mathcal{T}^{-1}(\Delta\mathbf{x})} W(\mathbf{x})\, d\mathbf{x}$$

However, this naive transportation outputs **are not correct** as the variance is not proportional to the integrated area

$$\sigma^2\left(\widetilde{W}(\Delta\mathbf{x})\right) = \frac{1}{\nu(\mathcal{T}^{-1}(\Delta\mathbf{x}))^2} \int_{\mathbf{x}\in\mathcal{T}^{-1}(\Delta\mathbf{x})} 1^2\, d\mathbf{x} = \frac{1}{\nu(\mathcal{T}^{-1}(\Delta\mathbf{x}))}$$

iii) Noise Transport Equation (Theoretical)

Noise transport equation for variance preservation

$$\widetilde{W}(A) = \int_{\mathbf{x}\in A} \frac{1}{|\nabla\mathcal{T}(\mathcal{T}^{-1}(\mathbf{x}))|^{\frac{1}{2}}} W(\mathcal{T}^{-1}(\mathbf{x}))\, d\mathbf{x}$$

where the **determinant of the Jacobian accounts for the variance change**.

A) Proof of variance preservation $\mathbf{y} = \mathcal{T}^{-1}(\mathbf{x})$, i.e. $|\nabla\mathcal{T}(\mathbf{y})|d\mathbf{y} = d\mathbf{x}$

- ***Ŵ(*A*)* is white noise**

$$\begin{aligned}\widetilde{W}(A) &= \int_{\mathbf{y}\in\mathcal{T}^{-1}(A)} \frac{1}{|\nabla\mathcal{T}(\mathbf{y})|^{\frac{1}{2}}} W(\mathbf{y})|\nabla\mathcal{T}(\mathbf{y})|d\mathbf{y} \\ &= \int_{\mathbf{y}\in\mathcal{T}^{-1}(A)} |\nabla\mathcal{T}(\mathbf{y})|^{\frac{1}{2}} W(\mathbf{y})d\mathbf{y} \\ &= \int_{\mathbf{y}\in\mathcal{T}^{-1}(A)} |\nabla\mathcal{T}(\mathbf{y})|^{\frac{1}{2}} dB(\mathbf{y}).\end{aligned}$$

- ***Ŵ(*A*)* variance**

$$\begin{aligned}\sigma^2 &= \int_{\mathbf{y}\in\mathcal{T}^{-1}(A)} \left(|\nabla\mathcal{T}(\mathbf{y})|^{\frac{1}{2}}\right)^2 d\mathbf{y} \\ &= \int_{\mathbf{y}\in\mathcal{T}^{-1}(A)} |\nabla\mathcal{T}(\mathbf{y})|\, d\mathbf{y} \\ &\boxed{= \int_{\mathbf{x}\in A} d\mathbf{x} = \nu(A),}\end{aligned}$$

iii) Noise Transport Equation (Theoretical)

B) Proof of warping characteristics

$$\frac{\widetilde{W}(\Delta\mathbf{x})}{\nu(\Delta\mathbf{x})^{\frac{1}{2}}} = \frac{W(\mathcal{T}^{-1}(\Delta\mathbf{x}))}{\nu(\mathcal{T}^{-1}(\Delta\mathbf{x}))^{\frac{1}{2}}}$$

$$\widetilde{W}(A) = \int_{\mathbf{x}\in A} \frac{1}{|\nabla\mathcal{T}(\mathcal{T}^{-1}(\mathbf{x}))|^{\frac{1}{2}}} W(\mathcal{T}^{-1}(\mathbf{x}))\, d\mathbf{x}$$

$$\widetilde{W}(\mathcal{T}(\Delta\mathbf{x})) = \int_{\mathbf{y}\in\Delta\mathbf{x}} |\nabla\mathcal{T}(\mathbf{y})|^{\frac{1}{2}} W(\mathbf{y})d\mathbf{y}.$$

Change of variable

$$\widetilde{W}(\mathcal{T}(\Delta\mathbf{x})) \simeq \int_{\mathbf{y}\in\Delta\mathbf{x}} \left(\frac{1}{\nu(\Delta\mathbf{x})}\int_{\mathbf{u}\in\Delta\mathbf{x}} |\nabla\mathcal{T}(\mathbf{u})|d\mathbf{u}\right)^{\frac{1}{2}} W(\mathbf{y})d\mathbf{y}$$

Constant Jacobian as its mean

$$= \left(\frac{1}{\nu(\Delta\mathbf{x})}\int_{\mathbf{u}\in\Delta\mathbf{x}} |\nabla\mathcal{T}(\mathbf{u})|d\mathbf{u}\right)^{\frac{1}{2}} \left(\int_{\mathbf{y}\in\Delta\mathbf{x}} W(\mathbf{y})d\mathbf{y}\right)$$

$$= \left(\frac{1}{\nu(\Delta\mathbf{x})}\int_{\mathbf{v}\in\mathcal{T}(\Delta\mathbf{x})} d\mathbf{v}\right)^{\frac{1}{2}} \left(\int_{\mathbf{y}\in\Delta\mathbf{x}} dB(\mathbf{y})\right)$$

White noise as Brownian

$$= \left(\frac{\nu(\mathcal{T}(\Delta\mathbf{x}))}{\nu(\Delta\mathbf{x})}\right)^{\frac{1}{2}} W(\Delta\mathbf{x}),$$

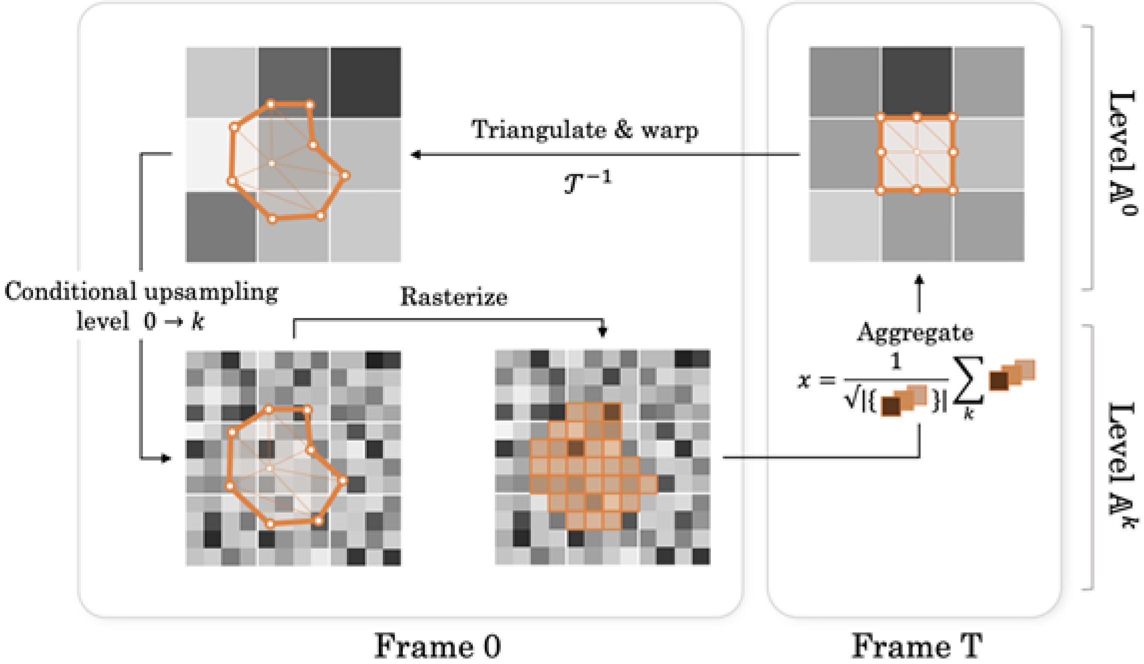
iii) Noise Transport Equation (Practical)

In practice, **the equation is intractable** due to the infinite nature of the white noise.

iii) Noise Transport Equation (Practical)

In practice, **the equation is intractable** due to the infinite nature of the white noise.

Warp a non-empty set of the domain using the inverse of deformation field, fetching values from the original white noise.



iii) Noise Transport Equation (Practical)

Discretization of the transport equation

$$G(\mathbf{p}) = \int_{\mathbf{x}\in A} \frac{1}{|\nabla\mathcal{T}(\mathcal{T}^{-1}(\mathbf{x}))|^{\frac{1}{2}}} W(\mathcal{T}^{-1}(\mathbf{x}))\, d\mathbf{x} \quad \Rightarrow \quad G(\mathbf{p}) = \frac{1}{\sqrt{|\Omega_{\mathbf{p}}|}} \sum_{A_i^k \in \Omega_{\mathbf{p}}} W_k(A_i^k)$$

Approximate using the constant Jacobian at **p**

$$\frac{\widetilde{W}(\Delta\mathbf{x})}{\nu(\Delta\mathbf{x})^{\frac{1}{2}}} = \frac{W(\mathcal{T}^{-1}(\Delta\mathbf{x}))}{\nu(\mathcal{T}^{-1}(\Delta\mathbf{x}))^{\frac{1}{2}}}$$

$$\begin{aligned} G(\mathbf{p}) &\simeq \left(\frac{\nu(A)}{\nu(\mathcal{T}^{-1}(A))}\right)^{\frac{1}{2}} W(\mathcal{T}^{-1}(A)) \\ &= \frac{1}{\nu(\mathcal{T}^{-1}(A))^{\frac{1}{2}}} W(\mathcal{T}^{-1}(A)) \qquad \text{since } \nu(A) = 1 \text{ by definition of } A. \end{aligned}$$

iii) Noise Transport Equation (Practical)

Approximate the warped pixel shape by its rasterized version at level *k*

$$\mathcal{T}^{-1}(A) \simeq \bigcup_{A_i^k \in \Omega_{\mathbf{p}}} A_i^k$$

$$G(\mathbf{p}) \simeq \frac{1}{\nu\left(\bigcup_{\Omega_{\mathbf{p}}} A_i^k\right)^{\frac{1}{2}}} W\left(\bigcup_{\Omega_{\mathbf{p}}} A_i^k\right)$$

Substituting the approximation

$$= \frac{1}{\left(\sum_{A_i^k \in \Omega_{\mathbf{p}}} \nu(A_i^k)\right)^{\frac{1}{2}}} \sum_{A_i^k \in \Omega_{\mathbf{p}}} W(A_i^k)$$

$$= \sqrt{\frac{N_k}{|\Omega_p|}} \sum_{A_i^k \in \Omega_{\mathbf{p}}} W(A_i^k) \qquad \text{since } \nu(A_i^k) = 1/N_k,$$

$$= \frac{1}{\sqrt{|\Omega_p|}} \sum_{A_i^k \in \Omega_{\mathbf{p}}} W_k(A_i^k) \qquad \text{since } W_k(A_i^k) = \sqrt{N_k}\, W(A_i^k),$$

iii) Noise Transport Equation (Practical)

Algorithm 2 Distribution-preserving noise warping (for a single pixel)

Input: G: discrete noise at anchor frame (in size $D \times D$)
A: pixel area in current frame
$\mathcal{T}$: deformation mapping between the anchor and current frame
k: noise upsampling factor
s: polygon subdivision steps

Output: pixel value x in current frame

$G_\infty \leftarrow \text{UPSAMPLE}_\infty(G, k)$
$(V, F) \leftarrow \text{TRIANGULATE\_AREA}(A, s)$
$V \leftarrow \text{WARP}_\infty(V, \mathcal{T})$
$\Omega \leftarrow \text{RASTERIZE}((V, F), G_\infty)$
$x \leftarrow \sum_{(i,j)\in\Omega} G_\infty(i, j)/\sqrt{\text{SIZE}(\Omega)}$

iiii) 1D Toy Experiment

1-D set of *i.i.d.* random variables $\{x_0, x_1, \ldots, x_n\} \sim \mathcal{N}(0,1)$

$$\mathcal{I} = \{0, \ldots, n\}$$

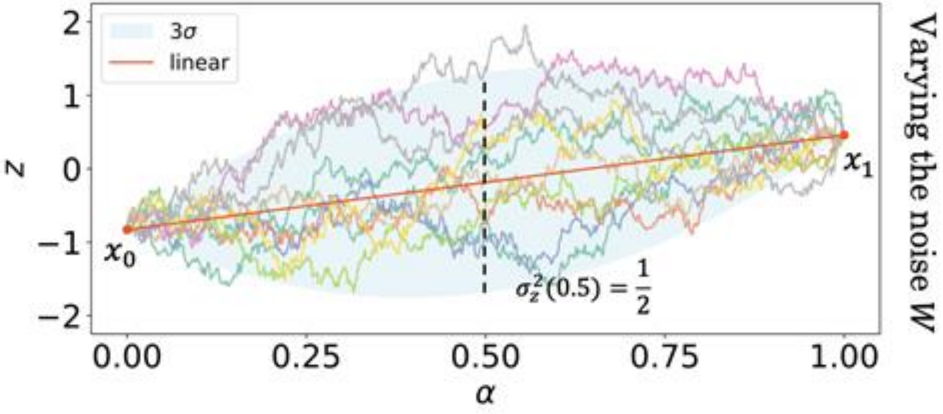
Mapping function: Translation $\mathcal{T}_{1D}^{-1}(i) = i - \alpha$

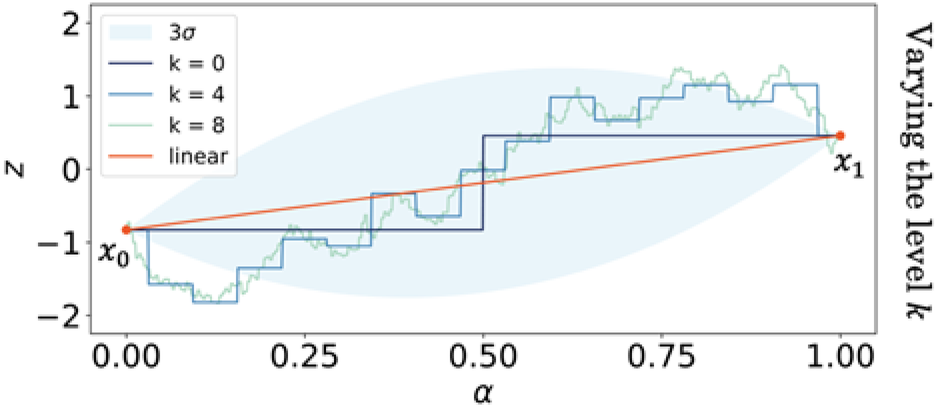
Transported values (Linear interpolation) - **Variance gets reduced.**

$$z_i = \alpha x_{i-1} + (1-\alpha) x_i, \quad z_i \sim \mathcal{N}(0, \sigma_z^2), \quad \text{with } \sigma_z^2 = \alpha^2 + (1-\alpha)^2$$

iiii) 1D Toy Experiment

Linear interpolation now becomes Brownian bridge.
The stochastic component compensates the diminished variance.

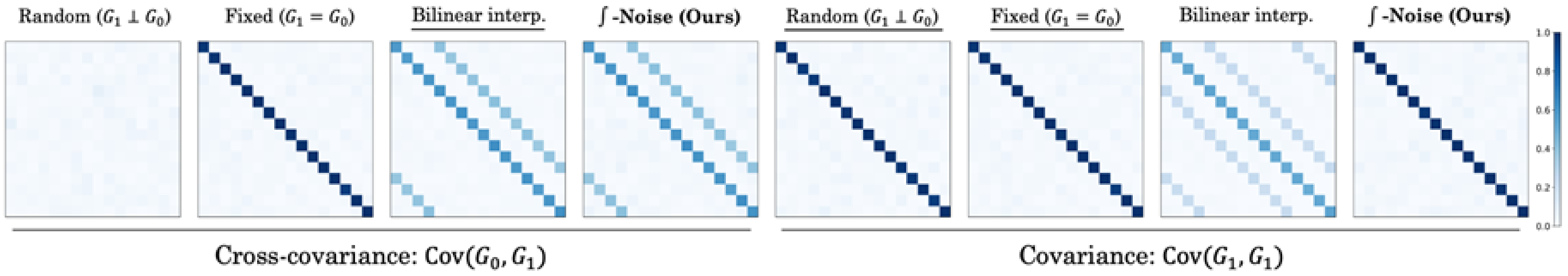




Experiments

Validating Integral Noise Prior

- Maximizing the correct correlation between the warped and the original sample
- Maintaining the independence of pixels within each sample



Random Noise: Different and independent noise samples for each frame
Fixed Noise: Same fixed set of noise samples for all frames

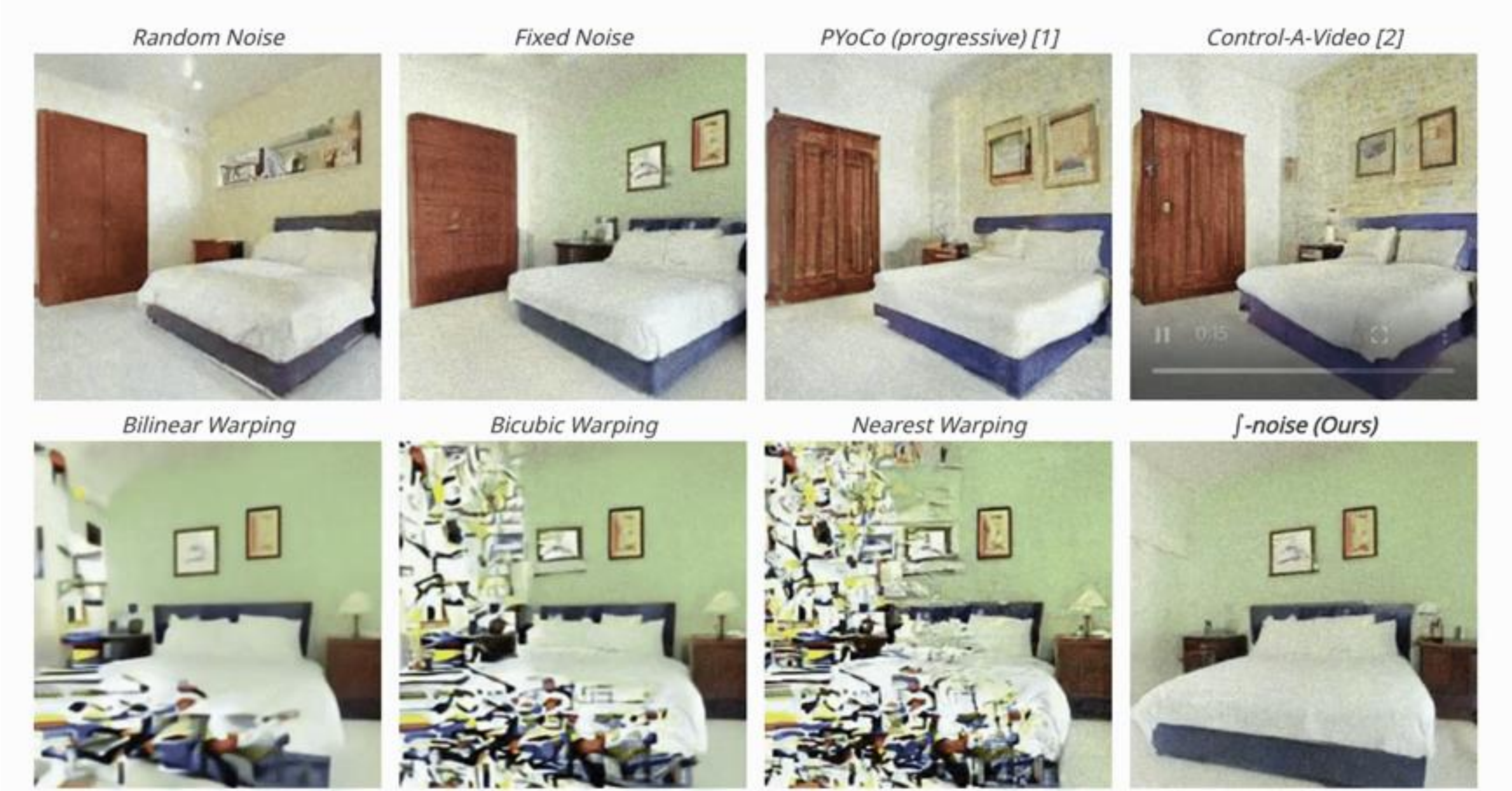
Applications

- Realistic appearance transfer with SDEdit
- Video restoration and super-resolution with I2SB
- Pose-to-person video generation
- Fluid simulation super-resolution

Baselines

- Fixed / Random noise
- Standard interpolation methods
- Control-A-Video: Residual based noise sampling
    - Residual: Noise is resampled at locations where temporal variations of RGB exceed a predefined threshold.
- PYoCo: Mixed / Progressive
    - Mixed: Linearly combines a frame-dependent noise sample and a noise shared across all frames
    - Progressive: The noise at the current frame is generated by perturbing the noise from the previous frame

i) Realistic appearance transfer with SDEdit

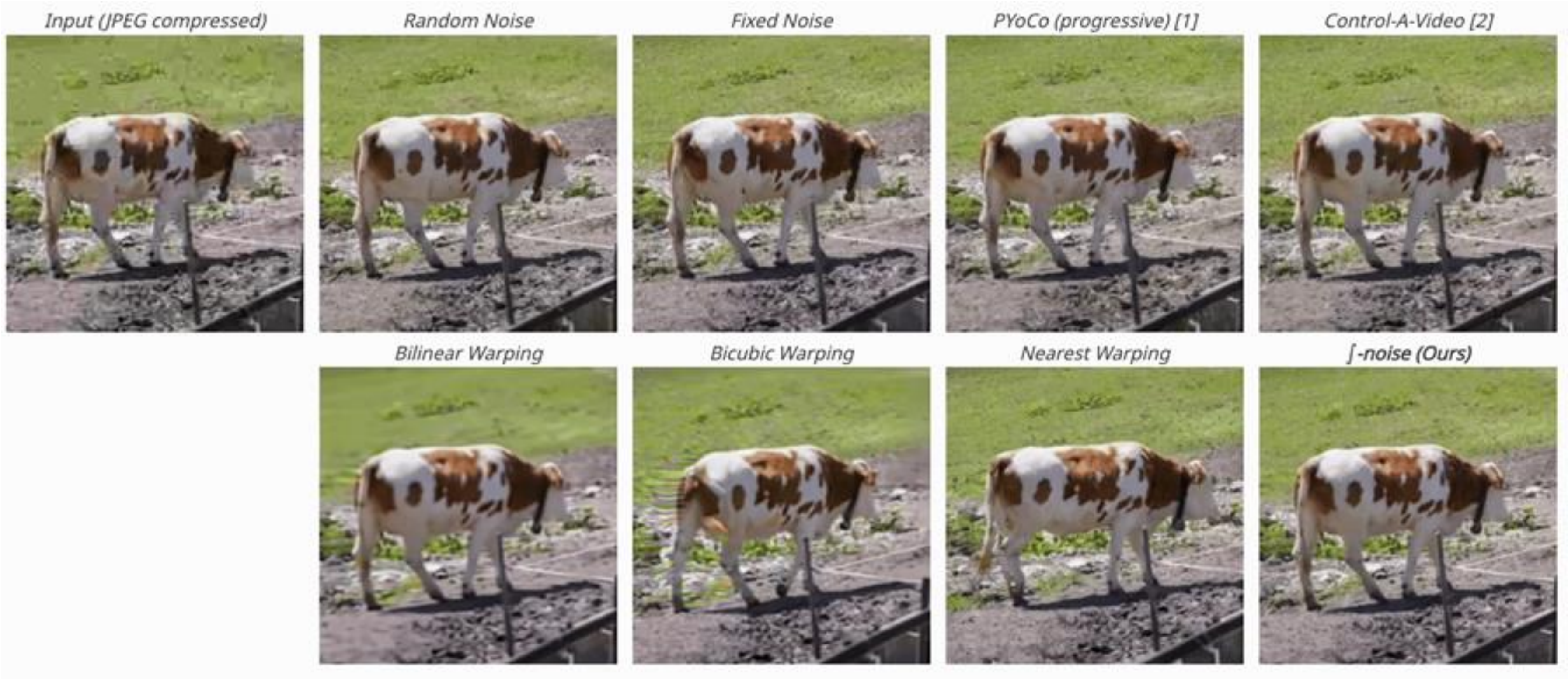


ii) Video restoration and super-resolution with I2SB

4× image super-resolution and JPEG compression restoration

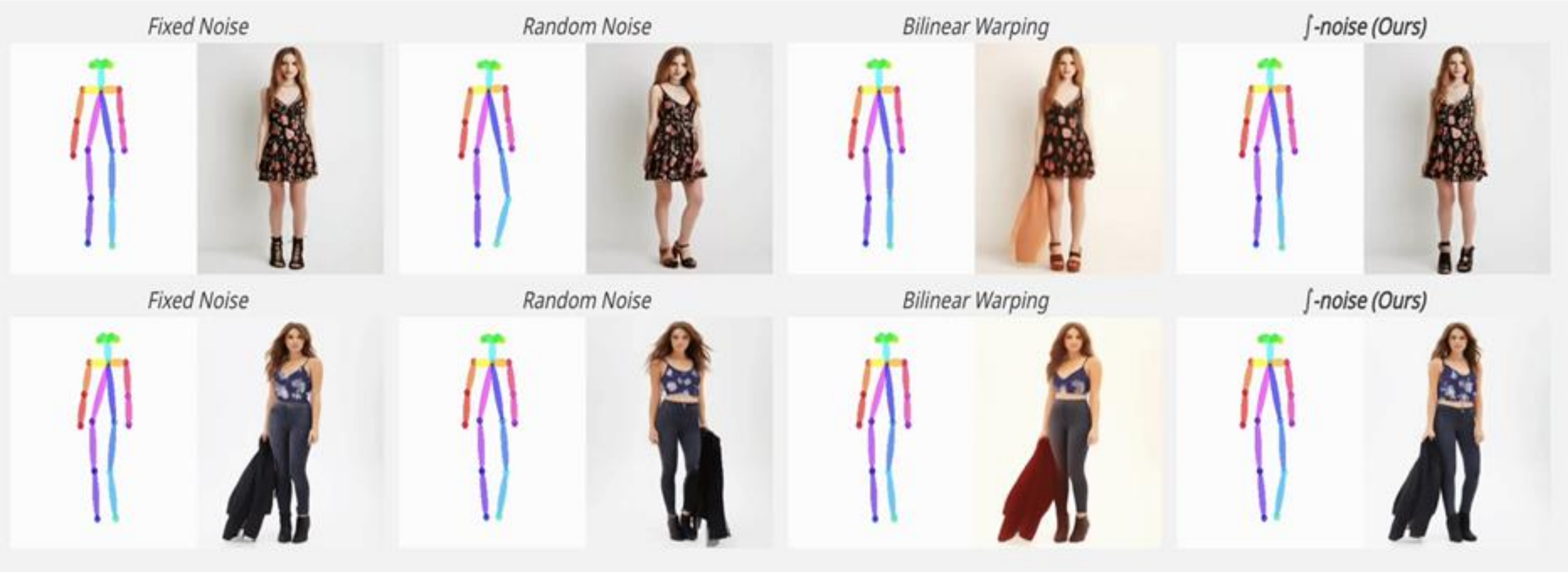
ii) Video restoration and super-resolution with I2SB

4× image super-resolution and **JPEG compression restoration**



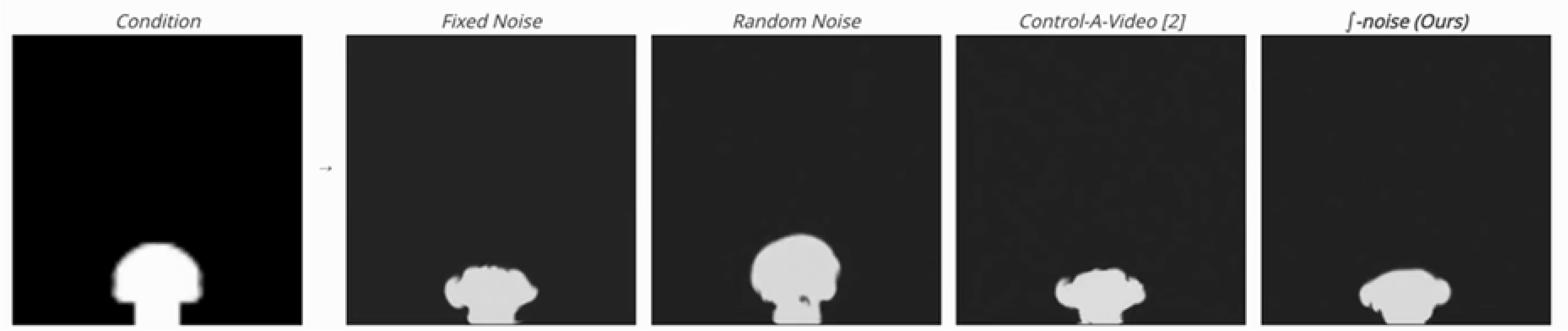
iii) Pose-to-person video

Run a few steps of diffusion with PIDM to get a blurry person outputs.
Extract the full body pose with DensePose and compute the optical flow.



iv) Fluid Simulation Super-Resolution

Unconditional diffusion model trained from scratch using 2D fluid simulation dataset



Quantitative Results

Method	Appearance Transfer			Video SR (4×)		JPEG restore		Pose-to-Person
	warp ($\times 10^{-3}$) ↓	FID ↓	Precision ↑	warp ↓	LPIPS ↓	warp ↓	LPIPS ↓	warp ↓
Random	10.00	74.75	0.719	8.91	0.192	8.28	0.163	34.22
Fixed	4.35	93.56	0.644	7.97	0.179	7.54	0.163	2.26
PYoCo (mixed)	7.26	81.60	0.674	8.48	0.190	7.77	0.166	20.92
PYoCo (prog.)	4.90	84.63	0.667	8.10	0.190	7.48	0.163	11.97
Control-A-Video	5.09	90.82	0.649	7.98	0.192	7.73	0.164	6.45
Bilinear	3.15	143.24	0.201	21.47	0.590	**5.26**	0.431	**2.13**
Bicubic	4.95	149.85	0.212	13.02	0.490	5.74	0.372	2.40
Nearest	15.10	154.73	0.344	14.30	0.329	7.91	0.213	9.21
$\int$-noise (ours)	**2.50**	92.63	0.661	**6.49**	0.196	5.98	0.165	2.92

Quantitative Results - Runtime

Method	Wall Time	CPU Time
Random	0.01	0.01
Fixed	0.01	0.01
PYoCo (mixed)	0.01	0.01
PYoCo (prog.)	0.01	0.01
Control-A-Video	6.08	95.46
Bilinear	5.26	76.76
Bicubic	6.00	87.73
Nearest	5.17	75.73
Root-bilinear	7.66	103.78
DDIM Inv. (20 Steps)	853.42	2226.6
DDIM Inv. (50 Steps)	2125.5	3608.3
$\int$-noise (ours, $k = 3, s = 4$)	629.01	2274.9

Table 3: **Runtime Comparisons of Different Noise Schemes.** The measurements are in milliseconds per frame at resolutions 256×256.

	$s = 1$	$s = 2$	$s = 3$	$s = 4$
$k = 0$	21.6	21.7	23.0	26.2
$k = 1$	23.6	23.5	23.8	25.1
$k = 2$	30.5	29.3	29.6	30.8
$k = 3$	58.8	55.4	55.0	53.7
$k = 4$	143.8	137.5	132.4	128.2

	$s = 1$	$s = 2$	$s = 3$	$s = 4$
$k = 0$	10.5	10.6	11.3	12.9
$k = 1$	10.5	10.6	10.8	11.3
$k = 2$	11.2	11.3	11.5	11.9
$k = 3$	15.5	15.3	15.5	15.6
$k = 4$	29.1	29.2	28.6	28.7

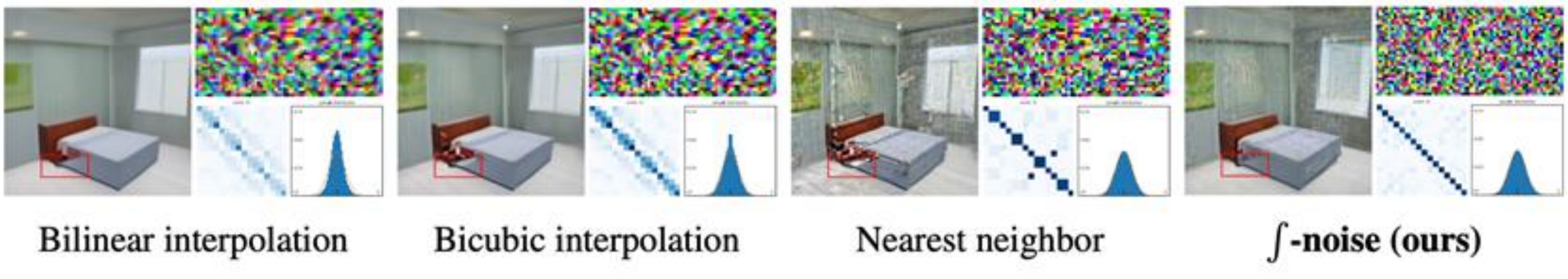
Table 4: **CPU time (top) and wall time (bottom) of $\int$-noise computation for different k and s parameters.** The measurements are for a video sequence of 24 frames at resolution 256×256, in seconds.

Ablation Studies

Standard interpolation methods

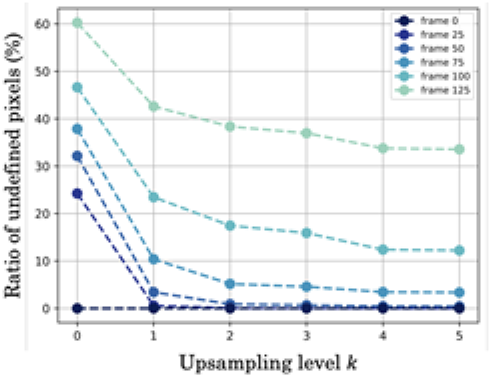
Bilinear / bicubic - Blurry outputs (Variance reduction)

Nearest neighbor - Duplicating artifacts



Noise upscaling factor *k*

Smaller *k* results in more undefined pixels after warping



Suboptimal results with LDMs

Possibly due to the latent space discrepancy

Conclusion & Limitations

- Computationally inefficient than previous noise prior sampling methods

- Temporally-correlated noise prior **DOES NOT GUARANTEE** better temporal coherency
    - **“More constrained models may be oblivious to noise initialization”**

- Extension to latent diffusion models

- Train a diffusion model with temporally consistent noise

“And that, dear reader, is how we warped your noise.”