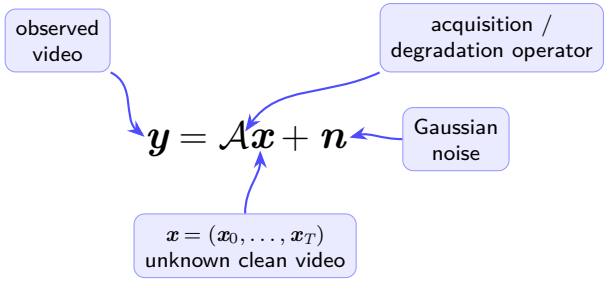
LATINO: LAtent Video consisTency INverse sOlver for High Definition Video Restoration

Alessio Spagnoletti
Andrés Almansa & Marcelo Pereyra

CNRS & Université Paris-Cité (MAP5) & Heriot Watt University

March 28, 2026

Video inverse problems



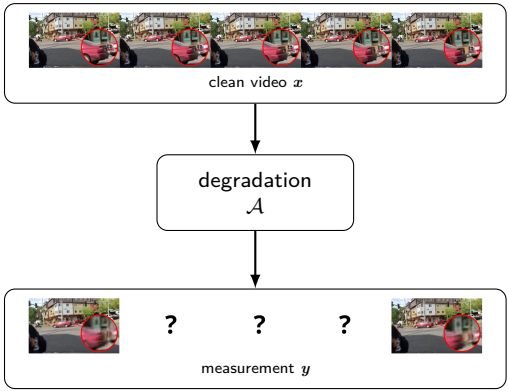
Goal: recover x from incomplete or corrupted measurements y.

Typical examples:

- spatial super-resolution,
- temporal super-resolution / frame interpolation,
- spatio-temporal degradations.

Key difficulty

Video restoration is usually **ill-posed**: there are multiple solutions that agree with the observed data.



Why video is harder than image restoration

A good reconstruction must recover

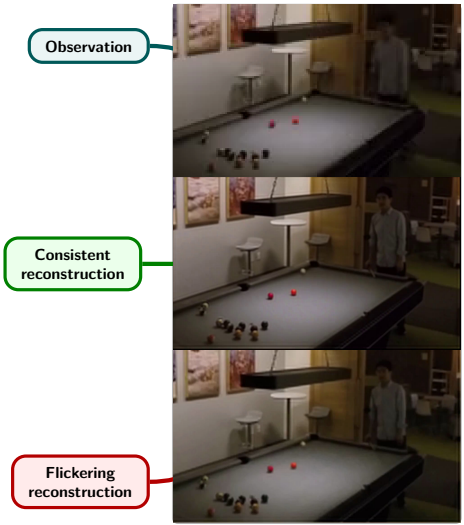
- **fine spatial detail** inside each frame,
- **temporal consistency** across frames,
- **motion-dependent structure** over time.

Restoring frames independently is often not enough:

- individual frames may look sharp,
- but the video can still exhibit **flicker**,
- and motion can become **incoherent**.

Extra challenges

- larger memory footprint,
- heavier computation,
- longer-range dependencies,
- causal / temporal modeling.



A Bayesian view of video restoration

Because the problem is ill-posed, we need prior information about plausible videos.

We model restoration through the posterior

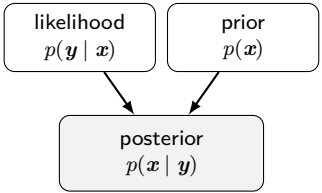
$$\underbrace{p(\boldsymbol{x} \mid \boldsymbol{y})}_{\text{target}} \propto \underbrace{p(\boldsymbol{y} \mid \boldsymbol{x})}_{\text{data fidelity}} \underbrace{p(\boldsymbol{x})}_{\text{video prior}}$$

where

- $p(\boldsymbol{y} \mid \boldsymbol{x})$ is the **likelihood**: does $\boldsymbol{x}$ explain the measurement?
- $p(\boldsymbol{x})$ is the **prior**: does $\boldsymbol{x}$ look like a realistic natural video?

The restoration task becomes a trade-off between

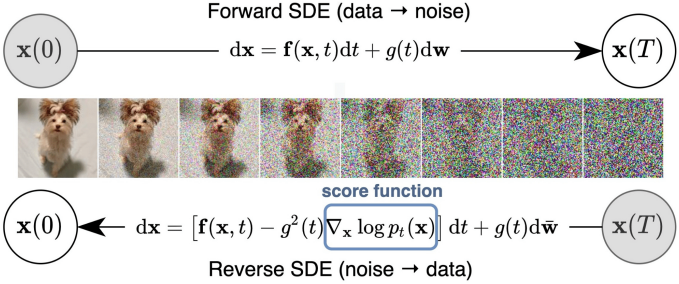
measurement consistency and natural video structure.



From diffusion models to video consistency models

Diffusion models (DMs) [Song et al. 2020; Ho et al. 2020]

- powerful generative priors,
- but often require many sampling steps.



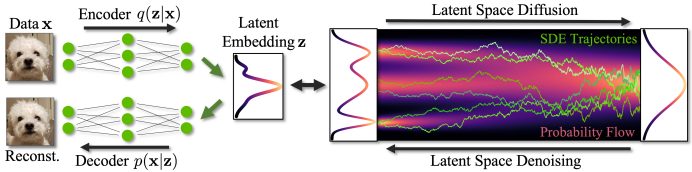
From diffusion models to video consistency models

Diffusion models (DMs) [Song et al. 2020; Ho et al. 2020]

- powerful generative priors,
- but often require many sampling steps.

Latent diffusion models (LDMs) [Rombach et al. 2021]

- operate in VAE latent space,
- reduce memory and compute cost.



From diffusion models to video consistency models

Diffusion models (DMs) [Song et al. 2020; Ho et al. 2020]

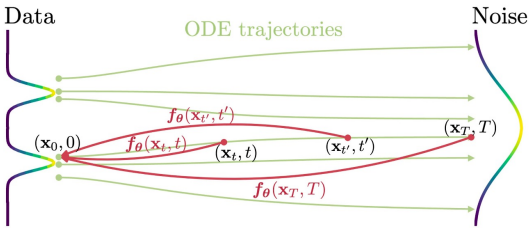
- powerful generative priors,
- but often require many sampling steps.

Latent diffusion models (LDMs) [Rombach et al. 2021]

- operate in VAE latent space,
- reduce memory and compute cost.

Consistency models (CMs) [Song et al. 2023]

- distilled from diffusion models,
- enable few-step inference.



From diffusion models to video consistency models

Diffusion models (DMs) [Song et al. 2020; Ho et al. 2020]

- powerful generative priors,
- but often require many sampling steps.

Latent diffusion models (LDMs) [Rombach et al. 2021]

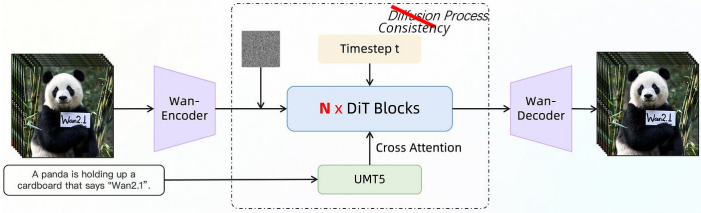
- operate in VAE latent space,
- reduce memory and compute cost.

Consistency models (CMs) [Song et al. 2023]

- distilled from diffusion models,
- enable few-step inference.

Video consistency models (VCMs) [Wang et al. 2023; Yin et al. 2024]

- based on large Video DMs (Wan2.1)
- provide fast video priors for inverse problems.



Where LAVTINO fits

Existing image-based inverse solvers can produce strong per-frame restorations, but may struggle to preserve temporal coherence in high-definition videos.

LAVTINO is designed to address this gap by combining

- a **video consistency prior (VCM)** for temporal structure,
- a **frame-wise image prior (ICM)** for spatial detail,
- and explicit **measurement consistency**.

Main idea

Use **fast generative priors** that are aware of video structure, while keeping inference scalable and **gradient-free**.

Desired properties

- high-resolution detail ← **from ICM**
- temporal smoothness ← **from VCM**
- few NFEs ← **from VCM & ICM**
- low memory usage ← **no backprop**
- plug-and-play use ← **training-free**

How? Product-of-experts Prior

Goal. Approximately sample from the posterior

$$p(\boldsymbol{x} \mid \boldsymbol{y}, c, \lambda) = \frac{p(\boldsymbol{y} \mid \boldsymbol{x})\, p(\boldsymbol{x} \mid c, \lambda)}{\int_{\mathbb{R}^n} p(\boldsymbol{y} \mid \boldsymbol{x})\, p(\boldsymbol{x} \mid c, \lambda)\, \mathrm{d}\boldsymbol{x}},$$

with data $\boldsymbol{y}$, prompt c, and spatiotemporal regularization $\lambda \in \mathbb{R}^3_+$.

Key traits. Zero-shot Langevin sampler for video; leverages *both* Video CMs (VCMs) and Image CMs (ICMs); few NFEs; gradient-free.

Product-of-experts prior.

$$p(\boldsymbol{x} \mid c, \lambda) \;\propto\; p_V^{\eta}(\boldsymbol{x} \mid c)\; p_I^{1-\eta}(\boldsymbol{x} \mid c)\; p_\phi(\boldsymbol{x} \mid \lambda), \quad \eta \in (0, 1).$$

- $p_V(\boldsymbol{x} \mid c)$: text-to-video LCM (encoder–decoder $(\mathcal{E}_V, \mathcal{D}_V)$, consistency f_ϑ^V) for long-range temporal structure.
- $p_I(\boldsymbol{x} \mid c)$: high-res text-to-image LCM (frame-wise $(\mathcal{E}_I, \mathcal{D}_I, f_\theta^I)$) for fine spatial detail.
- $p_\phi(\boldsymbol{x} \mid \lambda) \propto \exp\{-\phi_\lambda(\boldsymbol{x})\}$ with TV_3^λ:

$$\phi_\lambda(\boldsymbol{x}) = \sum_{t,c,i,j} \sqrt{\lambda_h^2 (D_h \boldsymbol{x})^2 + \lambda_v^2 (D_v \boldsymbol{x})^2 + \lambda_t^2 (D_t \boldsymbol{x})^2}\,.$$

LATINO [Spagnoletti et al. 2025]: embed CM priors via stochastic auto-encoders

Reminder: For image restoration LATINO uses an ICM prior to approximately sample from $p(\boldsymbol{x} \mid \boldsymbol{y}, c) \propto p(\boldsymbol{y} \mid \boldsymbol{x})\, p(\boldsymbol{x} \mid c)$.

Approximate Langevin splitting:

$$\underbrace{\boldsymbol{u} = \boldsymbol{x}_k + \int_0^{\delta_k} \nabla \log p(\tilde{\boldsymbol{x}}_s \mid c)\, ds + \sqrt{2}\, d\boldsymbol{w}_s}_{\text{prior half-step}} \quad \Longrightarrow \quad \underbrace{\boldsymbol{x}_{k+1} = \boldsymbol{u} + \delta_k \nabla \log p(\boldsymbol{y} \mid \boldsymbol{x}_{k+1})}_{\text{implicit likelihood step}}$$

or equivalently

$$\boldsymbol{x}_{k+1} = \operatorname{prox}_{\delta_k g_{\boldsymbol{y}}}(\boldsymbol{u}), \qquad g_{\boldsymbol{y}}(\boldsymbol{x}) = -\log p(\boldsymbol{y} \mid \boldsymbol{x}).$$

How do we realize the prior half-step with a consistency model?
Given a latent CM $(\mathcal{E}, \mathcal{D}, f_\theta)$, we replace the intractable prior diffusion by the *stochastic auto-encoder (SAE)* step

$$\boxed{t_k := t(\delta_k) \qquad \boldsymbol{z}_{t_k} = \sqrt{\alpha_{t_k}}\, \mathcal{E}(\boldsymbol{x}_k) + \sqrt{1-\alpha_{t_k}}\, \epsilon, \qquad \boldsymbol{u} = \mathcal{D}\big(f_\theta(\boldsymbol{z}_{t_k}, t_k, c)\big).} \Longrightarrow \boxed{\boldsymbol{u} = \mathrm{SAE}\big(\boldsymbol{x}_k;\, f_\theta, \mathcal{E}, \mathcal{D}, t_k, c\big)}$$

- **Encode + noise:** move $\boldsymbol{x}_k$ to a noisy latent $\boldsymbol{z}_{t_k}$.
- **Consistency decode:** map $\boldsymbol{z}_{t_k}$ back toward the data manifold through f_θ.
- **Role of t_k:** controls the strength of the contraction, depends on the Langevin step size δ_k.

Key idea reused in LVTINO

LVTINO keeps exactly this SAE principle, but applies it *twice per iteration*: first with a **video CM prior** (VCM SAE), then with a **frame-wise image CM prior** (ICM SAE), with proximal likelihood / TV corrections in between.

Split Scheme: VCM prior → prox → ICM prior → prox

Given step size $\delta > 0$ and weight η:

$$\textbf{(1) VCM prior:}\quad \boldsymbol{x}_{k+\frac{1}{4}} \approx \mathrm{SAE}_V\big(\boldsymbol{x}_k;\, f_\vartheta^V, \mathcal{E}_V, \mathcal{D}_V, t_k^{(V)}, c\big)$$

$$\textbf{(2) Likelihood \& TV half-step (implicit):}\quad \boldsymbol{x}_{k+\frac{1}{2}} \approx \mathrm{prox}_{\delta_k\eta\,[\,g_{\boldsymbol{y}}+\phi_\lambda\,]}\big(\boldsymbol{x}_{k+\frac{1}{4}}\big)$$

$$\textbf{(3) ICM prior:}\quad \boldsymbol{x}_{k+\frac{3}{4}} \approx \mathrm{SAE}_I\big(\boldsymbol{x}_{k+\frac{1}{2}};\, f_\theta^I, \mathcal{E}_I, \mathcal{D}_I, t_k^{(I)}, c\big)$$

$$\textbf{(4) Likelihood half-step (implicit):}\quad \boldsymbol{x}_{k+1} \approx \mathrm{prox}_{\delta_k(1-\eta)\,g_{\boldsymbol{y}}}\big(\boldsymbol{x}_{k+\frac{3}{4}}\big)$$

with $g_{\boldsymbol{y}}(\boldsymbol{x}) = -\log p(\boldsymbol{y}|\,\boldsymbol{x})$.

Why this split?

- Prior steps use *exact integration* via stochastic auto-encoding (few NFEs).
- Implicit (backward-Euler) likelihood steps ⇒ *numerically stable for large* δ_k (fast convergence, small bias).
- Prox view ⇒ simple solvers: CG for g_y, proximal splitting / warm-started Adam for $g_{\boldsymbol{y}} + \phi_\lambda$.

LΛTINO pipeline

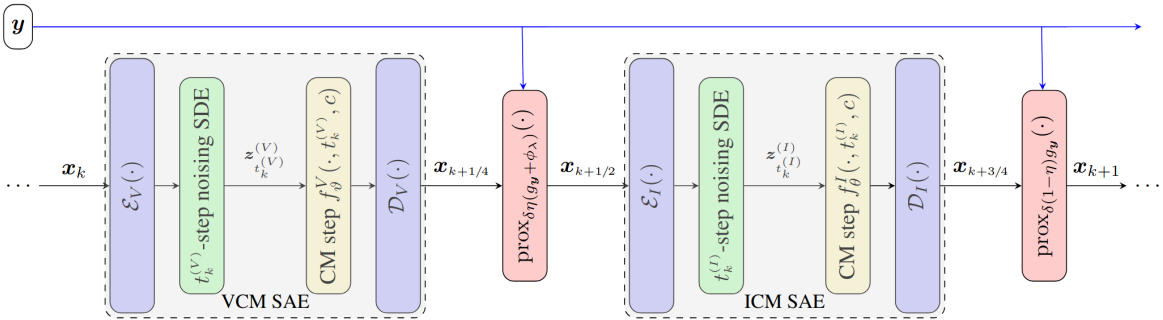
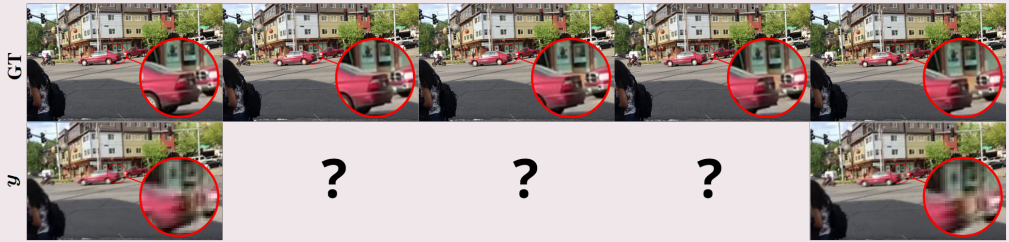


Figure: One step of the LΛTINO solver, a discretization of the Langevin SDE which targets the posterior $p(\boldsymbol{x}|\boldsymbol{y}, c, \lambda)$, involving two stochastic autoencoding (SAE) steps and two proximal steps.

Inverse problems considered

Measurement model: $y = Ax + n, \qquad n \sim \mathcal{N}(0, \sigma_n^2 \mathrm{Id}), \qquad \sigma_n = 0.001.$

Problem A: Temporal SR×4 + Spatial SR×4



$$\mathcal{A} = \underbrace{\text{temp. avg pool}_{\times 4}}_{\text{lower frame rate}} \circ \underbrace{\text{spatial downsample}_{\times 4}}_{\text{lower resolution}}$$

- Recover missing intermediate frames + invert motion blur (temporal avg pool).
- Recover fine spatial details.
- Strong reliance on temporal motion priors.

First setting: moderate temporal and spatial degradation.

Inverse problems considered

Measurement model: $y = Ax + n, \qquad n \sim \mathcal{N}(0, \sigma_n^2 \mathrm{Id}), \qquad \sigma_n = 0.001.$

Problem B: Temporal blur + Spatial SR×8

$$\mathcal{A} = \underbrace{\text{temporal blur}}_{\substack{\text{uniform kernel} \\ \text{size 7}}} \circ \underbrace{\text{spatial downsample}_{\times 8}}_{\text{very low resolution}}$$

- Simulates motion-corrupted observations.
- Very aggressive spatial information loss.
- Requires both deblurring and super-resolution.

Second setting: blur + very strong spatial degradation.

Inverse problems considered

Measurement model: $y = Ax + n, \qquad n \sim \mathcal{N}(0, \sigma_n^2 \mathrm{Id}), \qquad \sigma_n = 0.001.$

Problem C: Temporal SR×8 + Spatial SR×8

$$\mathcal{A} = \underbrace{\text{temp. avg pool}_{\times 8}}_{\text{strong temporal degradation}} \circ \underbrace{\text{spatial downsample}_{\times 8}}_{\text{strong spatial degradation}}$$

- Harder version of Problem A.
- Severe loss in both time and space.
- Particularly challenging for temporal consistency.

Third setting: strongest temporal and spatial degradation.

Problem A: Temporal SR×4 + Spatial SR×4

Problem B: Temporal blur + Spatial SR×8

Problem C: Temporal SR×8 + Spatial SR×8

Thank you for the attention !

References I

- Song, Yang et al. [2020]. "Score-Based Generative Modeling through Stochastic Differential Equations". In: *International Conference on Learning Representations*.
- Ho, Jonathan, Ajay Jain, and Pieter Abbeel [2020]. "Denoising diffusion probabilistic models". In: *Advances in neural information processing systems* 33, pp. 6840–6851.
- Rombach, Robin et al. [2021]. "High-Resolution Image Synthesis with Latent Diffusion Models". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685. URL: https://api.semanticscholar.org/CorpusID:245335280.
- Song, Yang et al. [2023]. "Consistency Models". In: *International Conference on Machine Learning*. URL: https://api.semanticscholar.org/CorpusID:257280191.
- Wang, Xiang et al. [2023]. "VideoLCM: Video Latent Consistency Model". In: *ArXiv* abs/2312.09109. URL: https://api.semanticscholar.org/CorpusID:266209871.
- Yin, Tianwei et al. [2024]. "From Slow Bidirectional to Fast Autoregressive Video Diffusion Models". In: *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974. URL: https://api.semanticscholar.org/CorpusID:274610175.
- Spagnoletti, Alessio et al. [2025]. *LATINO-PRO: LAtent consisTency INverse sOlver with PRompt Optimization*. arXiv: 2503.12615 [cs.CV]. URL: https://arxiv.org/abs/2503.12615.