This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

InstantViR: Real-Time Video Inverse Problem Solver with Distilled Diffusion Prior

Weimin Bai[1], Suzhe Xu[2], Yiwei Ren[1], Jinhua Hao[3], Ming Sun[3], Wenzheng Chen[1], He Sun[1†]
1. Peking University 2. Huaqiao University 3. Kuaishou Technology
https://ai4scientificimaging.org/instantvir

Abstract

*Video inverse problems are fundamental to streaming, telepresence, and AR/VR, where high perceptual quality must coexist with tight latency constraints. Diffusion-based priors currently deliver state-of-the-art reconstructions, but existing approaches either adapt image diffusion models with ad hoc temporal regularizers—leading to temporal artifacts—or rely on native video diffusion models whose iterative posterior sampling is far too slow for real-time use. We introduce **InstantViR**, an amortized inference framework for ultra-fast video reconstruction powered by a pre-trained video diffusion prior. We distill a powerful bidirectional video diffusion model (teacher) into a causal autoregressive student that maps a degraded video directly to its restored version in a single forward pass, inheriting the teacher's strong temporal modeling while completely removing iterative test-time optimization. The distillation is prior-driven: it only requires the teacher diffusion model and known degradation operators, and does not rely on externally paired clean/noisy video data. To further boost throughput, we replace the video-diffusion backbone VAE with a high-efficiency LeanVAE via an innovative teacher-space regularized distillation scheme, enabling low-latency latent-space processing. Across streaming random inpainting, Gaussian deblurring and super-resolution, InstantViR matches or surpasses the reconstruction quality of diffusion-based baselines while running at over **35 FPS** on NVIDIA A100 GPUs, achieving up to 100× speedups over iterative video diffusion solvers. These results show that diffusion-based video reconstruction is compatible with real-time, interactive, editable, streaming scenarios, turning high-quality video restoration into a practical component of modern vision systems.*

1. Introduction

Videos have become the dominant medium for communication, entertainment, and information in the digital age. Consequently, the task of solving video inverse problems—reconstructing a high-quality, clean video $\boldsymbol{x}$ from degraded measurements $\boldsymbol{y}$—is of pivotal importance, with applications ranging from enhancing consumer footage and restoring historical films to improving medical imaging and autonomous driving systems [3, 6, 9, 10, 32, 38]. This reconstruction task is inherently ill-posed, as a single measurement $\boldsymbol{y}$ can correspond to many possible clean videos $\boldsymbol{x}$. To solve this ambiguity, a standard approach is to cast video reconstruction as a Bayesian inverse problem, where the goal is to sample from the posterior distribution $p(\boldsymbol{x}|\boldsymbol{y})$:

$$p(\boldsymbol{x}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}). \tag{1}$$

This formulation elegantly decouples the problem into two components: (1) the likelihood $p(\boldsymbol{y}|\boldsymbol{x})$, governed by the physics of the degradation (e.g., blur, noise, or masking); and (2) the prior $p(\boldsymbol{x})$, which captures the high-dimensional spatiotemporal statistics of natural videos. The central challenge in video inverse problems is therefore how to represent and compute with this video prior $p(\boldsymbol{x})$ in a way that is both expressive and computationally tractable.

Recently, diffusion models [7, 20, 23] have become the dominant tool for learning powerful priors, yielding state-of-the-art results for image inverse problems [1, 5, 11, 14, 19]. A natural first step for videos is to repurpose pre-trained image diffusion models [7, 18] for video restoration. Methods along this line [6, 9, 10, 32] process videos frame-by-frame and enforce consistency only through empirical temporal regularizers (e.g., optical flow constraints or batched noise). While convenient, this image-native strategy has two key drawbacks: (1) it uses priors that ignore rich spatiotemporal dynamics, often leading to flicker and temporal inconsistency; and (2) it still relies on slow, iterative sampling to solve Eq. 1, which is ill-suited for real-time deployment.

A more principled alternative is to adopt native video diffusion models [24, 26], which are trained directly on videos and provide much stronger temporal priors, producing visually coherent sequences. However, when used for inverse problems, these models inherit and amplify the computational burden of diffusion: posterior sampling [5] in

†Corresponding author.

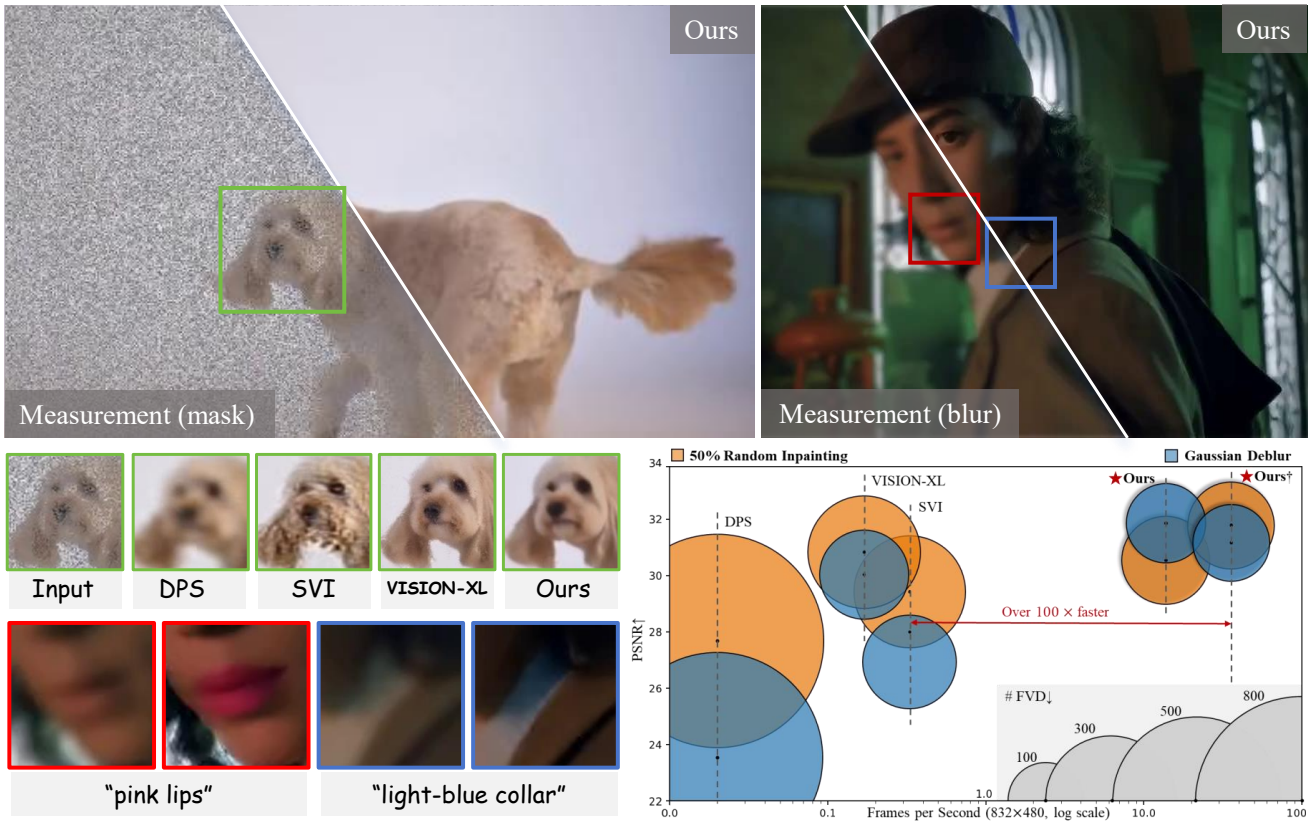


Figure 1. We introduce InstantViR, a real-time video inverse problem solver that drastically outperforms slow sampling-based methods in both speed and quality. **Bottom-right**: At 832×480 resolution, our amortized framework is over **100**× faster than sampling-based baselines like SVI [10], achieving over **35** FPS and the excellent quality. **Left and Bottom-left**: Qualitative examples demonstrate versatile, high-fidelity reconstruction for inpainting and deblurring, along with optional text-guided control (e.g., "pink lips", "light-blue collar").

Eq. 1 requires long iterative trajectories in a much higher-dimensional video space, making inference even more expensive. As a result, current diffusion-based approaches face a stark trade-off: either use image diffusion priors that are temporally weak and slow, or video diffusion priors that are temporally strong but prohibitively slower and incompatible with real-time, streaming applications.

In this work, we argue that this trade-off is not fundamental. We show that one can retain the rich, temporally consistent prior of a strong video diffusion model without inheriting its prohibitive sampling cost. To this end, we introduce InstantViR, a variational framework that recasts video reconstruction as amortized inference via video diffusion distillation [8, 11, 13, 28, 30, 31, 35]. Instead of solving a separate, slow optimization for $p(\boldsymbol{x}|\boldsymbol{y})$ at test time, InstantViR learns a universal one-step solver that maps a degraded video $\boldsymbol{y}$ directly to its reconstruction $\boldsymbol{x}$.

Our method uses an asymmetric teacher–student design. A slow bidirectional video diffusion model (e.g., Wan2.1 [24]) serves as the teacher, which together with the measurement operator (blur, masking, downsampling, text-guided edits, etc.) defines the target posterior. A fast causal autoregressive student is then trained to approximate this posterior in a single forward pass. Training is ground-truth-free: conditioned on the degraded measurement $\boldsymbol{y}$, we align the reconstruction with the teacher's prior while enforcing physics-based data fidelity, yielding a self-supervised approximation to the video posterior. As a result, the same framework seamlessly adapts to streaming random inpainting, Gaussian deblurring, super-resolution, and language-guided restoration/editing by simply changing the measurement model and the conditioning prompt of the video diffusion prior.

To make the system truly real-time, we also tackle the often-overlooked VAE bottleneck. We introduce a teacher-space regularized latent adaptation scheme that allows us to replace the original video VAE with LeanVAE [4]—an ultra-efficient spatiotemporal tokenizer with a lightweight Neighborhood-Aware Feedforward backbone and wavelet-based channel compression—while keeping the distilled solver consistent with the teacher prior. Combined with our one-step causal solver, this design allows InstantViR to deliver diffusion-level reconstructions with up to 100× speedup over iterative video diffusion priors, while maintaining high fidelity and temporal coherence.

Concretely, InstantViR contributes: (i) a ground-truth-free variational distillation framework that converts native video diffusion priors into one-step amortized solvers for diverse video inverse problems; (ii) a streaming causal inverse architecture with block-wise attention and KV caching for low-latency, high-fidelity reconstruction; and (iii) a teacher-space regularized LeanVAE integration that overcomes latent distribution shift and, to our knowledge, yields the first real-time, interactive pipeline for streaming video reconstruction with diffusion-level quality at over 35 FPS.

2. Related Works

2.1. Video Diffusion Models as Priors

Diffusion models [7, 20, 23] have emerged as the state-of-the-art to serve as powerful priors $p(\boldsymbol{x})$ that can accurately model the complex statistics of natural videos. These models define the prior $p(\boldsymbol{x})$ through a learned reverse-time stochastic differential equation (SDE) that denoises a simple Gaussian distribution $p(\boldsymbol{x}_T)$ back to the data distribution $p(\boldsymbol{x}_0)$. Concretely, for a clean video $\boldsymbol{x}_0$, a fixed forward process defines how noise is added:

$$d\boldsymbol{x}_t = f(t)\boldsymbol{x}_t\,dt + g(t)\,d\mathbf{w}_t, \tag{2}$$

where $t \in [0, T]$, $f(t)$ and $g(t)$ are drift and diffusion coefficients, and $\mathbf{w}_t$ is a standard Wiener process. The generative prior $p(\boldsymbol{x})$ is defined by the corresponding reverse-time SDE:

$$d\boldsymbol{x}_t = \left[f(t)\boldsymbol{x}_t - g(t)^2\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t)\right]dt + g(t)\,d\bar{\mathbf{w}}_t, \tag{3}$$

where the key component, the score function $\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t)$, is approximated by a neural network $s_\theta(\boldsymbol{x}_t, t)$. This network s_θ is trained by optimizing the denoising score matching (DSM) objective [22] with $\boldsymbol{x}_t = \alpha_t\boldsymbol{x}_0 + \sigma_t\boldsymbol{\epsilon}$:

$$\mathcal{L}_{\text{DSM}} = \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{\epsilon}}\Big[\lambda(t)\left\|s_\theta(\boldsymbol{x}_t, t) + \tfrac{\boldsymbol{\epsilon}}{\sigma_t}\right\|^2\Big]. \tag{4}$$

Recently, powerful video generative priors based on Diffusion Transformers (DiTs) [16], including Open-Sora [34], Wan2.1 [24, 26], MovieGen [17] have demonstrated unprecedented ability to capture complex spatiotemporal dynamics, paving the way toward world simulators [2]. These models achieve exceptional temporal coherence by employing bidirectional attention across all video frames. However, this architectural strength is also a critical bottleneck: the generation of a single frame requires processing the entire sequence, including future frames. This non-causal dependency introduces prohibitive latency and makes them fundamentally incompatible with real-time or streaming applications.

2.2. Diffusion Models for Inverse Problems

Using the diffusion priors defined above to solve inverse problems (Eq. 1) requires combining the pre-trained prior score s_θ with the likelihood $p(\boldsymbol{y}|\boldsymbol{x})$. While this Bayesian framework has achieved great success in static image inverse problems, the field is now moving towards the significantly more challenging video domain, which introduces the critical demand for spatiotemporal consistency.

Diffusion Models for Image Inverse Problems. In the image domain, sampling-based methods [5, 19, 21] work by injecting an approximation of the likelihood gradient, $\nabla_{\mathbf{x}}\log p(\boldsymbol{y}|\boldsymbol{x})$, into the reverse sampling process (Eq. 3). This is often achieved by approximating the intractable likelihood term $\nabla_{\mathbf{x}_t}\log p(y|\mathbf{x}_t)$ using the posterior mean $\hat{\mathbf{x}}_0$. Particle-based optimization methods [14, 37] avoid this approximation via variational inference. While effective for producing high-fidelity images, these methods are inherently slow, requiring hundreds or thousands of function evaluations to generate a single result.

Diffusion Models for Video Inverse Problems. Extending this framework to video inverse problems introduces a significant new challenge: maintaining temporal consistency. The field has primiarily focused on adapting the image-based solvers by adding temporal heuristics. For example, Warped Diffusion [6] uses optical flow to warp noise, SVI [10] and Vision-XL [9] treat time as a batch dimension, while STEP [32] tuned image diffusion model to capture temporal consistency. These methods are constrained by a weak image-native prior that struggles with complex dynamics and remain prohibitively slow. To address this speed bottleneck, one-step distillation methods [27, 36] have been proposed for specific tasks like super-resolution. However, they are highly task-specific and relies on massive, fully-paired $(\boldsymbol{x}, \boldsymbol{y})$ datasets (e.g., 10M video pairs), limiting its generality and scalability.

This leaves a clear research gap: existing methods force a trade-off between weak-but-slow image priors, or fast-but-task-specific models that require massive paired data. This motivates our work, InstantViR, which introduces an amortized paradigm to distill powerful, general-purpose video priors (e.g., Wan2.1 [24]) into a one-step, feed-forward solver that operates without paired data, enabling real-time, high-fidelity reconstruction for the first time.

3. Method

We now describe InstantViR, which transforms a slow, iterative video diffusion prior into a fast, single-step solver for video inverse problems. In Sec. 3.1, we introduce our variational amortized inference formulation and derive an objective that enables training without paired data. In Sec. 3.2, we present our causal video inverse solver architecture, which enables efficient block-wise streaming inference. In Sec. 3.3, we show how to replace the original video VAE with an ultra-efficient LeanVAE while carefully handling the induced la-

tent distribution shift, which is crucial for achieving true real-time performance.

3.1. Amortized Inference Framework

Modern video diffusion models typically operate in a latent space. A video $\boldsymbol{x}$ is first encoded into latent codes $\boldsymbol{z} = \mathcal{E}(\boldsymbol{x})$ by a video VAE, and then a Diffusion Transformer (DiT) backbone [16] learns a generative prior $p(\boldsymbol{z})$ over these latents. At sampling time, frames are decoded via $\boldsymbol{x} = \mathcal{D}(\boldsymbol{z})$, yielding high-fidelity, temporally coherent videos. Our goal is to leverage such a pre-trained video diffusion model as a prior to solve video inverse problems. Given degraded measurements $\boldsymbol{y}$ related to the clean video $\boldsymbol{x}$ via a known forward operator $\boldsymbol{y} = \mathcal{A}(\boldsymbol{x})$, we wish to sample from the posterior $p(\boldsymbol{x}|\boldsymbol{y})$ defined in Eq. 1.

Traditional sampling-based approaches [6, 9, 10] inject the likelihood $p(\boldsymbol{y}|\boldsymbol{x})$ (defined by $\mathcal{A}$) into the diffusion process and iteratively perturb the generation trajectory to enforce measurement consistency. This yields high quality but is extremely slow, and repeatedly backpropagating through the decoder $\mathcal{D}$ adds further overhead.

Our amortized inference approach, instead, directly learns a single-step feed-forward solver, $q_\phi(\boldsymbol{x}|\boldsymbol{y})$, as a variational approximation to the posterior $p(\boldsymbol{x}|\boldsymbol{y})$. We minimize the expected KL divergence

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{y}\sim p(\boldsymbol{y})}\Big[D_{\mathrm{KL}}(q_\phi(\boldsymbol{x}|\boldsymbol{y}) \,\|\, p(\boldsymbol{x}|\boldsymbol{y}))\Big], \tag{5}$$

which, up to an additive constant, can be decomposed as

$$\mathbb{E}_{\boldsymbol{y}}\Big\{\mathbb{E}_{\boldsymbol{x}\sim q_\phi(\boldsymbol{x}|\boldsymbol{y})}\Big[-\log p(\boldsymbol{y}|\boldsymbol{x})\Big] + D_{\mathrm{KL}}(q_\phi(\boldsymbol{x}|\boldsymbol{y}) \,\|\, p(\boldsymbol{x}))\Big\}, \tag{6}$$

where the first term enforces data fidelity through the known degradation model, and the second term regularizes the solver to remain on the natural video manifold defined by the diffusion prior $p(\boldsymbol{x})$ (e.g., Wan2.1 [24]).

Because the diffusion prior is defined in the latent space, we work with $\boldsymbol{z}$ instead of $\boldsymbol{x}$, where $\boldsymbol{x} = \mathcal{D}(\boldsymbol{z})$ and $\boldsymbol{z} \sim p(\boldsymbol{z})$. We parameterize the solver as $q_\phi(\boldsymbol{z}|\boldsymbol{y})$ and obtain the latent-space objective

$$\mathbb{E}_{\boldsymbol{y}}\big\{\underbrace{\mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{y})}\left[-\log p\big(\boldsymbol{y}|\mathcal{D}(\boldsymbol{z})\big)\right]}_{\mathcal{L}_{\text{likelihood}}} + \underbrace{D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{y}) \,\|\, p(\boldsymbol{z}))}_{\mathcal{L}_{\text{prior}}}\big\}. \tag{7}$$

The prior term $\mathcal{L}_{\text{prior}}$ involves a KL divergence to the diffusion prior $p(\boldsymbol{z})$, which is only implicitly defined via its score function $s_\theta(\boldsymbol{z}_t, t)$. Following score-distillation methods [11, 13, 30], we approximate this KL by a score-matching loss:

$$\mathcal{L}_{\text{prior}} \approx \mathbb{E}_{t,\boldsymbol{\epsilon},\boldsymbol{z}\sim q_\phi(\cdot|\boldsymbol{y})}\big[w(t)\,\|s_\theta(\boldsymbol{z}_t, t) - s_{q_\phi}(\boldsymbol{z}_t, t)\|^2\big], \tag{8}$$

where $\boldsymbol{z}_t = \alpha_t \boldsymbol{z} + \sigma_t \boldsymbol{\epsilon}$ is the noised latent at time t, $w(t)$ is a weighting schedule, and s_{q_ϕ} denotes the score implied by the solver (implemented via a small auxiliary head s_φ).

Optimizing the combined objective in Eq. 7 trains $q_\phi(\boldsymbol{z}|\boldsymbol{y})$ to provide an amortized approximation of the true posterior $p(\boldsymbol{z}|\boldsymbol{y})$, and thus $p(\boldsymbol{x}|\boldsymbol{y})$ via $\mathcal{D}$. This training requires only (i) the degraded measurements $\boldsymbol{y}$ and (ii) the frozen pre-trained diffusion model s_θ. No paired $(\boldsymbol{x}, \boldsymbol{y})$ ground-truth dataset is needed, making the framework naturally scalable and flexibly conditioned on arbitrary measurement operators.

3.2. Causal Autoregressive Solver

A key distinction between offline and streaming video processing is causality. In offline generation, the model can attend to the entire video sequence, whereas in streaming reconstruction it only has access to past and current frames at time step n, not future ones. As a result, a standard DiT with full spatiotemporal attention cannot be directly used as a streaming video inverse solver.

Following this constraint, we design q_ϕ as a causal autoregressive solver that operates on temporal blocks of T frames (Fig. 2), inspired by [31]. Each block of degraded frames $\boldsymbol{y}_n$ is encoded to latents, and the model processes these blocks sequentially using a block-causal attention mechanism with two modes:

- **Intra-block bidirectional attention.** Within the current block n, all tokens attend to each other to model rich local spatiotemporal structure. For a query $\mathbf{Q}_i$ from token i in block n, with keys $\mathbf{K}_n$ and values $\mathbf{V}_n$ from all tokens in that block,

$$\mathrm{Att}_{\text{intra}}(\mathbf{Q}_i, \mathbf{K}_n, \mathbf{V}_n) = \mathrm{softmax}\left(\frac{\mathbf{Q}_i \mathbf{K}_n^\top}{\sqrt{d_k}}\right)\mathbf{V}_n. \tag{9}$$

- **Inter-block causal attention.** Across blocks, tokens in block n are allowed to attend only to previous reconstructed blocks $\hat{\boldsymbol{z}}_{<n}$. For the same query $\mathbf{Q}_i$, with keys $\mathbf{K}_{<n}$ and values $\mathbf{V}_{<n}$ from all past blocks,

$$\mathrm{Att}_{\text{inter}}(\mathbf{Q}_i, \mathbf{K}_{<n}, \mathbf{V}_{<n}) = \mathrm{softmax}\left(\frac{\mathbf{Q}_i \mathbf{K}_{<n}^\top}{\sqrt{d_k}}\right)\mathbf{V}_{<n}, \tag{10}$$

In practice, we implement the inter-block attention with standard autoregressive KV caching: after reconstructing each block, we store its attention keys and values and reuse them as $\mathbf{K}_{<n}$ and $\mathbf{V}_{<n}$ for subsequent blocks. This avoids redundant computation over past frames while preserving exact causal behavior and keeping the per-frame cost low.

3.3. Latent-Space Adaptation with Compact VAE

The combination of variational diffusion distillation and our causal architecture already yields a strong solver, achieving about 15 FPS at 832×480 resolution with the original video VAE. At this point, however, the DiT backbone is no longer the main bottleneck—the heavy video decoder becomes the limiting factor for real-time performance, consistent with prior observations [27, 36] in latent video diffusion models.

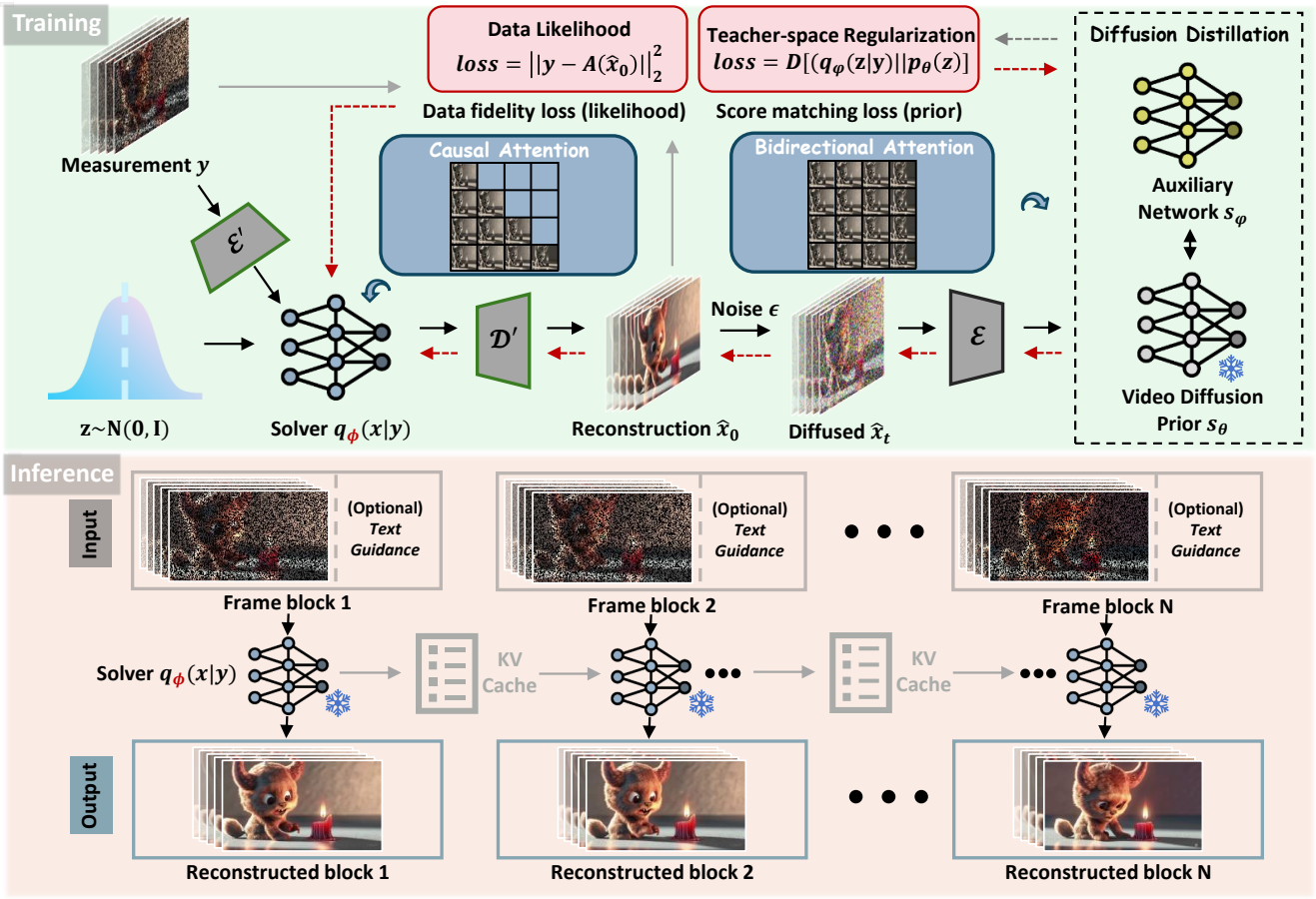


Figure 2. **Overview of the InstantViR framework. (Top) Training:** We train a single-step solver q_ϕ using only degraded measurements $\boldsymbol{y}$. The solver is optimized with two objectives: a data fidelity loss (ensuring the reconstruction matches the measurement $\boldsymbol{y}$) and a prior distillation loss (using a frozen video diffusion prior [24], s_θ, to enforce temporal consistency and realism). **(Bottom) Inference:** The trained solver q_ϕ operates as a fast, feed-forward network, processing the video in a causal, block-wise, and autoregressive manner. This enables real-time, single-step streaming reconstruction with optional text guidance.

A natural attempt to accelerate the system would replace the original VAE with a more efficient architecture. However, this breaks the semantics of the latent space: the diffusion prior $p(\boldsymbol{z})$ and score s_θ are trained in the original latent space $\boldsymbol{z}$ induced by $(\mathcal{E}, \mathcal{D})$, while a new VAE $(\mathcal{E}', \mathcal{D}')$ defines a different latent space $\boldsymbol{z}'$. Directly distilling a solver in $\boldsymbol{z}'$ without accounting for this distribution shift causes severe mismatch with the teacher prior.

To address this, we propose a teacher-space regularized distillation scheme that explicitly bridges the two latent spaces. We train a video inverse solver $q'_\phi(\boldsymbol{z}'|\boldsymbol{y})$ in the $\boldsymbol{z}'$-space using the objective

$$\mathcal{L}(q'_\phi) = \mathbb{E}_{\boldsymbol{y}}\mathbb{E}_{\boldsymbol{z}'\sim q'_\phi(\boldsymbol{z}'|\boldsymbol{y})}\big[-\log p\big(\boldsymbol{y}|\mathcal{D}'(\boldsymbol{z}')\big)\big] + \mathbb{E}_{t,\boldsymbol{\epsilon},\,\boldsymbol{z}'\sim q'_\phi(\cdot|\boldsymbol{y})}\Big[w(t)\,\big\|s_\theta(\boldsymbol{z}_t,t) - s_{q'}(\boldsymbol{z}_t,t)\big\|^2\Big], \tag{11}$$

where we map the new latent $\boldsymbol{z}'$ back into the *teacher* latent space via

$$\boldsymbol{x} = \mathcal{D}'(\boldsymbol{z}'), \quad \boldsymbol{z} = \mathcal{E}(\boldsymbol{x}), \quad \boldsymbol{z}_t = \alpha_t \boldsymbol{z} + \sigma_t \boldsymbol{\epsilon}.$$

Thus, the likelihood term is evaluated in the new latent space, while the prior term uses the teacher diffusion score s_θ in the original latent space $\boldsymbol{z}$ after passing through $\mathcal{D}'$ and $\mathcal{E}$. This constrains the new latent space $\boldsymbol{z}'$ so that, when decoded and re-encoded, it remains aligned with the teacher prior, enabling effective one-step distillation under a new VAE.

After exploring several efficient VAEs, we adopt LeanVAE [4] as our new video tokenizer. LeanVAE is an ultra-efficient spatiotemporal VAE built on a lightweight NAF backbone with wavelet-based channel compression. Plugging LeanVAE into our teacher-space regularized distillation scheme yields an additional $> 2\times$ speedup over the original VAE, pushing InstantViR beyond 35 FPS while preserving diffusion-level fidelity and temporal coherence for streaming video inverse problems.

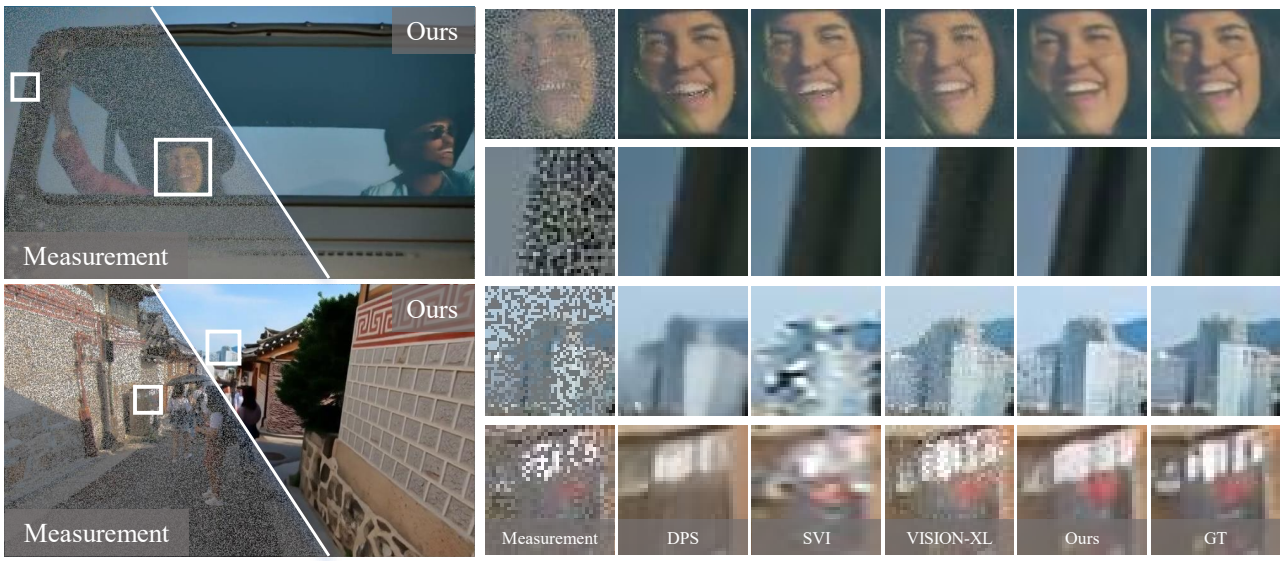


Figure 3. **Qualitative comparison for video random inpainting. (Top)** On the Open-Sora dataset [12], our model reconstructs a high-fidelity and temporally consistent video from a 50% masked measurement. The zoomed-in face sequence demonstrates the stability and fine detail of our single-step result. **(Bottom)** We demonstrate strong zero-shot generalization on the REDS dataset [33], our method generates a sharp, coherent video that is perceptually close to the ground truth.

4. Experiment

We evaluate InstantViR on a range of video inverse problems, compare it against strong diffusion-based baselines, analyze the contribution of its key components (amortized solver, causal architecture, LeanVAE), and demonstrate extended capabilities such as text-guided video reconstruction and editing. Further implementation details and additional results are provided in the supplementary material.

4.1. Experimental Setup

Datasets and Base Model. Our primary training dataset is Open-Sora-v1.1 [12]. We use 6,000 video clips for training, without using any associated text labels. Following our methodology in Sec. 3.1, we do not require paired $(\mathbf{x}, \mathbf{y})$ data: degraded measurements $\mathbf{y}$ are generated on-the-fly from clean videos $\mathbf{x}$ via the known forward operators, and only $\mathbf{y}$ is only used to query the teacher diffusion model. For evaluation, we use a held-out set of 500 Open-Sora videos and the standard REDS30 dataset [15] to test zero-shot generalization. All videos are processed at 832×480 resolution. Our video diffusion prior is the Wan2.1-1.3B model [24]. Training takes two weeks on 8 NVIDIA A100.

Baselines. We compare InstantViR against a suite of strong diffusion-based video inverse problem solvers: DPS [5] as a foundational diffusion posterior sampler; SVI [10] and VISION-XL [9], which adapt image diffusion priors with batch-consistent sampling; and DiffIR2VR [29], which uses hierarchical latent warping for video restoration. We use official implementations for all baselines to ensure high-quality generation and fair comparisons.

Table 1. **Quantitative results of temporal quality and inference speed.** The former is evaluated with FVD ↓ and the latter is with FPS ↑. Best results are in **bold**, suboptimal are underlined.

Method	FVD ↓			Avg. FPS ↑
	Inpainting	Super-Res.	Deblur	
DPS	375.81	711.61	783.10	<0.02
DiffIR2VR	-	311.61	-	0.12
SVI	219.90	176.60	154.38	0.29
VISION-XL	224.74	172.79	138.79	<0.17
InstantViR (Ours)	136.06	**153.13**	110.51	13.91
InstantViR† (Ours)	**132.59**	156.43	**103.45**	**35.56**

Metrics. We evaluate performance using four complementary criteria: (1) Per-frame reconstruction quality: PSNR and SSIM; (2) Perceptual quality: LPIPS [33]; (3) Temporal consistency: Fréchet Video Distance (FVD) [25], which measures the distributional similarity and temporal coherence across frames; (4) Inference speed: frames per second (FPS) on a single NVIDIA A800 80GB GPU.

4.2. Standard Video Inverse Problems

We first evaluate InstantViR on three canonical video inverse problems: random inpainting, Gaussian deblurring, and $4\times$ super-resolution. Table 2 summarizes per-frame reconstruction and perceptual quality, while Table 1 reports temporal consistency and runtime.

As shown in Table 2, our one-step solver **InstantViR** (using the original VAE) and its accelerated variant **InstantViR**† (with LeanVAE) achieve state-of-the-art or highly competitive performance across all tasks, surpass-



Figure 4. **Results for text-guided video reconstruction.** InstantViR not only restores high-fidelity details from severely masked measurements (top row) but also uses text prompts to control the reconstruction. **(Left)** Given a masked input, our method can add specific details like a *"black headband"* or *"wear glasses"* as guided by the text. **(Right)** The framework can generate multi-modal, semantically distinct, and plausible outputs from the same masked input, such as reconstructing the subject with either *"close eyes"* or *"open eyes"*, demonstrating fine-grained controllability.

Table 2. **Quantitative results of spatial quality.** Metrics include PSNR, SSIM, LPIPS. Best results are in **bold**, suboptimal are underlined.

Method	50% Random Inpainting			4× Super-Resolution			Gaussian Deblur		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
DPS	27.68	0.92	0.32	22.78	0.91	0.46	23.54	0.88	0.46
DiffIR2VR	-	-	-	33.44	0.92	0.33	-	-	-
SVI	29.42	0.90	0.17	33.85	0.96	**0.17**	26.93	0.89	0.31
VISION-XL	30.83	0.95	0.25	**35.69**	**0.98**	0.24	30.03	0.93	0.28
InstantViR (Ours)	30.54	**0.97**	0.12	34.91	0.96	0.23	**31.85**	**0.97**	0.17
InstantViR† (Ours)	**31.78**	0.96	**0.13**	27.04	0.95	0.22	31.16	0.97	**0.15**

ing iterative methods such as SVI and VISION-XL in PSNR, SSIM, and LPIPS. This confirms that the amortized student solver successfully inherits the high-fidelity prior of the teacher video diffusion model.

Table 1 and Fig. 1 (Bottom-Right) highlight the substantial inference speedup achieved by InstantViR and InstantViR†. Sampling-based baselines [5, 9, 10] operate at less than 1 FPS and thus are impractical for streaming. In contrast, InstantViR achieves superior temporal consistency (lowest FVD) on all tasks while running at $\sim$ 14 FPS. By replacing the original VAE with LeanVAE via our teacher-space regularized distillation scheme (Sec. 3.3), InstantViR† reaches over **35 FPS**, achieving up to over **100**× speedup over SVI while further improving temporal coherence.

Random Inpainting. Fig. 3 shows qualitative results for 50% random inpainting. On the Open-Sora test set, InstantViR recovers fine, high-frequency details (e.g., individual teeth) that are blurred or missing in the baselines. Moreover, the reconstructions remain temporally stable across frames. On the unseen REDS [15] dataset, our method still produces sharp, coherent distant city views, whereas baseline methods often exhibit blur and temporal jitter, demonstrating strong zero-shot generalization.

Gaussian Deblurring. For Gaussian deblurring (Fig. 5), sampling-based methods such as SVI [10] tend to suffer from noticeable temporal flicker, leading to unstable videos. InstantViR, by contrast, produces sharp and consistent reconstruction across frames, restoring fine details, such as the athlete's earrings, without sacrificing temporal coherence.

4× Super-Resolution. In the 4× super-resolution setting (Fig. 6), InstantViR remains highly competitive with or superior to SVI and VISION-XL. While these baselines are spatially stable, they tend to produce unrealistic jagged edges or

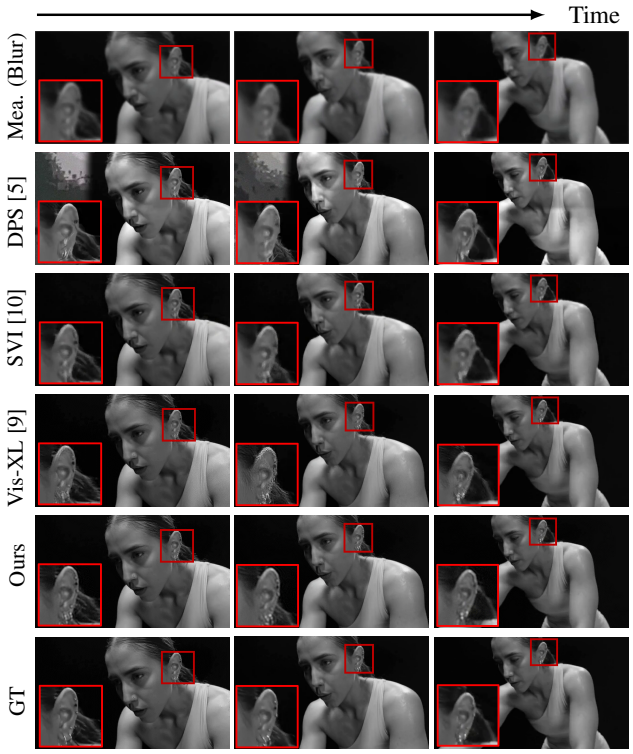


Figure 5. Qualitative comparison for video Gaussian deblurring.

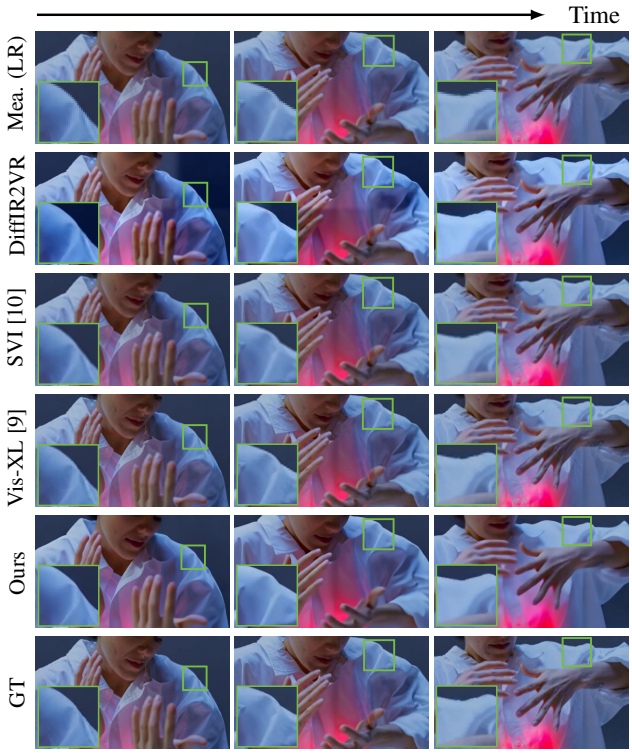


Figure 6. Qualitative comparison for video super-resolution (4x).

overly smoothed details. Our method recovers sharp textures that are both perceptually detailed and temporally aligned with the ground truth.

4.3. Text-Guided Video Reconstruction

Because our framework builds on a text-conditional video diffusion prior (Wan2.1), InstantViR and InstantViR† can naturally generalize to text-guided video reconstruction and editing by conditioning the amortized inference objective (Eq. 7) on prompts. As illustrated in Fig. 1 and 4, this enables fine-grained semantic control over the reconstructed content.

Given a partially masked input video and a text prompt, the model generates plausible, semantically aligned completions in the missing regions. For the same input sequence, different prompts (e.g., *"wearing glasses"* vs. *"wearing a headband"*) produce distinct but temporally consistent edits of the subject. In another example, we can control whether a musician *"closes eyes"* or *"keeps eyes open"* while playing the piano. These results demonstrate that InstantViR not only reconstructs measurement-consistent content (likelihood term), but can also hallucinate coherent new details guided by text (prior term).

This unique combination of high speed, causal streaming, and prompt-level controllability opens up new possibilities for real-time interactive applications, such as live broadcast enhancement, streaming video restoration, and on-the-fly creative editing.

5. Conclusion

In this paper, we presented **InstantViR**, a novel amortized inference framework that resolves the core trade-off between reconstruction quality and speed in video inverse problems. By distilling a powerful bidirectional video diffusion prior into a lightweight, one-step causal autoregressive solver, InstantViR achieves both robust temporal consistency and real-time performance. Our framework eliminates the need for paired ground truth, replaces the heavy video VAE with a lightweight alternative via teacher-space regularization, and bypasses the slow, iterative posterior sampler. We demonstrate these gains on various streaming video tasks, attaining state-of-the-art temporal quality while surpassing 35 FPS.

Limitations and Future Directions. Despite these advances, our accelerated model with latent-space adaptation still slightly lags behind the distilled model using the original VAE in reconstruction quality. This suggests that latent distribution shift remains a limiting factor; jointly fine-tuning the lightweight VAE to better align its latent space with the teacher's original space may further close this gap.

Beyond natural video restoration, the label-free, amortized nature of our framework makes it well suited for scaling to larger unlabeled video corpora and for deployment in other real-time domains such as medical video enhancement. We view InstantViR as a step toward practical, diffusion-level video reconstruction that is both fast and controllable.

Acknowledgement

This work was supported by the National Key Research and Development Program of China (2024YFC3406400), the Shanghai Municipal Science and Technology Major Project (2025SHZDZX026D03) and the National Natural Science Foundation of China (32450631, 62371007). Additionally, we acknowledge the support from the Biomedical Computing Platform of National Biomedical Imaging Center, Peking University.

References

[1] Weimin Bai, Siyi Chen, Wenzheng Chen, and He Sun. Blind inversion using latent diffusion priors. *arXiv preprint arXiv:2407.01027*, 2024.

[2] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

[3] Pascal Chang, Jingwei Tang, Markus Gross, and Vinicius C Azevedo. How i warped your noise: a temporally-correlated noise prior for diffusion models. *arXiv preprint arXiv:2504.03072*, 2025.

[4] Yu Cheng and Fajie Yuan. Leanvae: An ultra-efficient reconstruction vae for video diffusion models. *arXiv preprint arXiv:2503.14325*, 2025.

[5] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

[6] Giannis Daras, Weili Nie, Karsten Kreis, Alex Dimakis, Morteza Mardani, Nikola Borislavov Kovachki, and Arash Vahdat. Warped diffusion: Solving video inverse problems with image diffusion models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020.

[8] Zemin Huang, Zhengyang Geng, Weijian Luo, and Guojun Qi. Flow generator matching. *arXiv preprint arXiv:2410.19310*, 2024.

[9] Taesung Kwon and Jong Chul Ye. Vision-xl: High definition video inverse problem solver using latent image diffusion models, 2024.

[10] Taesung Kwon and Jong Chul Ye. Solving video inverse problems using image diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[11] Sojin Lee, Dogyun Park, Inho Kong, and Hyunwoo J Kim. Diffusion prior-based amortized variational inference for noisy inverse problems. In *European Conference on Computer Vision*, pages 288–304. Springer, 2024.

[12] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024.

[13] Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. *Advances in Neural Information Processing Systems*, 36:76525–76546, 2023.

[14] Morteza Mardani, Jiaming Song, Jan Kautz, and Arash Vahdat. A variational perspective on solving inverse problems with diffusion models. *arXiv preprint arXiv:2305.04391*, 2023.

[15] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[16] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023.

[17] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

[18] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[19] Litu Rout, Negin Raoof, Giannis Daras, Constantine Caramanis, Alex Dimakis, and Sanjay Shakkottai. Solving linear inverse problems provably via posterior sampling with latent diffusion models. *Advances in Neural Information Processing Systems*, 36:49960–49990, 2023.

[20] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015.

[21] Bowen Song, Soo Min Kwon, Zecheng Zhang, Xinyu Hu, Qing Qu, and Liyue Shen. Solving inverse problems with latent diffusion models via hard data consistency. *arXiv preprint arXiv:2307.08123*, 2023.

[22] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[23] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[24] Wan Team. Wan: Open and advanced large-scale video generative models. 2025.

[25] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019.

[26] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang,

Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[27] Jianyi Wang, Shanchuan Lin, Zhijie Lin, Yuxi Ren, Meng Wei, Zongsheng Yue, Shangchen Zhou, Hao Chen, Yang Zhao, Ceyuan Yang, et al. Seedvr2: One-step video restoration via diffusion adversarial post-training. *arXiv preprint arXiv:2506.05301*, 2025.

[28] Yifei Wang, Weimin Bai, Weijian Luo, Wenzheng Chen, and He Sun. Integrating amortized inference with diffusion models for learning clean distribution from corrupted images. *arXiv preprint arXiv:2407.11162*, 2024.

[29] Chang-Han Yeh, Chin-Yang Lin, Zhixiang Wang, Chi-Wei Hsiao, Ting-Hsuan Chen, Hau-Shiang Shiu, and Yu-Lun Liu. Diffir2vr-zero: Zero-shot video restoration with diffusion-based image restoration models. *arXiv preprint arXiv:2407.01519*, 2024.

[30] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6613–6623, 2024.

[31] Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. *arXiv preprint arXiv:2412.07772*, 2, 2024.

[32] Bingliang Zhang, Zihui Wu, Berthy T Feng, Yang Song, Yisong Yue, and Katherine L Bouman. Step: A general and scalable framework for solving video inverse problems with spatiotemporal diffusion priors. *arXiv preprint arXiv:2504.07549*, 2025.

[33] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[34] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

[35] Mingyuan Zhou, Huangjie Zheng, Zhendong Wang, Mingzhang Yin, and Hai Huang. Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In *Forty-first International Conference on Machine Learning*, 2024.

[36] Junhao Zhuang, Shi Guo, Xin Cai, Xiaohui Li, Yihao Liu, Chun Yuan, and Tianfan Xue. Flashvsr: Towards real-time diffusion-based streaming video super-resolution. *arXiv preprint arXiv:2510.12747*, 2025.

[37] Nicolas Zilberstein, Morteza Mardani, and Santiago Segarra. Repulsive latent score distillation for solving inverse problems. *arXiv preprint arXiv:2406.16683*, 2024.

[38] Zihao Zou, Jiaming Liu, Shirin Shoushtari, Yubo Wang, and Ulugbek S Kamilov. Flair: A conditional diffusion framework with applications to face video restoration. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5228–5238. IEEE, 2025.