Real-Time Video Inverse Problem Solver with Distilled Diffusion Prior

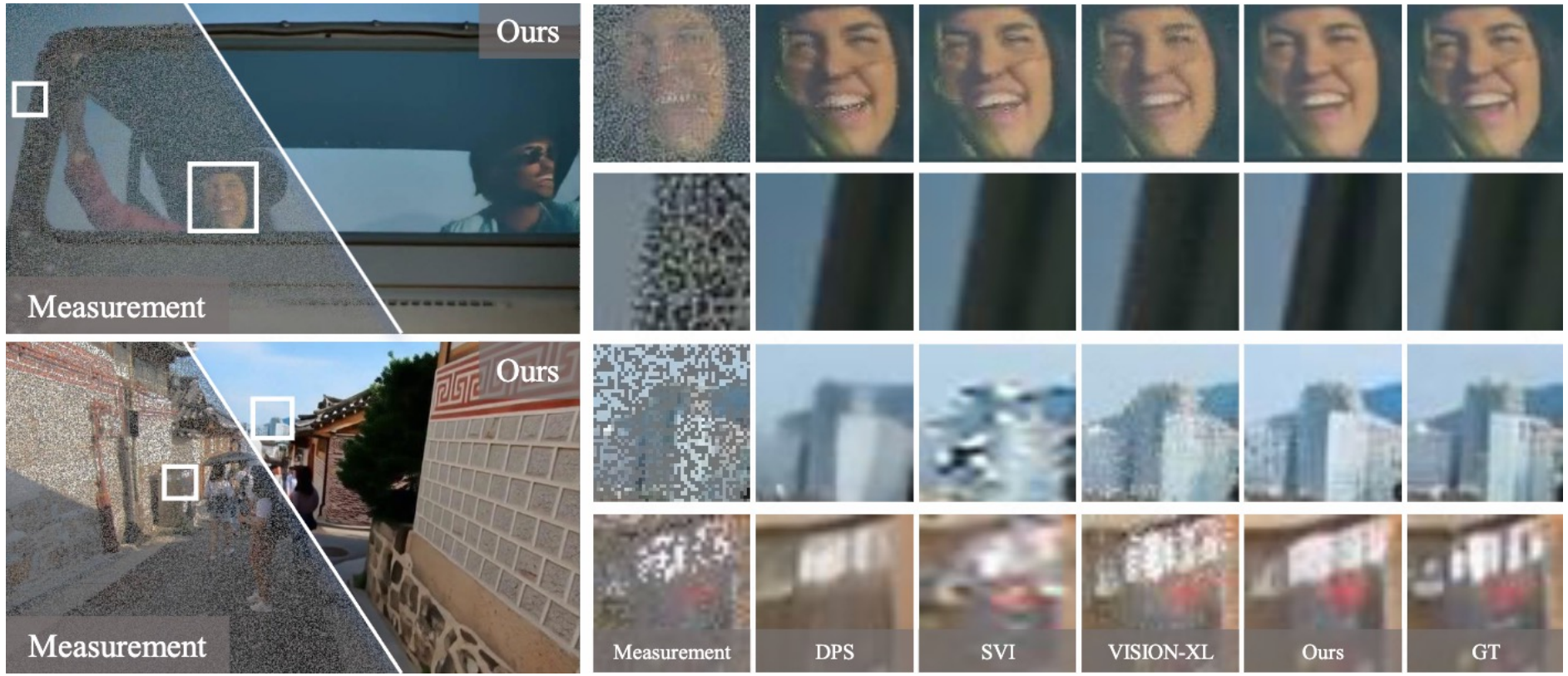
Weimin Bai[1], Suzhe Xu[2], Yiwei Ren[1], Jinhua Hao[3], Ming Sun[3], Wenzheng Chen[1], He Sun[1]

1.Peking University 2.Huaqiao University 3.Kuaishou Technology

Background: Video Inverse Problem

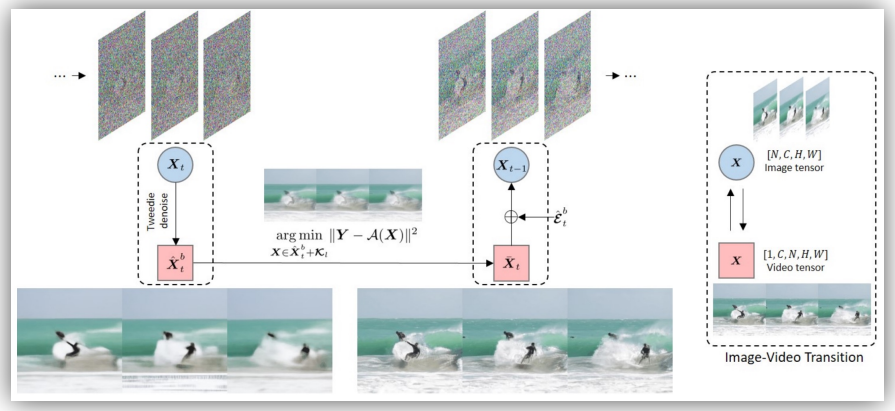
北京大学 PEKING UNIVERSITY

- Reconstructing **clean video** from **degraded input** is a critical and widespread challenge



Challenge 1: Per-Frame Quality
Challenge 2: Temporal Consistency
Challenge 3: Time Consumption

Related Works



2022-2024

Image Inverse Problem – Image Diffusion

DPS, ICLR 2023 (Spotlight)

RED-Diff, ICLR 2024

2024-2025

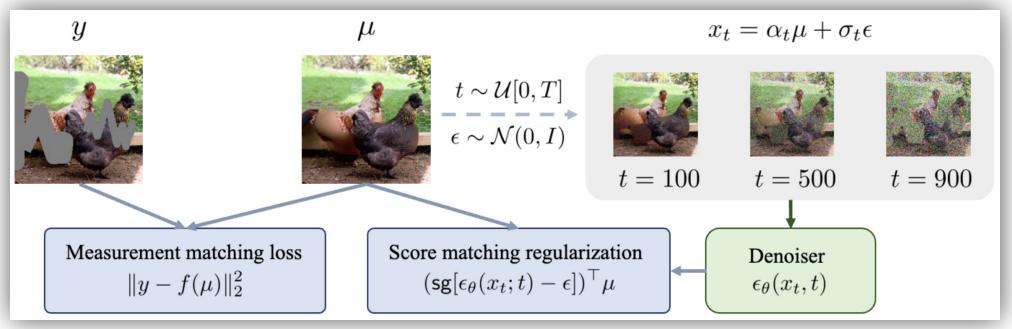
Video Inverse Problem – Image Diffusion

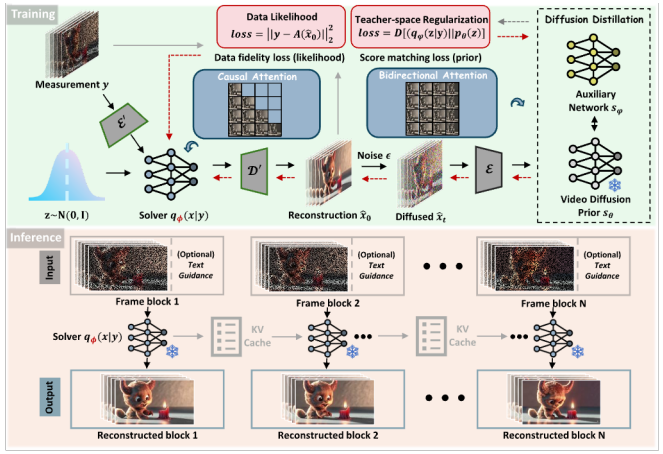
SVI & Vision-XL

STEP

Video Inverse Problem – Video Diffusion

InstantViR(Ours)





Challenge 1: Per-Frame Quality

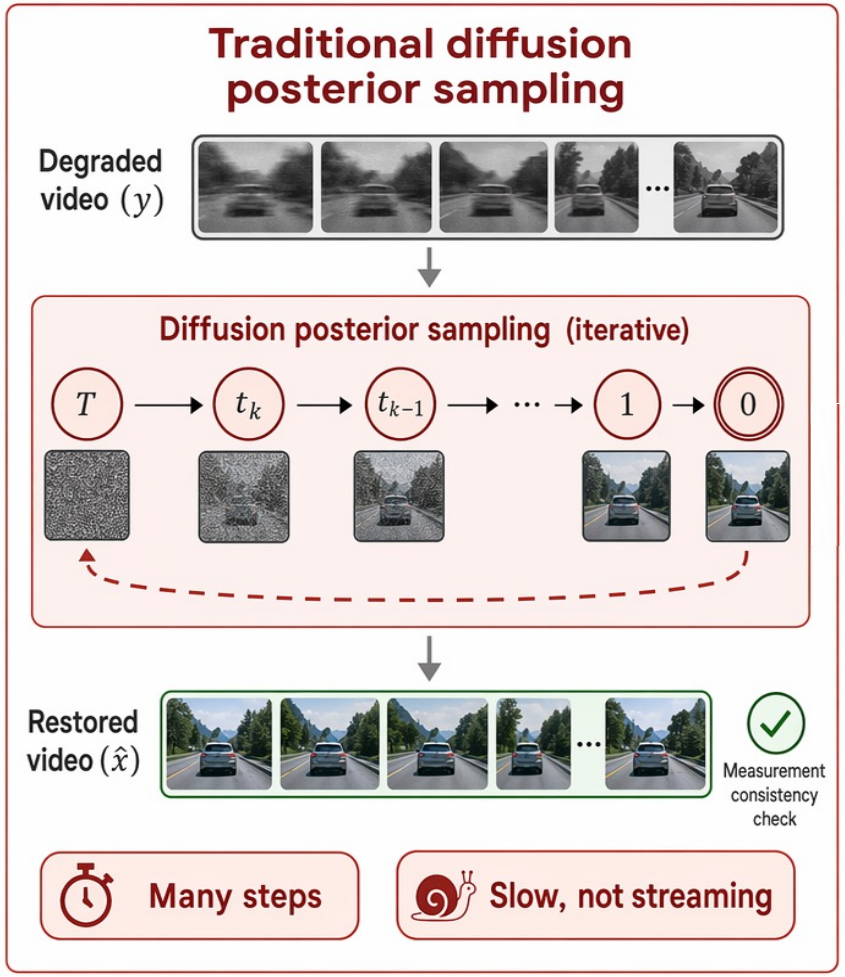
Challenge 2: Temporal Consistency

Challenge 3: Time Consumption

Method

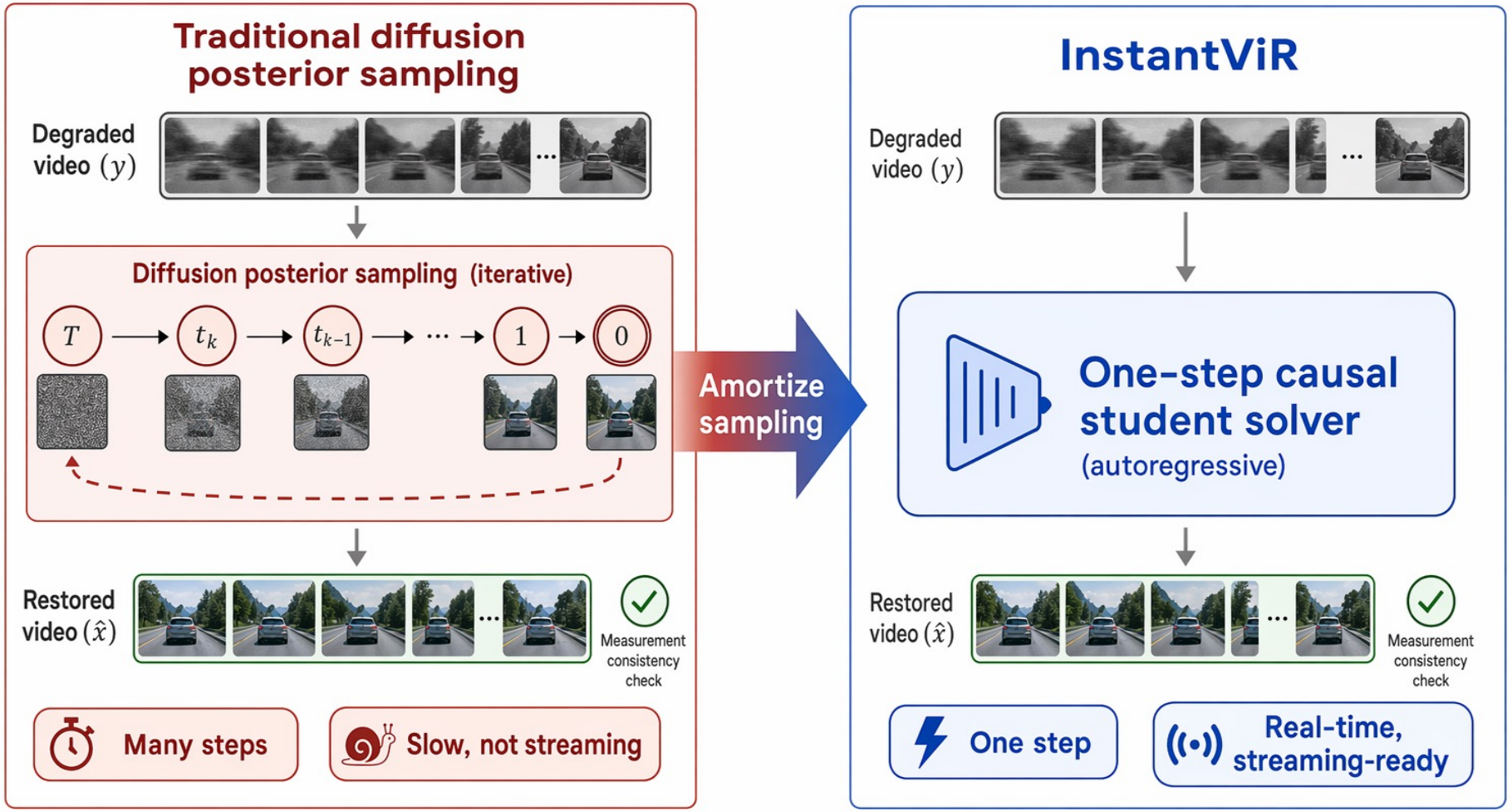
北京大学 PEKING UNIVERSITY

Amortize posterior sampling into one step



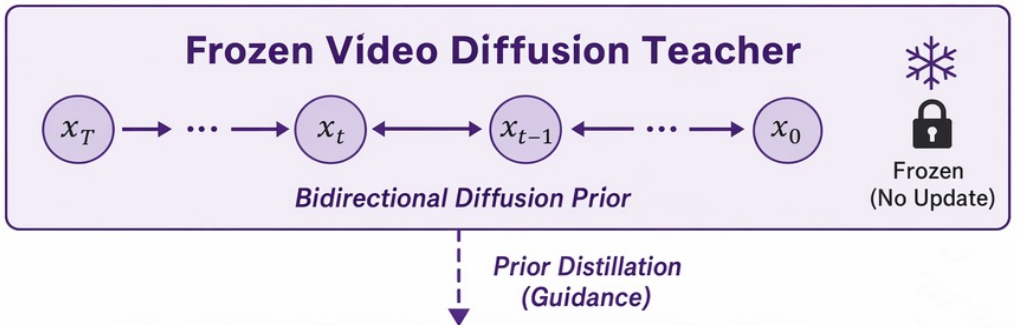
Method

Amortize posterior sampling into one step



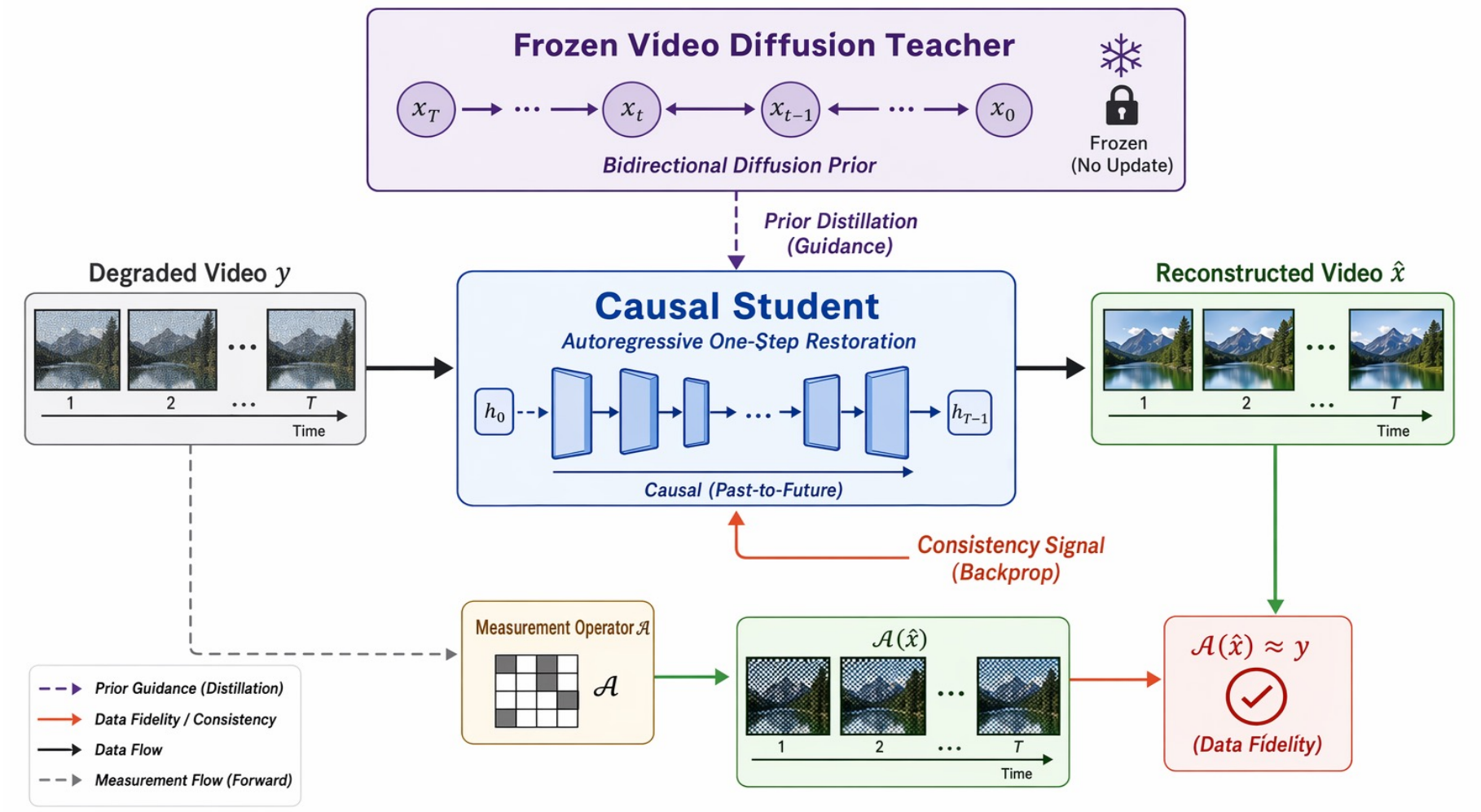
Method

Training: Distill a Frozen Video Diffusion Prior into a Causal Student



Method

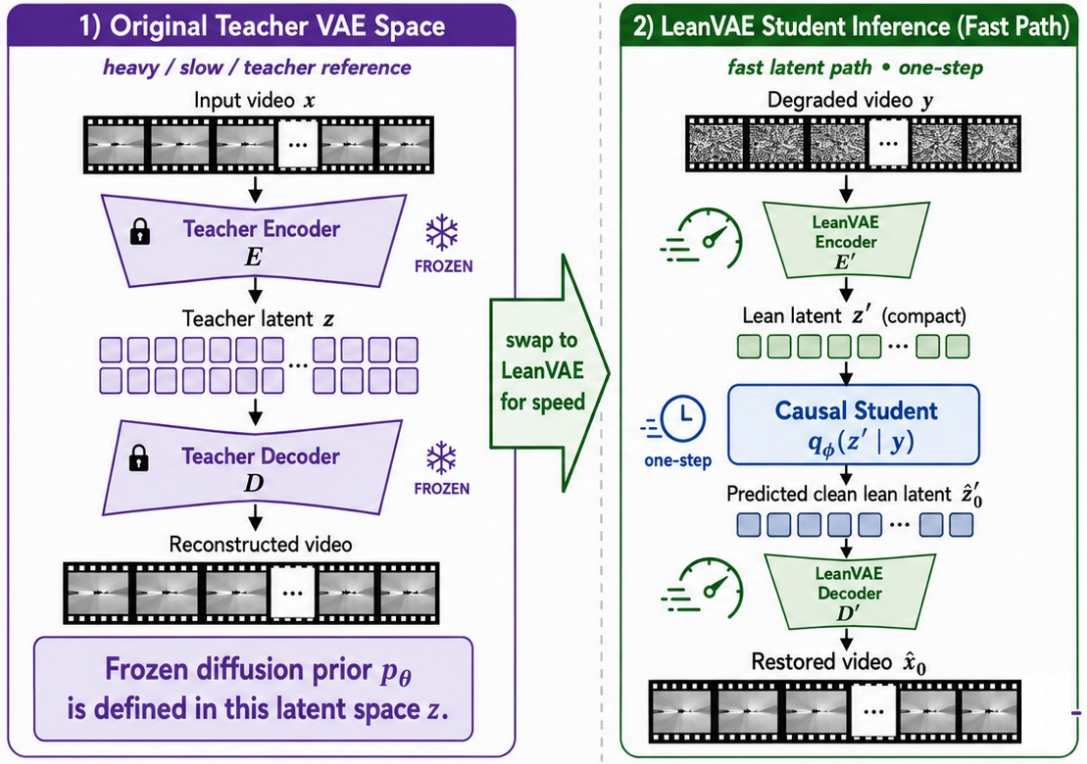
Training: Distill a Frozen Video Diffusion Prior into a Causal Student



PEKING UNIVERSITY

Replacing Heavy (Wan)VAE with LeanVAE

Teacher-space regularization preserves the frozen diffusion prior after the VAE swap.

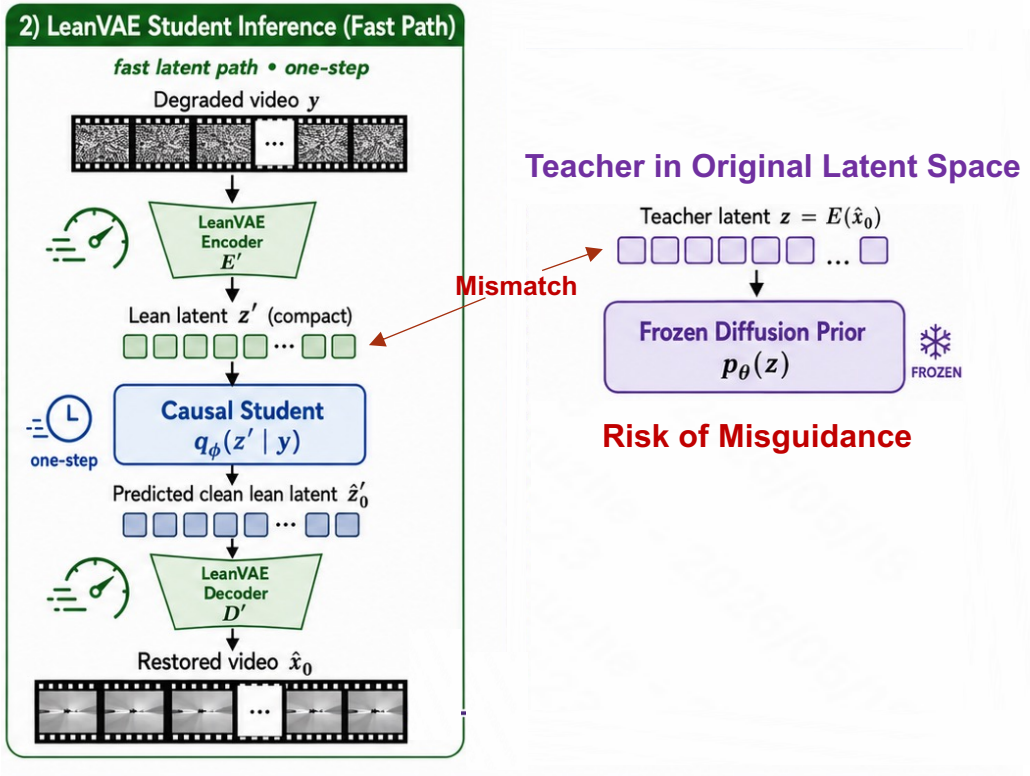


LeanVAE gives speed; teacher-space regularization keeps the student aligned with the original diffusion prior.

北京大学 PEKING UNIVERSITY

Replacing Heavy (Wan)VAE with LeanVAE

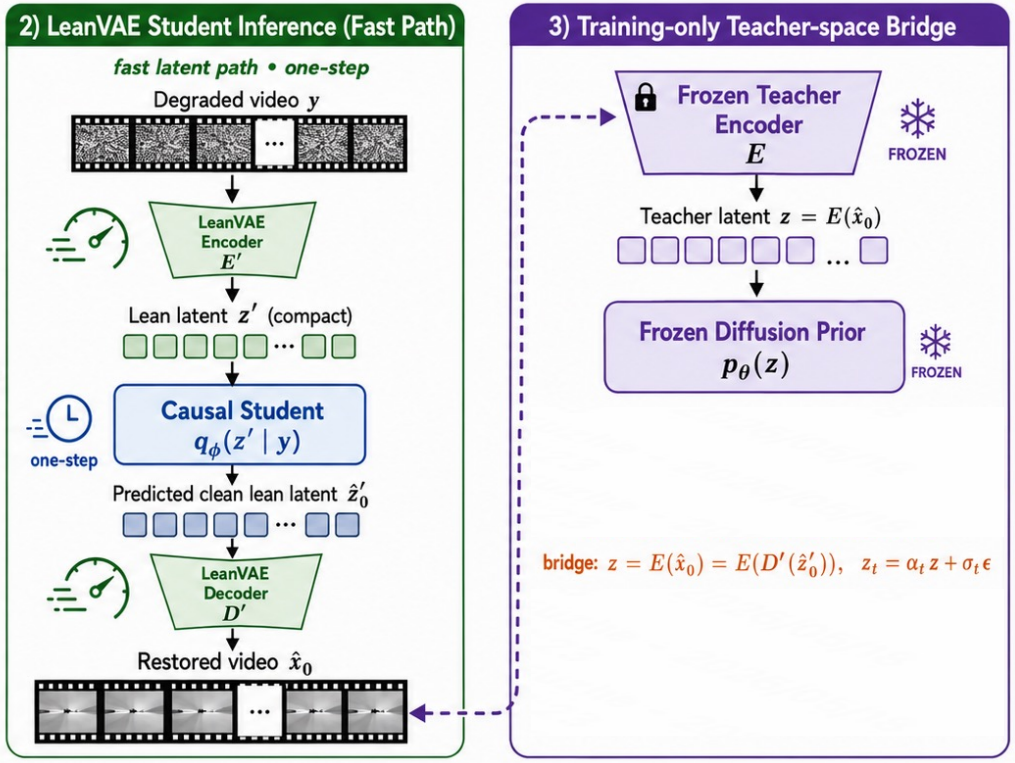
Teacher-space regularization preserves the frozen diffusion prior after the VAE swap.



北京大学 PEKING UNIVERSITY

Replacing Heavy (Wan)VAE with LeanVAE

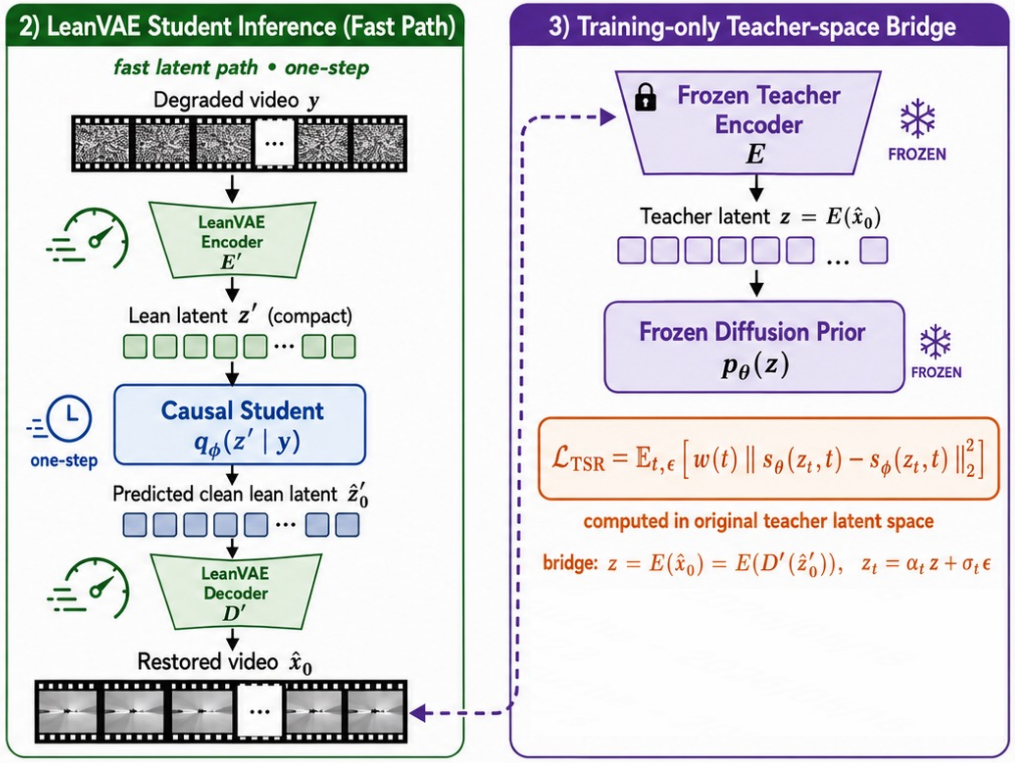
Teacher-space regularization preserves the frozen diffusion prior after the VAE swap.



北京大学 PEKING UNIVERSITY

Replacing Heavy (Wan)VAE with LeanVAE

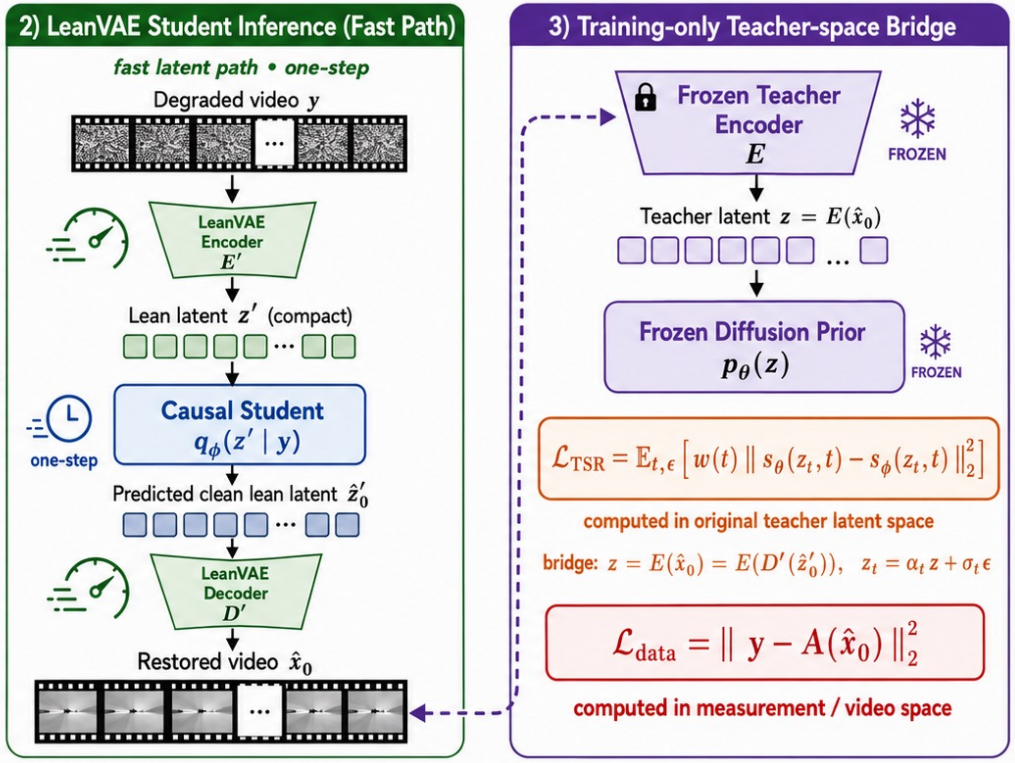
Teacher-space regularization preserves the frozen diffusion prior after the VAE swap.



北京大学 PEKING UNIVERSITY

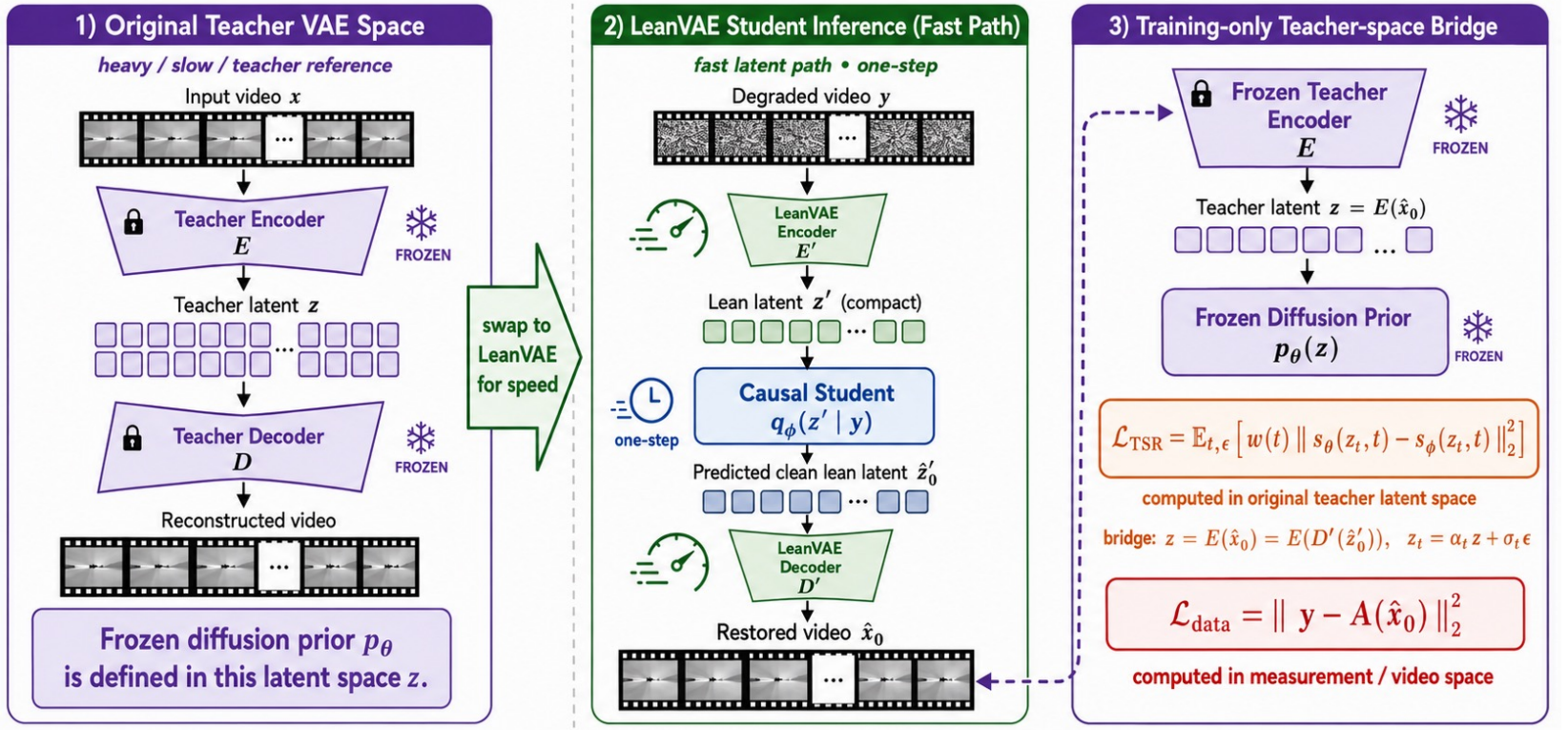
Replacing Heavy (Wan)VAE with LeanVAE

Teacher-space regularization preserves the frozen diffusion prior after the VAE swap.



Replacing Heavy (Wan)VAE with LeanVAE

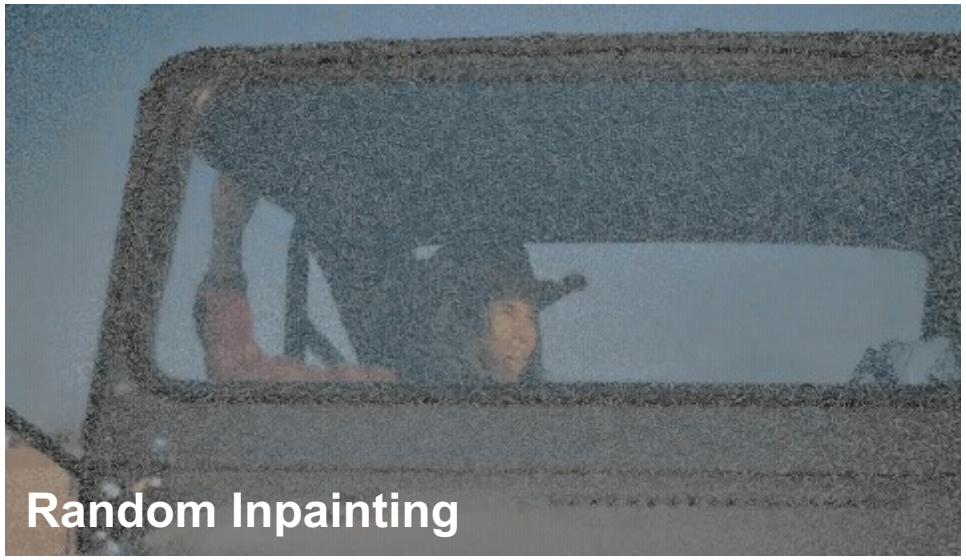
Teacher-space regularization preserves the frozen diffusion prior after the VAE swap.



Discarded After Training

LeanVAE accelerates inference in z' space; teacher-space regularization maps $z' \to \hat{x}_0 \to z$ so the frozen diffusion prior still supervises in its original latent space.

Results



Results

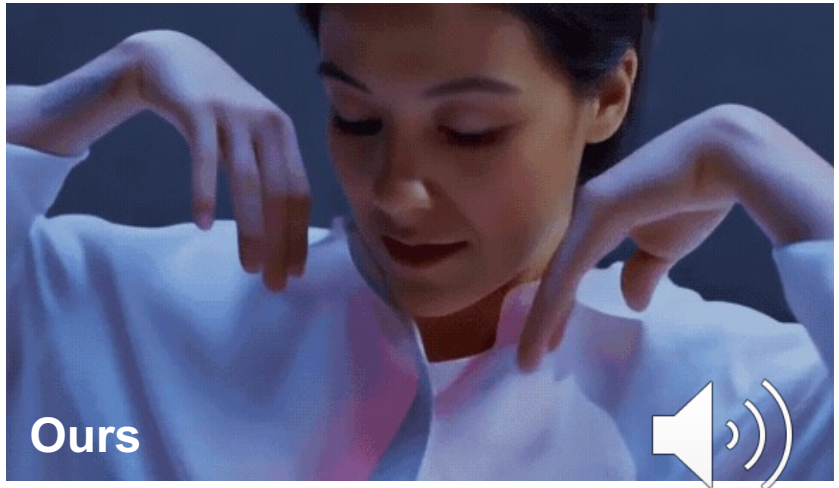
One solver for multiple inverse problems

北京大学 PEKING UNIVERSITY

Random inpainting

Gaussian deblurring

4× video SR



Results

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**

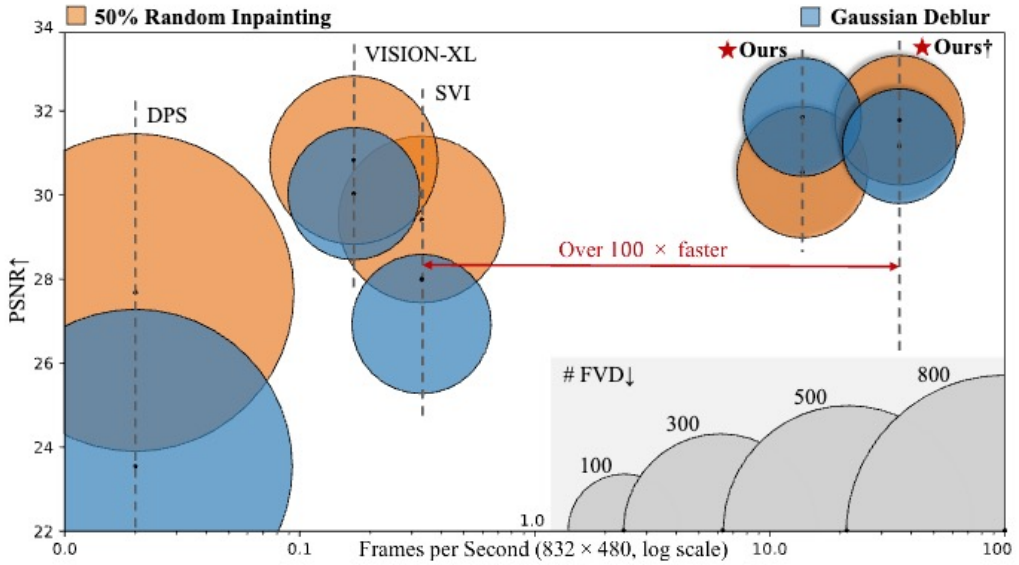


Table 2. **Quantitative results of spatial quality.** Metrics include PSNR, SSIM, LPIPS. Best results are in **bold**, suboptimal are underlined.

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**

Results

Summary

- **Distilling** a frozen video diffusion prior into a **causal one-step student**.
  - **Measurement consistency**
  - **Streaming-ready inference**

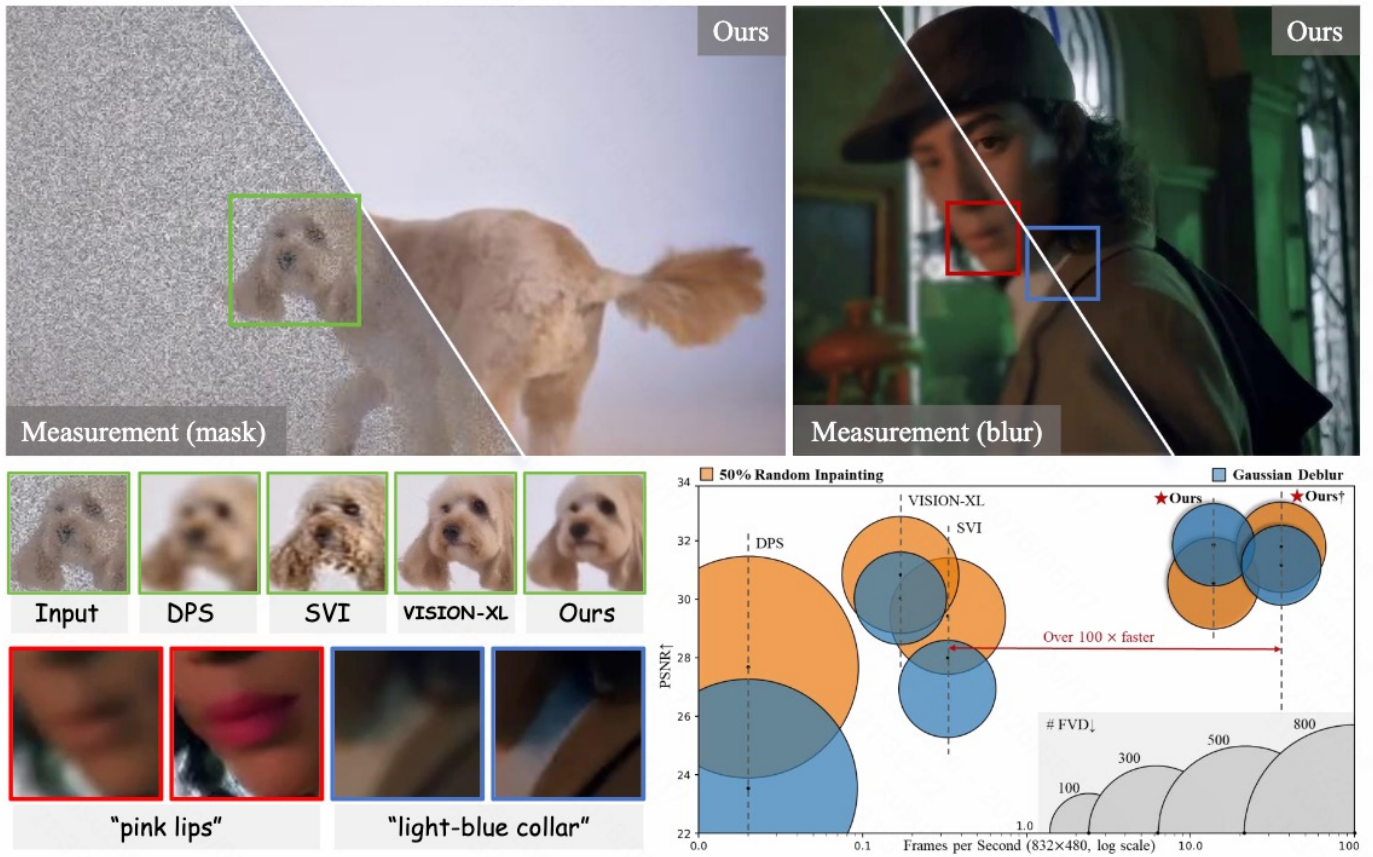


Figure 1. We introduce InstantViR, a real-time video inverse problem solver that drastically outperforms slow sampling-based methods in both speed and quality. **Bottom-right**: At 832×480 resolution, our amortized framework is over **100**× faster than sampling-based baselines like SVI [10], achieving 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").

The End

北京大学
PEKING UNIVERSITY

THANK YOU For Listening