NEURAL INFORMATION PROCESSING SYSTEMS

DOVE: Efficient One-Step Diffusion Model for Real-World Video Super-Resolution

Zheng Chen[1*], Zichen Zou[1*], Kewei Zhang[1], Xiongfei Su[3], Xin Yuan[4], Yong Guo[5], Yulun Zhang[1†]

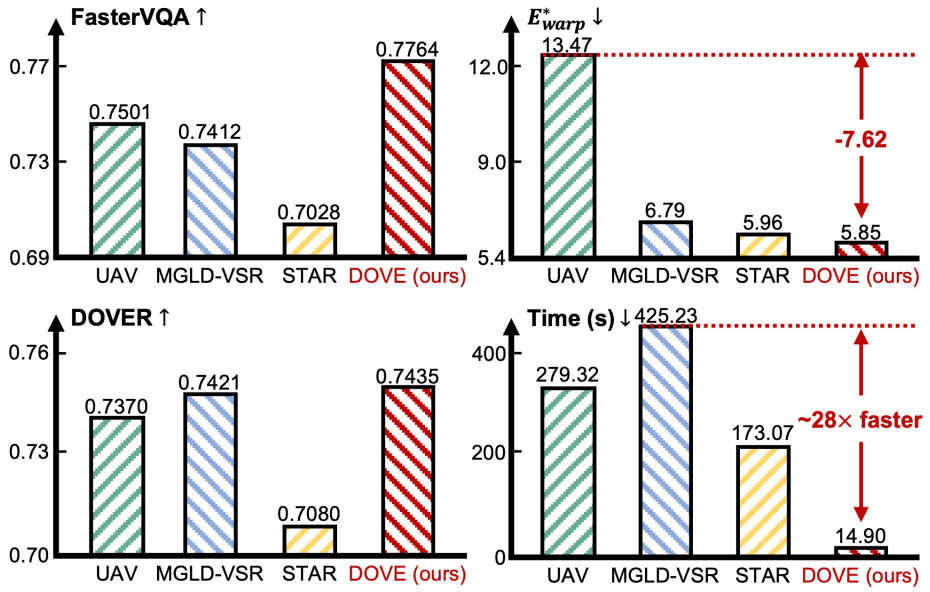
[1]School of Computer Science, Shanghai Jiao Tong University, [2]Zhiyuan College, Shanghai Jiao Tong University,

[3]China Mobile Research Institute, [4]Westlake University, [5]Huawei Consumer Business Group

Introduction

Overview

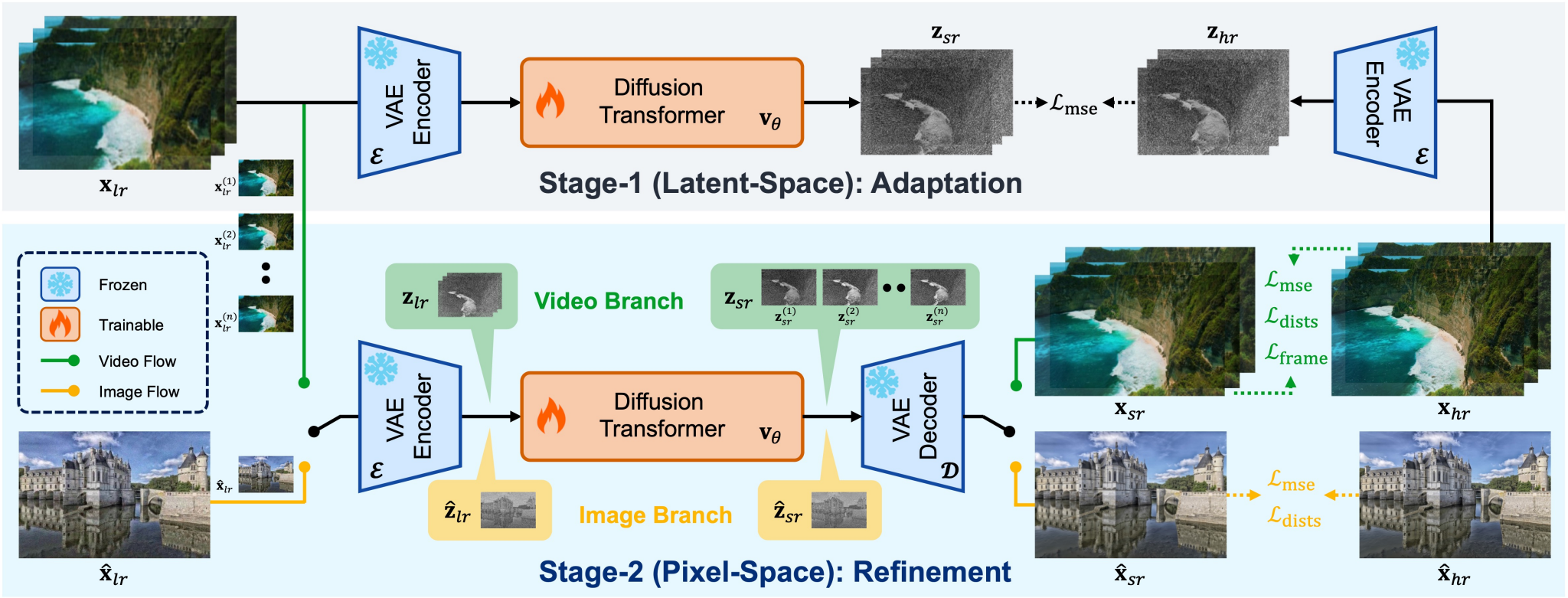
- Diffusion models show strong potential in video super-resolution (VSR).
- Existing methods suffer from **multi-step sampling** and **extra modules**.
- We propose **DOVE**, an efficient one-step diffusion model for VSR.
- DOVE delivers up to **28×** faster than previous diffusion-based methods.



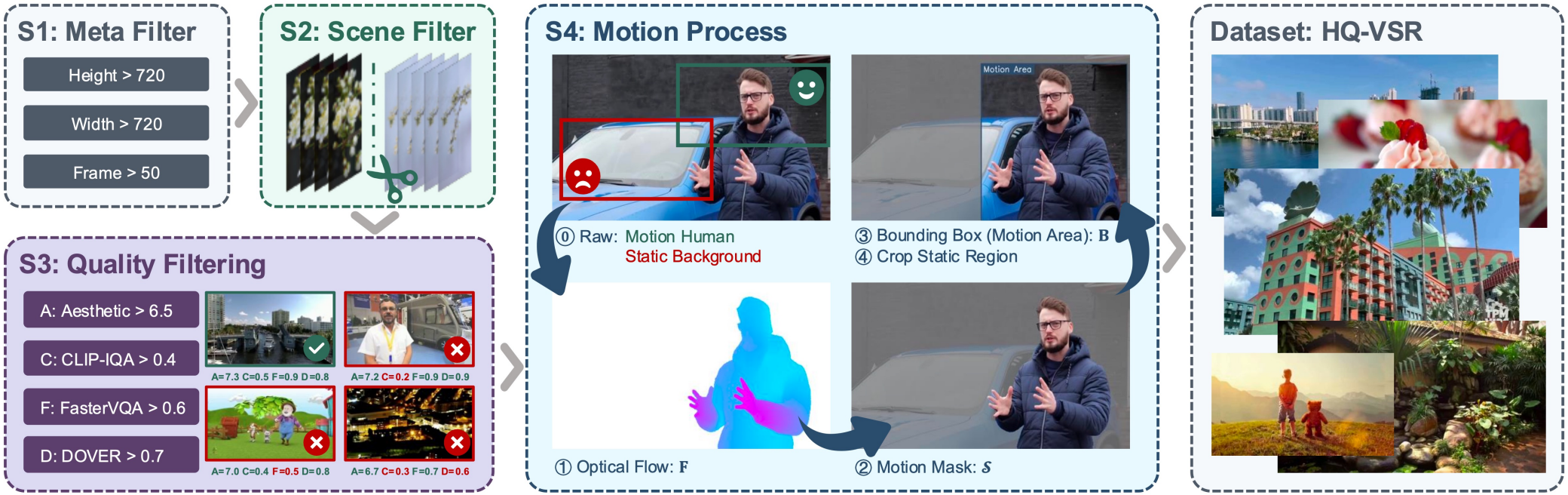
Method

Latent-Pixel Training Strategy

- **Stage 1 (latent-space)**: Learn LR→HR mapping by minimizing latent differences.
- **Stage 2 (pixel-space)**: Refine details via mixed **image**/**video** training in pixel level.



Method

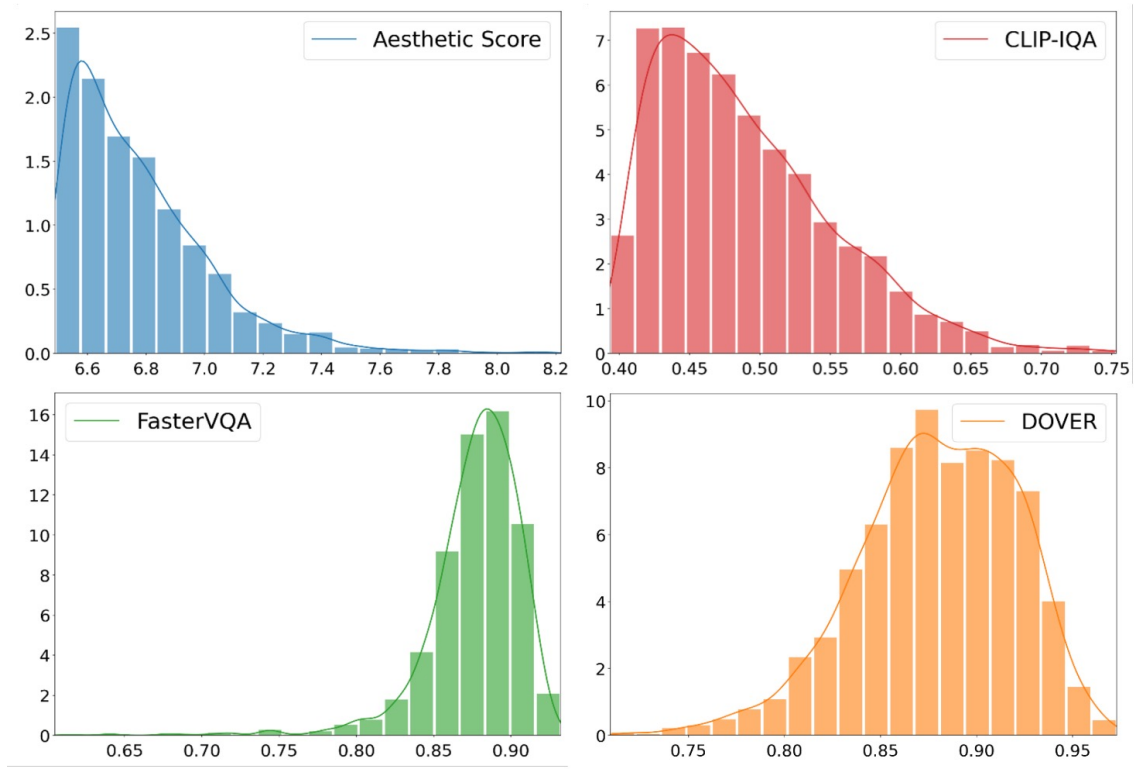


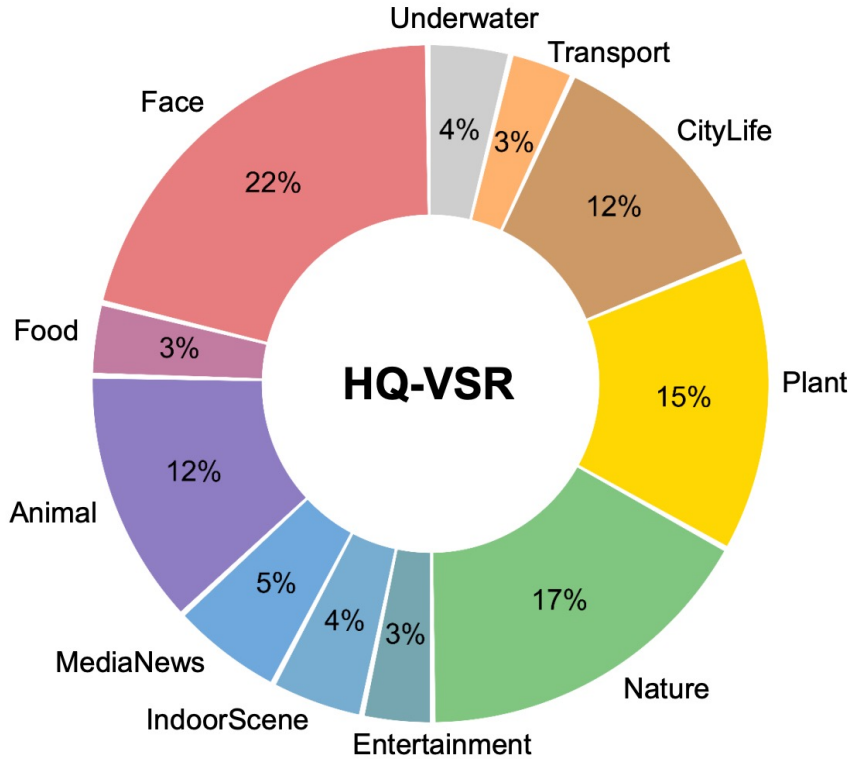
Video Processing Pipeline

- A systematic **pipeline** to curate high-quality videos for VSR fine-tuning.
- Include **four steps**: metadata, scene, quality, and motion filtering.
- Construct **HQ-VSR**, a dataset of 2,055 high-quality videos.

Video Processing Pipeline

- **Quality diversity:** High and diverse scores across multiple metrics.
- **Scene coverage:** Covers **11** scene categories, improving model generalization.





Experiments

Training Stage	S1	S1+S2-I	S1+S2-I/V
PSNR ↑	**27.20**	26.39	26.48
LPIPS ↓	0.3037	0.2784	**0.2696**
CLIP-IQA ↑	0.3236	0.5085	**0.5107**
DOVER ↑	0.6154	0.7694	**0.7809**

Training Strategy

- **Stage-1:** produce smooth results.
- **Stage-2:** better perceptual quality.
- **Mixed training:** enhances temporal consistency performance.

Image Ratio

Mixed image/video training ($\varphi = 0.8$) achieves the best balance of quality and stability.

Image Ratio	0% (video)	20%	50%	80%	100% (image)
PSNR ↑	26.41	26.41	26.44	**26.48**	26.39
LPIPS ↓	0.2624	**0.2617**	0.2686	0.2696	0.2784
CLIP-IQA ↑	0.4800	0.5012	0.5027	**0.5107**	0.5085
DOVER ↑	0.7647	0.7701	0.7751	**0.7809**	0.7694

Experiments

Dataset	PSNR ↑	LPIPS ↓	CLIP-IQA ↑	DOVER ↑
YouHQ	26.88	0.3383	0.2496	0.3965
OpenVid-1M	27.04	0.3376	0.2683	0.4363
HQ-VSR	**27.20**	**0.3037**	**0.3236**	**0.6154**

Training Dataset

- HQ-VSR achieves the best performance
- Quality matters more than quantity.

Processing Pipeline

- Motion-based cropping for dynamic regions.
- Each step progressively improves performance.

Pipeline	PSNR ↑	LPIPS ↓	CLIP-IQA ↑	DOVER ↑
OpenVid-1M	27.04	0.3376	0.2683	0.4363
+Filter	27.09	0.3236	0.2894	0.5357
+Motion	**27.20**	**0.3037**	**0.3236**	**0.6154**

Experiments

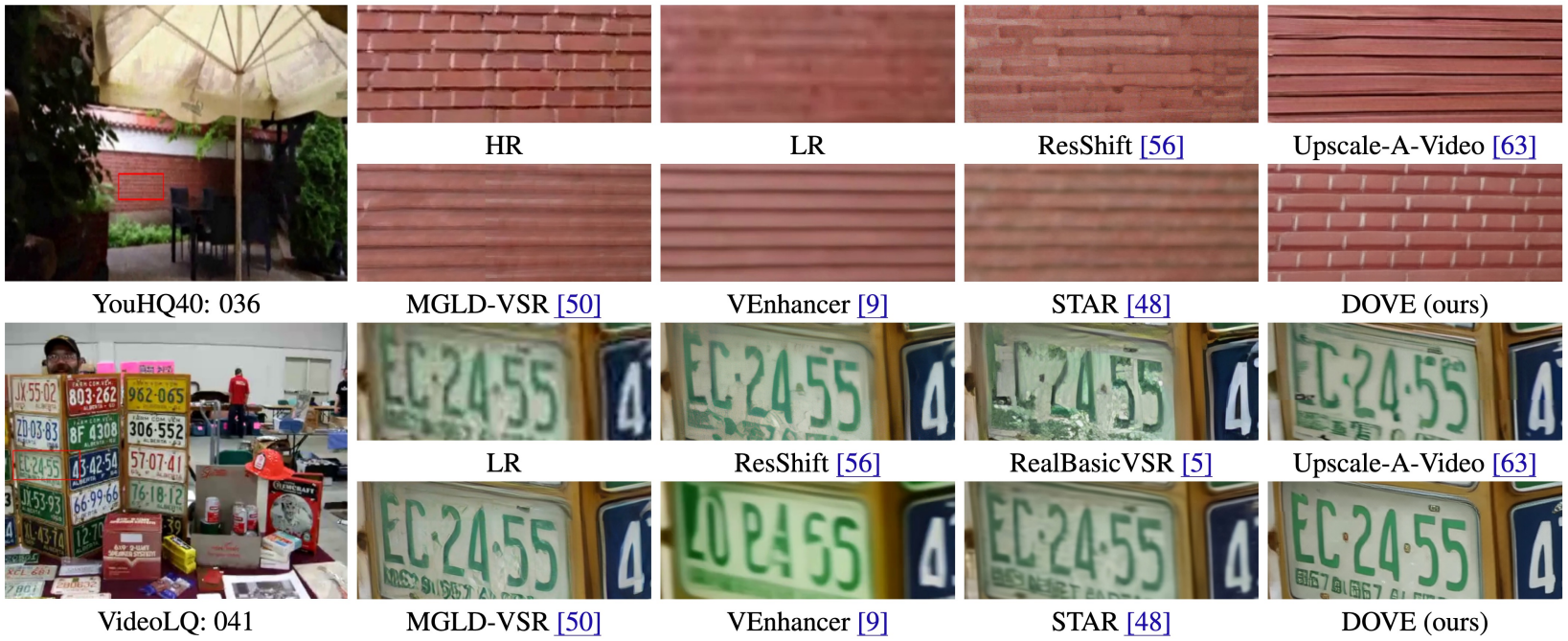
Dataset	Metric	RealESRGAN [38]	ResShift [56]	RealBasicVSR [5]	Upscale-A-Video [63]	MGLD-VSR [50]	VEnhancer [9]	STAR [48]	DOVE (ours)
UDM10	PSNR ↑	24.04	23.65	24.13	21.72	24.23	21.32	23.47	26.48
	SSIM ↑	0.7107	0.6016	0.6801	0.5913	0.6957	0.6811	0.6804	0.7827
	LPIPS ↓	0.3877	0.5537	0.3908	0.4116	0.3272	0.4344	0.4242	0.2696
	DISTS ↓	0.2184	0.2898	0.2067	0.2230	0.1677	0.2310	0.2156	0.1492
	CLIP-IQA ↑	0.4189	0.4344	0.3494	0.4697	0.4557	0.2852	0.2417	0.5107
	FasterVQA ↑	0.7386	0.4772	0.7744	0.6969	0.7489	0.5493	0.7042	0.8064
	DOVER ↑	0.7060	0.3290	0.7564	0.7291	0.7264	0.4576	0.4830	0.7809
	E^*_{warp} ↓	4.83	6.12	3.10	3.97	3.59	1.03	2.08	1.77
MVSR4x	PSNR ↑	22.47	21.58	21.80	20.42	22.77	20.50	22.42	22.42
	SSIM ↑	0.7412	0.6473	0.7045	0.6117	0.7418	0.7117	0.7421	0.7523
	LPIPS ↓	0.4534	0.5945	0.4235	0.4717	0.3568	0.4471	0.4311	0.3476
	DISTS ↓	0.3021	0.3351	0.2498	0.2673	0.2245	0.2800	0.2714	0.2363
	CLIP-IQA ↑	0.4396	0.5003	0.4118	0.6106	0.3769	0.3104	0.2674	0.5453
	FasterVQA ↑	0.3371	0.4723	0.7497	0.7663	0.6764	0.3584	0.2840	0.7742
	DOVER ↑	0.2111	0.3255	0.6846	0.7221	0.6214	0.3164	0.2137	0.6984
	E^*_{warp} ↓	1.64	3.89	1.69	5.10	1.55	0.62	0.61	0.78
VideoLQ	CLIP-IQA ↑	0.3617	0.4049	0.3433	0.4132	0.3465	0.3031	0.2652	0.3484
	FasterVQA ↑	0.7381	0.5909	0.7586	0.7501	0.7412	0.6769	0.7028	0.7764
	DOVER ↑	0.7310	0.6160	0.7388	0.7370	0.7421	0.6912	0.7080	0.7435
	E^*_{warp} ↓	7.58	7.79	5.97	13.47	6.79	6.495	5.96	5.85

Method	Upscale-A-Video [22]	MGLD-VSR [16]	VEnhancer [2]	STAR [14]	DOVE-2B (ours)	DOVE (ours)
Inference Step	30	50	15	15	1	1
Parameters (M)	1,086.75	1,564.66	2,496.59	2,492.90	1,910.28	5,787.19
MACs (T)	9,084.73	8,528.7	3,056.16	4,281.67	461.38	504.81
Running Time (s)	279.32	425.23	121.27	173.07	14.88	14.90

Quantitative

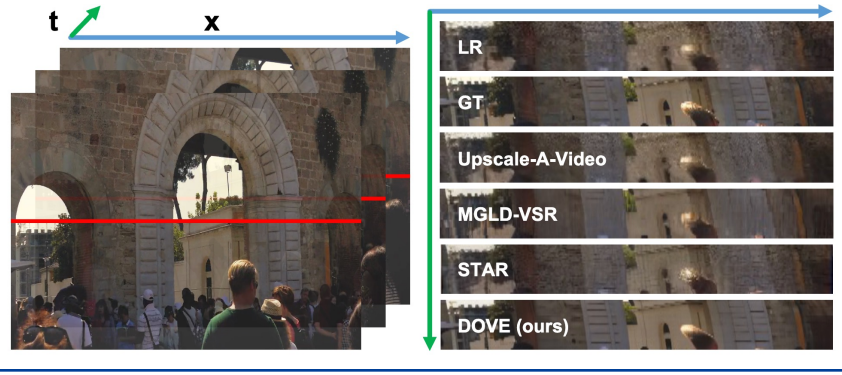
- **State-of-the-art performance:** DOVE surpasses existing VSR methods on most benchmarks.
- **Significant acceleration:** DOVE achieves up to **28×** faster inference than multi-step methods.

Experiments



Qualitative

- **Visual quality:** DOVE produces realistic details.
- **Consistency:** Achieves superior **temporal** and **spatial** coherence.



Conclusion

Contribution

- Propose **DOVE**, the first **one-step** diffusion model for real-world VSR.
- Design a latent-pixel **training strategy** enabling efficient fine-tuning.
- Build **HQ-VSR**, a high-quality dataset tailored for video restoration.
- Achieve state-of-the-art fidelity with up to **28×** faster inference speed.

Poster

- Exhibit Hall C,D,E
- Wed 3 Dec 4:30 p.m. PST — 7:30 p.m. PST

Project

Thanks!