Learning Spatial Adaptation and Temporal Coherence in Diffusion Models for Video Super-Resolution

Zhikai Chen[†], Fuchen Long[§], Zhaofan Qiu[§], Ting Yao[§], Wengang Zhou[†], Jiebo Luo[‡] and Tao Mei[§]

† University of Science and Technology of China, Hefei, China
§ HiDream.ai, Beijing, China
‡ University of Rochester, Rochester, NY USA

ID

Video Super-Resolution

- **Definition**
  - VSR aims to restore a sequence of high-resolution (HR) frames from their low-resolution (LR) counterparts.
  - HR space - the natural video space
  - HR video - high perceptual quality

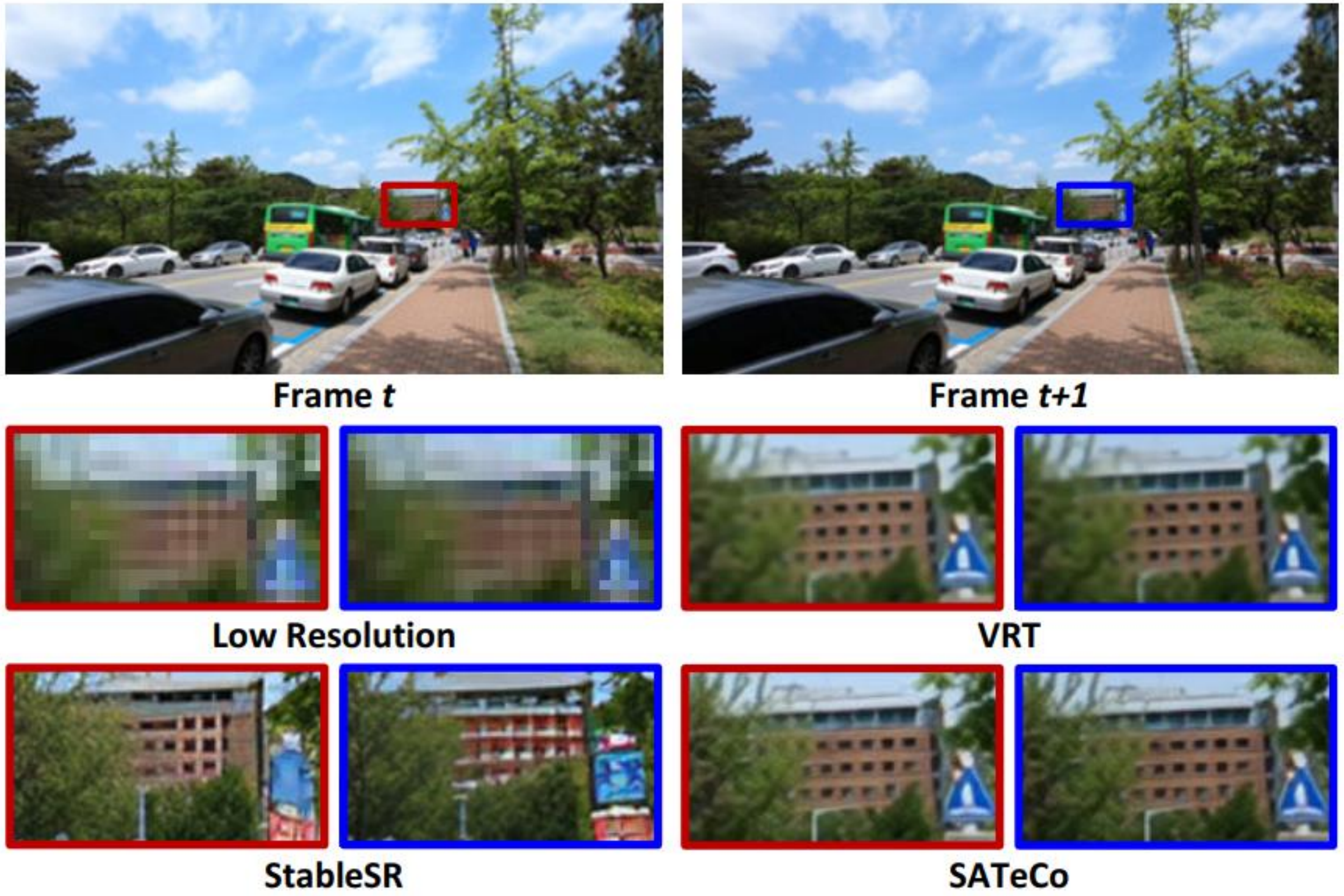
LR Video (320 x 180)

Zoomed LR Video (640 x 360)

HR Video (640 x 360)

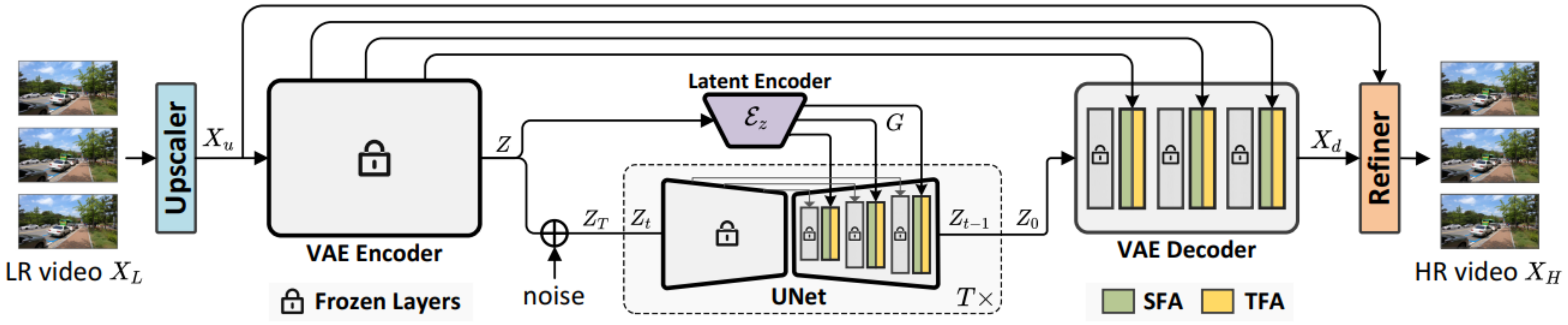
SATeCo

- **Motivation:** exploiting image prior knowledge encapsulated in pre-trained diffusion model for video enhancement
  - How to alleviate the stochasticity in diffusion process to preserve **visual appearance**
  - How to guarantee the **temporal consistency** across frames in the HR videos



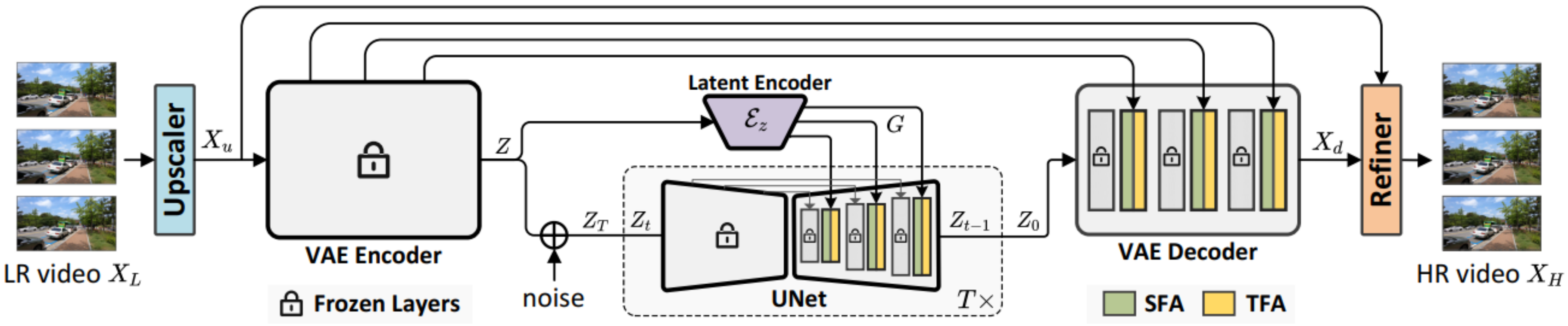
SATeCo

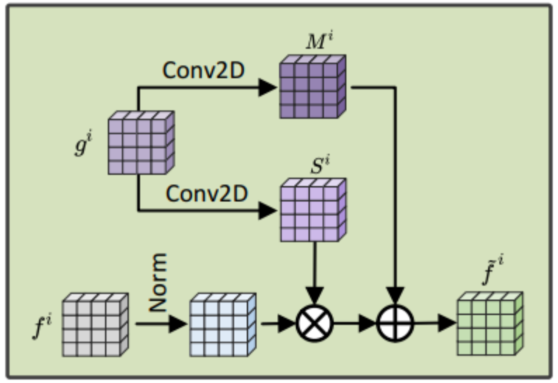
- **Proposal**
  - Spatial Feature Adaptation (SFA) and Temporal Feature Alignment module (TFA)
  - Inserting these two modules in UNet/VAE for **latent-space** video denoising and **pixel-space** video reconstruction



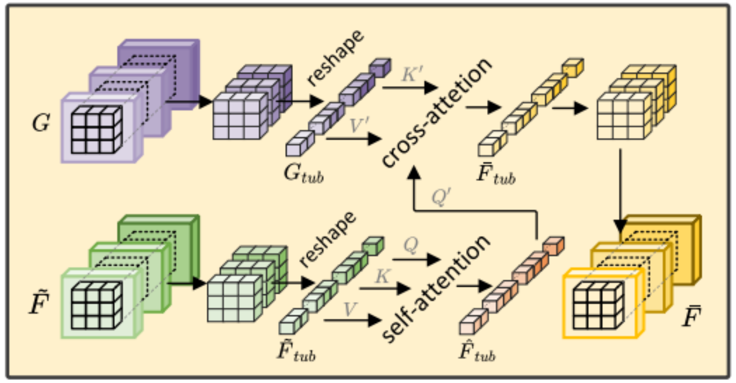
SATeCo – SFA/TFA

- **Proposal**
  - Spatial Feature Adaptation (SFA) and Temporal Feature Alignment module (TFA)
  - Inserting these two modules in UNet/VAE for **latent-space** video denoising and **pixel-space** video reconstruction





(c) Spatial Feature Adaption Module



(d) Temporal Feature Alignment Module

SFA

$$
\begin{cases}
M^i = \mathrm{Conv2D}(g^i), \quad S^i = \mathrm{Conv2D}(g^i). \\
\tilde{f}^i = S^i \odot \dfrac{f^i - \mu^i}{\sigma^i} + M^i,
\end{cases}
$$

TFA

$$
\begin{cases}
Q, K, V = \mathrm{Conv3D}(\tilde{F}_{tub}), \\
\hat{F}_{tub} = \mathrm{Attention}(Q, K, V), \\
Q' = \mathrm{Conv3D}(\hat{F}_{tub}), \quad K', V' = \mathrm{Conv3D}(G_{tub}), \\
\bar{F}_{tub} = \mathrm{Attention}(Q', K', V'),
\end{cases}
$$

Quantitative Results

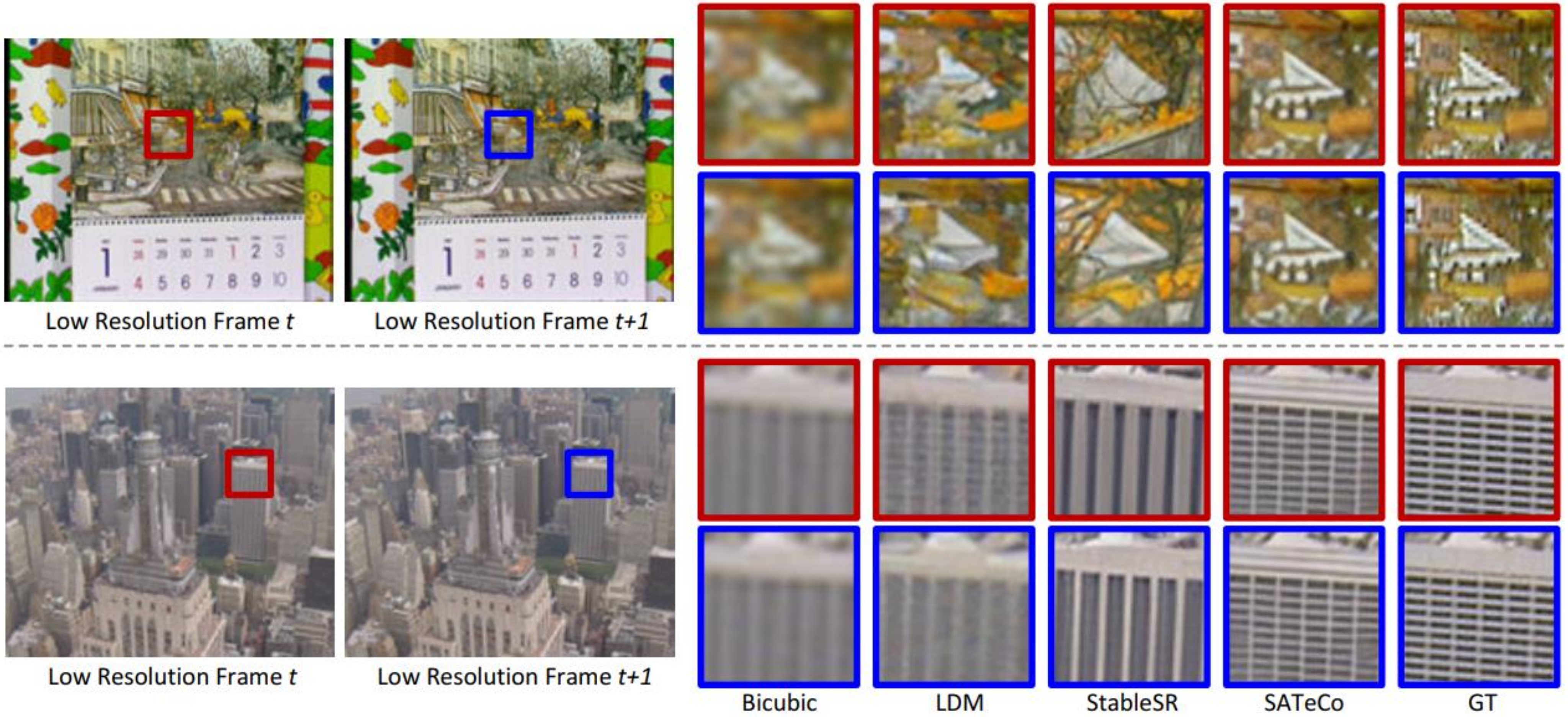
CVPR SEATTLE, WA JUNE 17-21, 2024

- **Evaluation Datasets:**
  - REDS4: videos with 100 frames, HR at 1280x720
  - Vid4: videos with 40 frames, HR at 720x512

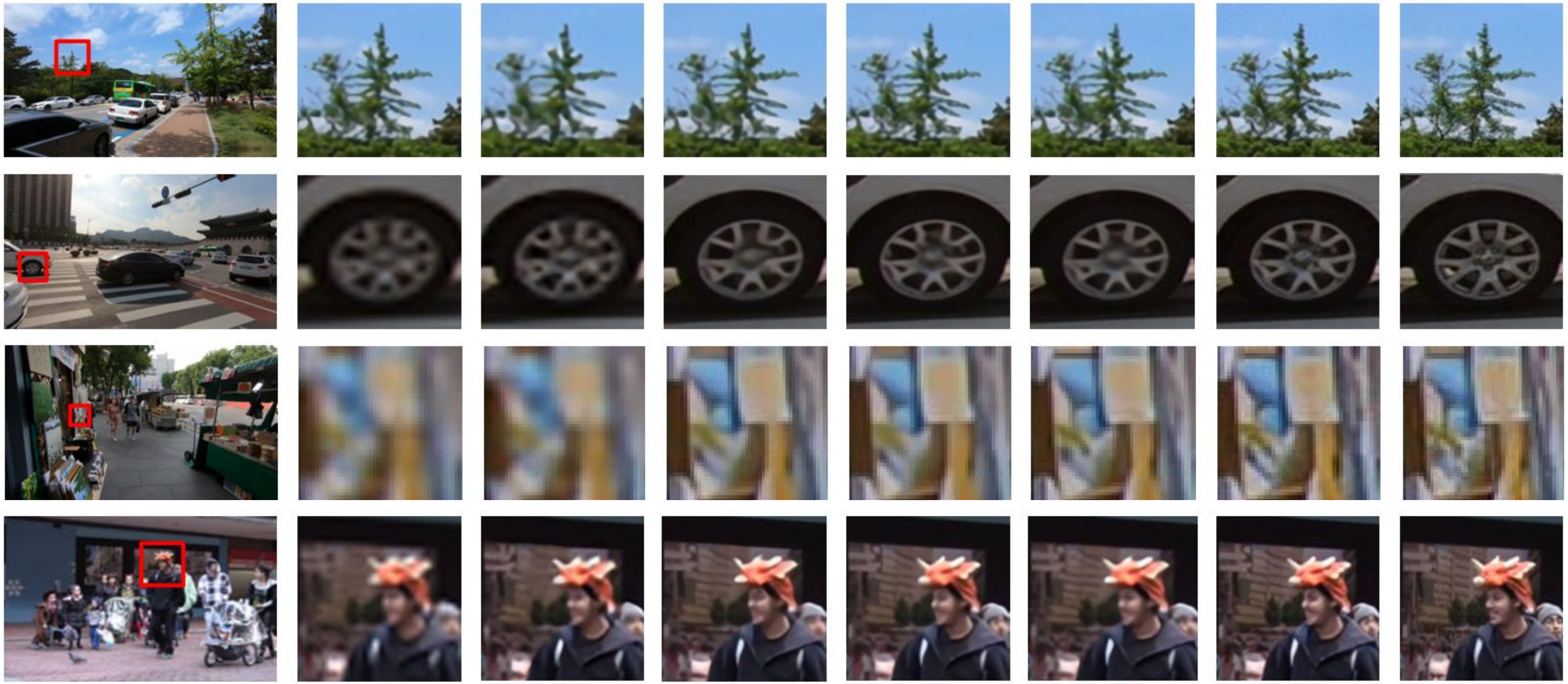
- **Evaluation Metrics:**
  - Pixel-based metrics: PSNR and SSIM
  - Perception-based metrics: LPIPS, DISTS, NIQE and CLIP-IQA

Datasets	Metrics	Bicubic	StableSR [41]	TOFlow [46]	EDVR-M [43]	BasicVSR [2]	VRT [22]	IconVSR [2]	SATeCo
REDS4	PSNR↑	26.14	24.79	27.98	30.53	31.42	31.60	**31.67**	<u>31.62</u>
	SSIM↑	0.7292	0.6897	0.7990	0.8699	0.8909	0.8888	**0.8948**	<u>0.8932</u>
	LPIPS↓	0.3519	0.2412	0.3104	0.2312	0.2023	0.2077	<u>0.1939</u>	**0.1735**
	DISTS↓	0.1876	<u>0.0755</u>	0.1468	0.0943	0.0808	0.0823	0.0762	**0.0607**
	NIQE↓	7.257	<u>4.116</u>	6.260	4.544	4.197	4.252	4.117	**4.104**
	CLIP-IQA↑	0.6045	<u>0.6579</u>	0.6176	0.6382	0.6353	0.6379	0.6162	**0.6622**
Vid4	PSNR↑	23.78	22.18	25.89	27.10	27.24	**27.93**	27.39	<u>27.44</u>
	SSIM↑	0.6347	0.5904	0.7651	0.8186	0.8251	**0.8425**	0.8279	<u>0.8420</u>
	LPIPS↓	0.3947	0.3670	0.3386	0.2898	0.2811	<u>0.2723</u>	0.2739	**0.2291**
	DISTS↓	0.2201	0.1385	0.1776	0.1468	0.1442	<u>0.1372</u>	0.1406	**0.1015**
	NIQE↓	7.536	<u>5.237</u>	7.229	5.528	5.340	5.242	5.392	**5.212**
	CLIP-IQA↑	0.6817	**0.7644**	0.7365	0.7380	0.7410	0.7434	0.7411	<u>0.7451</u>

Visualization Results on Vid4



Visualization Results on REDS4



Input | Bicubic | TOFlow | EDVR-M | VRT | IconVSR | SATeCo | GT

Thanks!