DUO-VSR: Dual-Stream Distillation for One-Step Video Super-Resolution

Supplementary Material

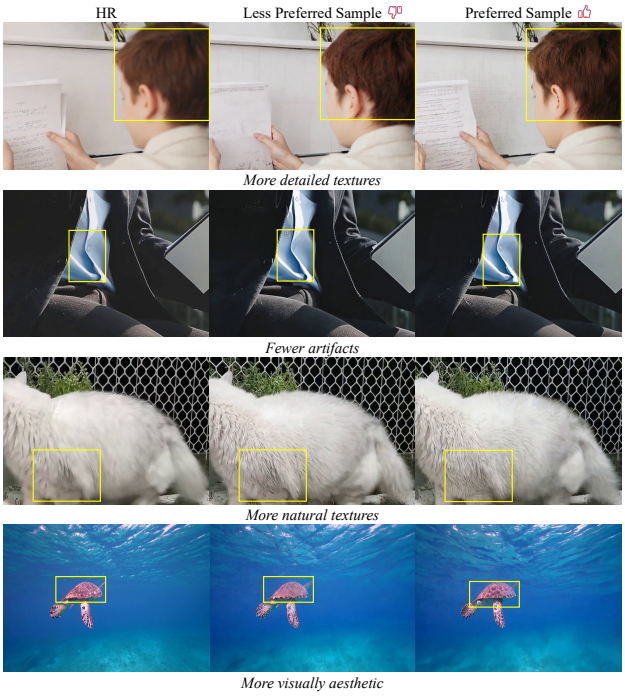

Figure 8. Examples of preferred and less-preferred samples in the constructed preference dataset. Zoom in for details.

7. Further Implementation Details

7.1. Algorithm for Dual-Stream Distillation

The detailed procedure of the dual-stream distillation strategy is outlined in Algorithm 1, comprising interleaved Auxiliary and Student updates. In our implementation, we set the update interval $N = 3$ by default.

7.2. Construction of Preference Dataset

In the preference-guided refinement stage, we construct a preference dataset for Direct Preference Optimization. Specifically, for each LR video, we generate five candidate reconstructions using the second-stage model. We then evaluate these candidates using the LPIPS [90], MUSIQ [16] and DOVER [68] metrics and rank them according to their combined quality scores. The highest-scoring output is selected as the preferred sample, while the lowest-scoring one serves as the less preferred sample. As illustrated in Fig. 8, the preferred samples typically exhibit richer, more natural, and aesthetically pleasing textures. In total, we construct 2000 preference pairs for fine-tuning.

8. Further Discussions and Ablation Analyses.

In the ablation study presented in Sec. 4.3 of the main text, we analyzed the effectiveness of the three-stage distillation framework and the two branches in the Dual-Stream Distillation, namely the DMD stream and the RFS-GAN stream, along with the exploration of different optimization strategies. In this section, we provide additional discussions on the design and training of the RFS-GAN.

Noise-Perturbed Sample Input in RFS-GAN. Different

Algorithm 1: Dual-Stream Distillation Strategy

Input: Frozen Real Score model θ_{R}; trainable Fake Score model θ_{F}; student θ_{S}; discriminator heads H_ϕ; loss weights $\lambda_{\mathrm{DMD}}, \lambda_{\mathrm{GAN}}, \lambda_{\mathrm{FM}}$; interval N.

while *not converged* **do**

  for $i \leftarrow 1$ **to** N **do**

    `/* Auxiliary update */`

    Sample $(\boldsymbol{z}^{LR}, \boldsymbol{z}^{HR}, \boldsymbol{c}), t, \epsilon$;

    $\hat{\boldsymbol{z}}_0^S \leftarrow \epsilon - \boldsymbol{v}_{\theta_{\mathrm{S}}}(\epsilon, t, \boldsymbol{z}^{LR}, \boldsymbol{c})$;

    $\hat{\boldsymbol{z}}_t^S \leftarrow q_t(\hat{\boldsymbol{z}}_0^S),\ \boldsymbol{z}_t^{HR} \leftarrow q_t(\boldsymbol{z}^{HR})$;

    `// Diffusion loss for` θ_{F}

    Compute target $\boldsymbol{v}$ at $(\hat{\boldsymbol{z}}_t^S, t, \boldsymbol{z}^{LR}, \boldsymbol{c})$;

    $\mathcal{L}_{\mathrm{Diff}} \leftarrow \left\|\boldsymbol{v}_{\theta_{\mathrm{F}}}(\hat{\boldsymbol{z}}_t^S, t, \boldsymbol{z}^{LR}, \boldsymbol{c}) - \boldsymbol{v}\right\|^2$;

    `// GAN discriminator loss for` ϕ `with stop_grad backbones`

    $\mathbf{h}^S \leftarrow \mathrm{concat}\big(\mathrm{Feat}_{\theta_{\mathrm{R}}}(\hat{\boldsymbol{z}}_t^S), \mathrm{Feat}_{\theta_{\mathrm{F}}}(\hat{\boldsymbol{z}}_t^S)\big)$;

    $\mathbf{h}^{HR} \leftarrow \mathrm{concat}\big(\mathrm{Feat}_{\theta_{\mathrm{R}}}(\boldsymbol{z}_t^{HR}), \mathrm{Feat}_{\theta_{\mathrm{F}}}(\boldsymbol{z}_t^{HR})\big)$;

    $D_S \leftarrow H_\phi\big(\mathrm{sg}[\mathbf{h}^S]\big)$;

    $D_{HR} \leftarrow H_\phi\big(\mathrm{sg}[\mathbf{h}^{HR}]\big)$;

    $\mathcal{L}_{\mathrm{D}} \leftarrow \mathbb{E}[\max(0, 1 - D_{HR})] + \mathbb{E}[\max(0, 1 + D_S)]$;

    Update θ_{F} by descending $\nabla_{\theta_{\mathrm{F}}} \mathcal{L}_{\mathrm{Diff}}$;

    Update ϕ by descending $\nabla_\phi \mathcal{L}_{\mathrm{D}}$;

  `/* Student update (after every` N `Auxiliary steps) */`

  Sample $(\boldsymbol{z}^{LR}, \boldsymbol{z}^{HR}, \boldsymbol{c}), t, \epsilon$;

  $\hat{\boldsymbol{z}}_0^S \leftarrow \epsilon - \boldsymbol{v}_{\theta_{\mathrm{S}}}(\epsilon, t, \boldsymbol{z}^{LR}, \boldsymbol{c})$;

  $\hat{\boldsymbol{z}}_t^S \leftarrow q_t(\hat{\boldsymbol{z}}_0^S),\ \boldsymbol{z}_t^{HR} \leftarrow q_t(\boldsymbol{z}^{HR})$;

  `// DMD loss`

  $\hat{\boldsymbol{z}}_0^R \leftarrow \hat{\boldsymbol{z}}_0^R(\hat{\boldsymbol{z}}_t^S; \theta_{\mathrm{R}}), \quad \hat{\boldsymbol{z}}_0^F \leftarrow \hat{\boldsymbol{z}}_0^F(\hat{\boldsymbol{z}}_t^S; \theta_{\mathrm{F}})$;

  $\mathrm{Grad} \leftarrow \dfrac{\hat{\boldsymbol{z}}_0^F - \hat{\boldsymbol{z}}_0^R}{\mathrm{mean}\big(\mathrm{abs}(\hat{\boldsymbol{z}}_0^S - \hat{\boldsymbol{z}}_0^R)\big)}$;

  $\mathcal{L}_{\mathrm{DMD}} \leftarrow \left\|\hat{\boldsymbol{z}}_0^S - \mathrm{sg}[\hat{\boldsymbol{z}}_0^S - \mathrm{Grad}]\right\|^2$;

  `// GAN generator loss`

  $\mathbf{h}^S \leftarrow \mathrm{concat}\big(\mathrm{Feat}_{\theta_{\mathrm{R}}}(\hat{\boldsymbol{z}}_t^S), \mathrm{Feat}_{\theta_{\mathrm{F}}}(\hat{\boldsymbol{z}}_t^S)\big)$;

  $D(\hat{\boldsymbol{z}}_t^S) \leftarrow H_\phi\big(\mathrm{sg}[\mathbf{h}^S]\big)$;

  $\mathcal{L}_{\mathrm{G}} \leftarrow -\mathbb{E}[D(\hat{\boldsymbol{z}}_t^S)]$;

  `// Feature matching loss`

  $\mathcal{L}_{\mathrm{FM}} \leftarrow \left\|\mathbf{h}^S - \mathbf{h}^{HR}\right\|^2$;

  $\mathcal{L}_{\mathrm{S}} \leftarrow \lambda_{\mathrm{DMD}}\mathcal{L}_{\mathrm{DMD}} + \lambda_{\mathrm{GAN}}\mathcal{L}_{\mathrm{G}} + \lambda_{\mathrm{FM}}\mathcal{L}_{\mathrm{FM}}$;

  Update θ_{S} by descending $\nabla_{\theta_{\mathrm{S}}} \mathcal{L}_{\mathrm{S}}$;

Figure 9. Noise-perturbed samples stabilize adversarial training and suppress artifacts. Zoom in for details.

from SeedVR2 [60], which directly feeds the clean outputs of the student into the discriminator, we observe that such a design often leads to training instability and occasionally produces grid-like artifacts, as shown in Fig. 9. We hypothesize that this instability stems from a discriminator–generator imbalance, where an overly strong discriminator can easily distinguish real samples from fake ones. Inspired by the perturbation strategy in DMD [83], which intentionally blurs the boundary between real and fake data distributions, we similarly add random noise with varying intensity to both real and fake inputs of the discriminator. This modification effectively stabilizes the adversarial learning while preserving its enhancement effect.

Furthermore, using noisy real and fake samples enables sharing the intermediate features from real and fake score computation for the GAN loss calculation, requiring only an additional extraction of features from real samples and thus reducing the number of forward passes.

Cross-Model and Multi-Layer Feature in RFS-GAN In RFS-GAN, both the real score model and the fake score model are employed as the backbones of the discriminator. As illustrated in Fig. 10, intermediate representations are extracted from the 9th, 18th, and 27th layers of the DiT architecture (consisting of 30 layers in total). RFS-GAN effectively integrates shallow features that capture structural and semantic information with deeper representations that encode richer and more fine-grained details. Furthermore, the two score models are optimized over distinct data distributions: the real score model is intrinsically aligned with the real (teacher) distribution, providing high-quality discriminative guidance, whereas the fake score model dynamically reflects the evolving distribution of the student. The complementarity between these two models substantially enhances the representational capacity of the discriminator, thereby delivering stronger and more reliable gradient feedback to the student model.

Ablation studies are performed on the second-stage model to assess the effectiveness of discriminator features extracted from the real and fake score models. As shown in Tab. 5, the discriminator that combines the real and fake score models achieves the best performance in perceptual metrics, demonstrating the effectiveness of RFS-GAN.

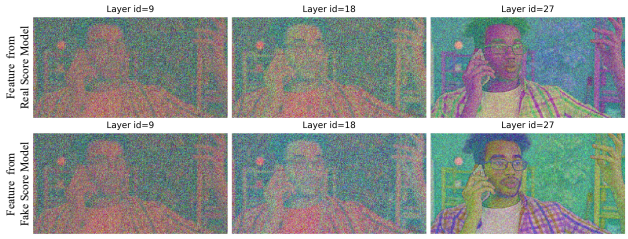

Figure 10. Discriminator features from the real and fake score models used for the RFS-GAN loss computation, reduced to three dimensions via t-SNE [39] for visualization.

Table 5. Ablation study on the discriminator design of RFS-GAN.

Method	NIQE↓	MUSIQ ↑	CLIPIQA ↑	DOVER ↑
Real Score Model only	4.71	62.79	0.456	87.95
Fake Score Model only	4.98	62.98	0.475	87.76
RFS-GAN	**4.64**	**63.36**	**0.487**	**88.01**

9. Additional Evaluation Results

9.1. Additional Visual Comparisons

Comparison with the base model. We first compare DUO-VSR with its base model to examine the effectiveness of the distillation framework, as shown in the Fig. 11. The results indicate that our method achieves a comparable ability to generate textures (first row), while producing more natural and visually coherent details (third and fourth rows).

Comparison with other methods. We present additional visual quality comparisons with VEnhancer [9], UAV [96], STAR [73], DLoRAL [54], DOVE [6], and SEEDVR2 [60] in Fig. 12. These results further demonstrate the advantages of our method when dealing with challenging regions that involve fine textures.

9.2. Discussion of Concurrent Works

We note that several concurrent works [8, 93, 98] have explored efficient video super-resolution, some of which also employ DMD for one-step inference. Both InfVSR [93] and FlashVSR [98] adopt DMD and causal DiT architectures to achieve one-step streaming VSR, focusing primarily on reformulating full-sequence diffusion into a causal structure, where DMD mainly serves as a step-distillation mechanism. Earlier, UltraVSR [30] also employs distribution matching distillation to facilitate one-step VSR, but focuses on degradation-aware scheduling and leverages an image diffusion backbone (extended Stable Diffusion [46] for VSR). In contrast, our DUO-VSR takes an orthogonal perspective by revisiting the intrinsic limitations of DMD in VSR and introducing an effective dual-stream distillation strategy to mitigate them. This design offers a complementary pathway that could potentially be integrated with existing DMD-based frameworks to further enhance their robustness and visual quality.

Recently, both FlashVSR [98] and UltraVSR [30] have made their official implementations publicly available, and

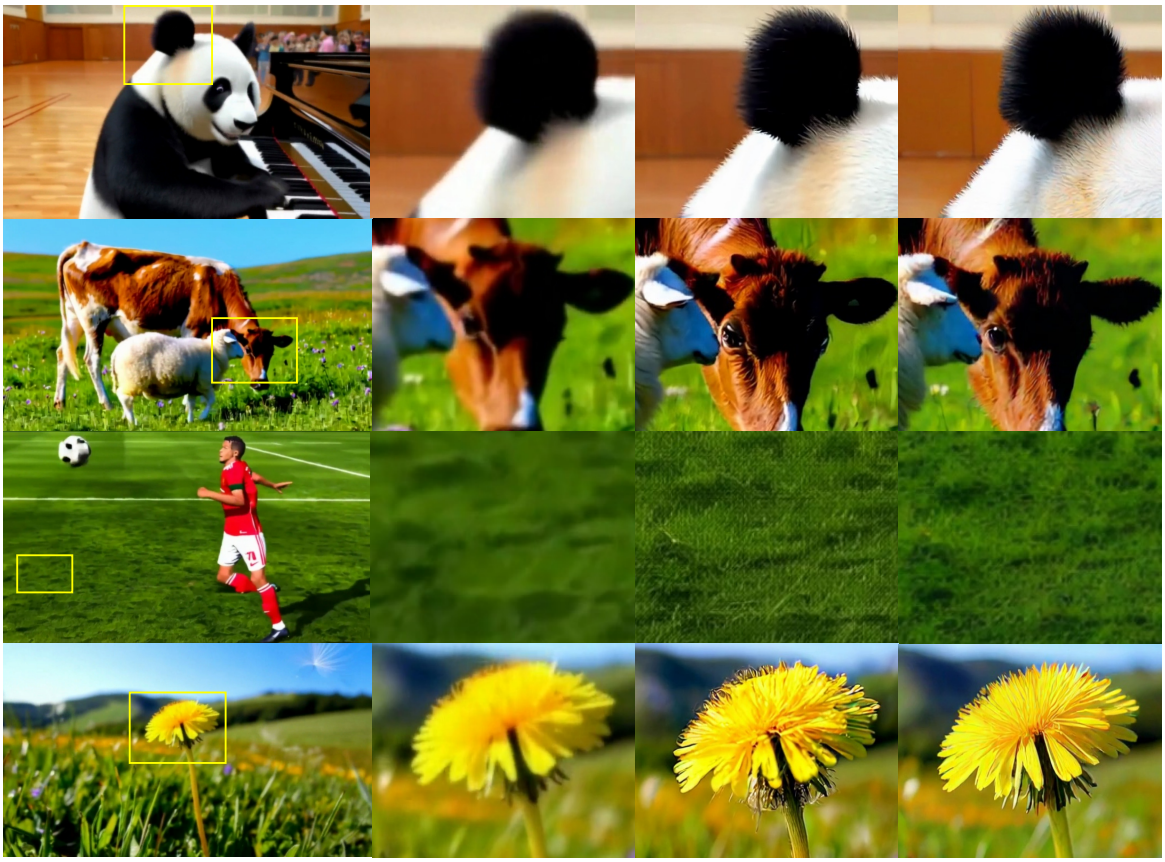

Figure 11. Visual comparison with base VSR model. Zoom in for details.

we include comparative results in this supplementary material. To ensure a fair comparison in terms of performance and quality, we use FlashVSR-Full for evaluation. As shown in Fig. 12, our method produces more realistic and natural details than FlashVSR and UltraVSR. Specifically, in the first case, DUO-VSR reconstructs finer and smoother fur textures on the fox; in the second case, the woman's eyebrows and eyes appear more natural; and in the fourth case, the wheat spikes exhibit more faithful and visually convincing structures. Tab. 6 presents a quantitative comparison between DUO-VSR and these two methods on the AIGC60 dataset. It can be seen that DUO-VSR achieves superior performance in perceptual metrics while exhibiting comparable inference efficiency to FlashVSR-Full.

Table 6. Quantitative comparison on the AIGC60 dataset.

Metric	UltraVSR	FlashVSR	DUO-VSR
NIQE ↓	5.58	<u>4.67</u>	**4.42**
MUSIQ ↑	58.23	<u>63.11</u>	**63.68**
CLIP-IQA ↑	0.4434	<u>0.4690</u>	**0.4886**
DOVER ↑	86.45	<u>87.49</u>	**88.15**
E^*_{warp} ↓	<u>1.54</u>	1.76	**1.08**
Time (s)	126.5	**10.7**	<u>11.3</u>
Params (B)	<u>1.9</u>	**1.3**	**1.3**

9.3. User Study

Following APT [26] and SeedVR2 [60], we conducted a blind user study using the GSB test to more comprehensively assess the subjective visual quality of our method. Specifically, the preference score is computed as $\frac{G-B}{(G+B+S)}$, where G denotes the number of samples judged as good, B as bad, and S as similar. The score ranges from -100% to 100%, with 0% indicating equal performance. We randomly selected 30 samples from the VideoLQ and AIGC60 datasets. The evaluation primarily compared our approach with recent one-step video super-resolution methods, including SeedVR2-7B [60], DOVE [6], DLoRAL [54], UltraVSR [30], and FlashVSR-Full [98]. Participants rated three aspects: visual fidelity, visual quality, and overall quality. Twenty researchers with computer vision backgrounds took part in the evaluation. As shown in Tab. 7, DUO-VSR achieves higher subjective preference scores than previous methods.

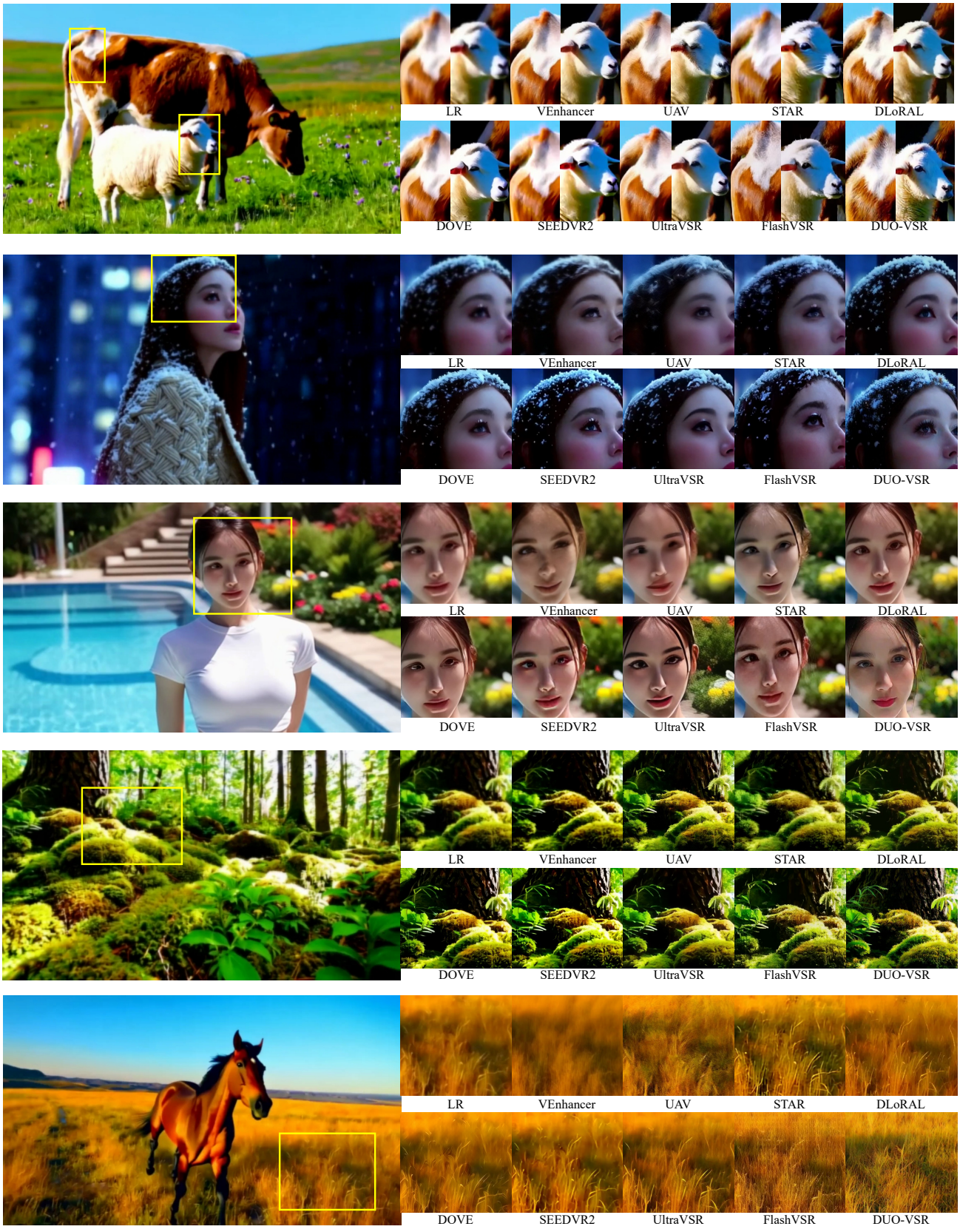

Figure 12. Visual comparison of different VSR methods. DUO-VSR consistently reconstructs finer textures. Zoom in for details.

Table 7. Blind user study results based on GSB test.

Method	Overall Quality	Visual Fidelity	Visual Quality
DUO-VSR	0%	0%	0%
Our Base VSR model	-1.3%	-3.7%	+2.3%
SeedVR2-7B	-32.7%	-13.3%	-39.2%
DOVE	-29.3%	-8.0%	-36.2%
DLoRAL	-34.0%	-10.8%	-37.8%
UltraVSR	-39.8%	-16.7%	-43.3%
FlashVSR-Full	-25.5%	-6.7%	-28.2%

10. Limitations and Future Work

Limitations. Despite the strong efficiency and perceptual quality achieved by our one-step framework, several limitations remain. Since our method is trained in the latent space, the underlying VAE applies an aggressive spatiotemporal compression ($8\times$ spatial and $4\times$ temporal), which can hinder the reconstruction of extremely fine-grained details such as tiny text. In addition, the video VAE becomes the dominant computational bottleneck during inference, accounting for more than 90% of the total runtime.

Future Work. In future work, we plan to explore more efficient or task-specific video VAEs that not only preserve high-frequency details and temporal coherence but also significantly accelerate the decoding process, thereby reducing the overall inference latency of our one-step framework.