DTG-Restore: Training-Free Diffusion Refinement for Generative Video Super-Resolution

Supplementary Material

A. Details on User Study

We conducted a user study with 50 participants who evaluated a total of 60 generated videos using a standardized interface, an example of which is shown in Fig. 6. Participants viewed each video independently and answered three specific questions on a 1–5 Likert scale: **(i) "How would you rate the perceived clarity and sharpness of the video?" (ii) "How would you rate the smoothness of the motion across frames?" (iii) "How would you rate the overall aesthetic quality of the video (i.e., visual appeal)?"** The scoring interface explicitly clarified the meaning of each score range to ensure consistent evaluations across users (see page 1 of the provided questionnaire form). The aggregated results are reported in Table 5, where our DTG-Restore achieves the highest scores across all three metrics, indicating that it was consistently preferred by participants. Venhancer follows as the second-best method but maintains a notable gap from ours, particularly in motion-related quality. SeedVR2 demonstrates strong sharpness but shows instability in motion smoothness, whereas RealViformer presents more balanced but less competitive results. STAR and Upscale-A-Video receive the lowest scores, reflecting their struggle with temporal artifacts and perceptual inconsistencies. These findings confirm that DTG-Restore not only achieves state-of-the-art quantitative results in the creative upscaling task but also aligns strongly with human subjective preferences.

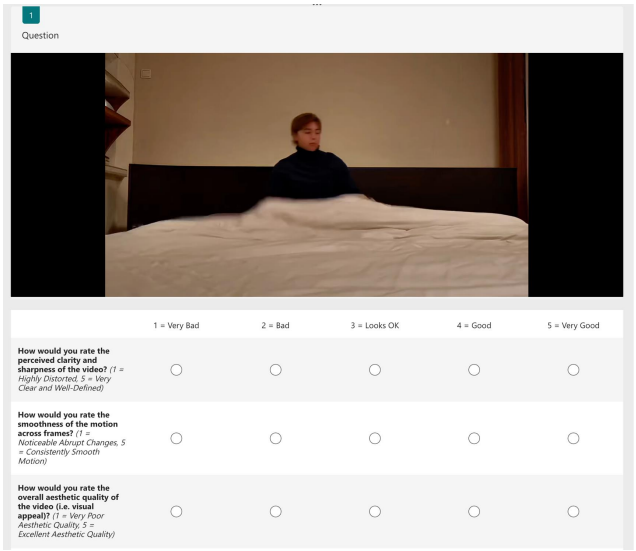

Figure 6. Example of the user study interface used in our evaluation. Each participant watched one video at a time and subsequently rated its sharpness/clearness, motion smoothness, and overall aesthetic quality using a 1–5 Likert scale.

B. Additional Ablations

Figure 7 presents an ablation on DTG and its plug-and-play detail enhancement stage. DTG alone already corrects large geometric distortions and improves temporal stability, indicating that the core gains come from decoupled time guidance during sampling. When DTG is followed by restoration backbones, fine textures and local details are further improved while preserving the corrected structure. This trend is consistent across diverse scenes, showing that DTG serves as a robust first-stage geometry prior and that downstream enhancers primarily contribute detail refinement rather than structural repair.

C. Limitations and Conclusion

While our approach demonstrates strong creative upsampling capabilities, it remains inherently dependent on the representational biases and reconstruction limits of the pretrained diffusion backbone. Since our framework does not retrain the underlying generative model, failure cases may arise in regions where the base model lacks sufficient priors or where degradation is severe. Moreover, the hallucination–reconstruction balance can still be challenging in extreme motions or textures that deviate from the pretrained model's distribution. Despite these constraints, our results show that structured creative upsampling within a diffusion–transformer architecture provides a robust path toward high-quality, temporally coherent video enhancement.

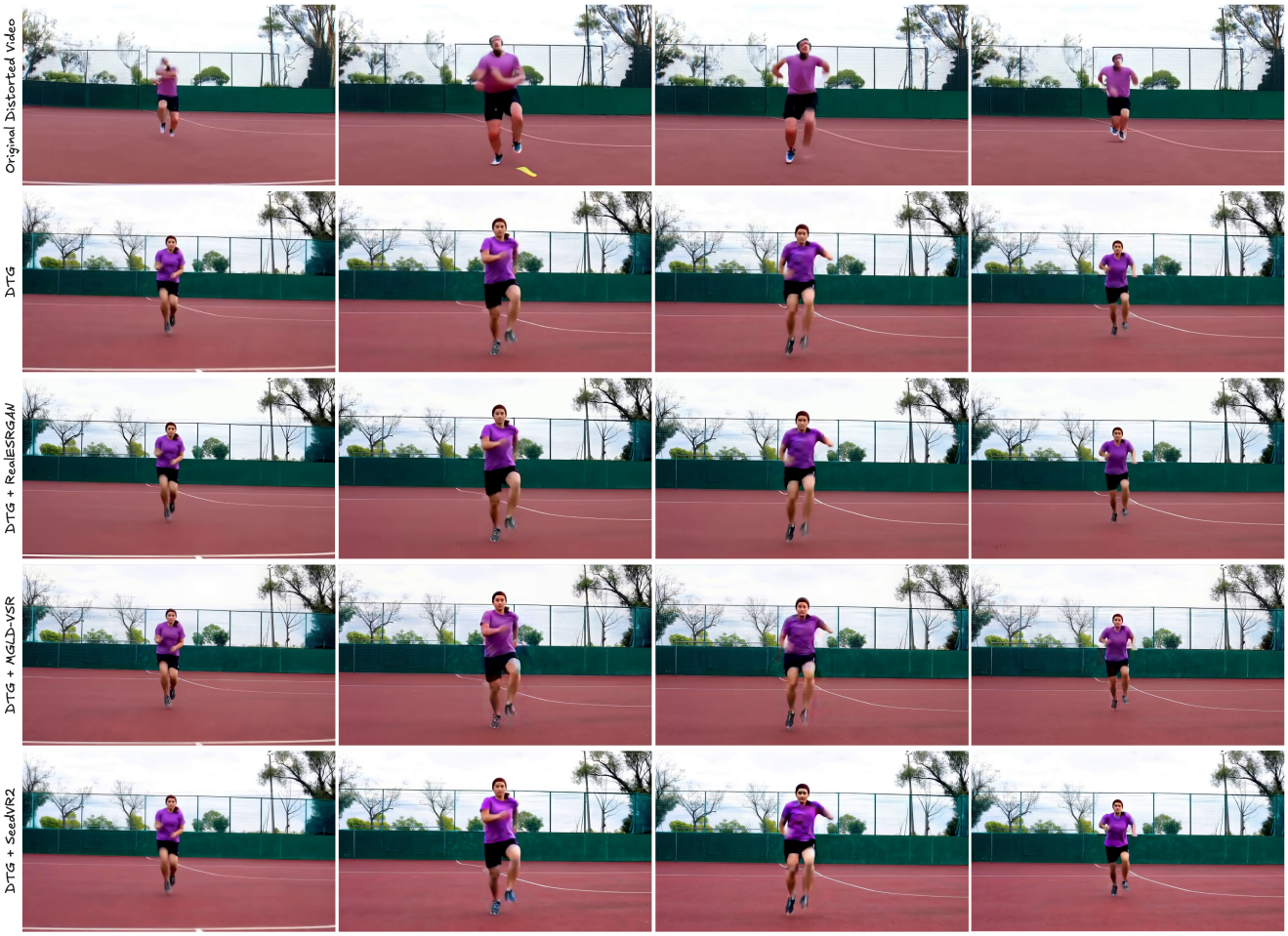

Figure 7. Ablation study of DTG and its plug-and-play detail modules. DTG alone corrects geometry and stabilizes motion, while combining DTG with various restoration models further enhances fine details.