Bias for Action: Video Implicit Neural Representations with Bias Modulation Supplementary Document

Alper Kayabaşı[1], Anil Kumar Vadathya[3], Guha Balakrishnan[2], Vishwanath Saragadam[1]
[1]University of California Riverside, [2]Rice University
[3] Neal Cancer Center, Houston Methodist Hospital
{akaya003,vishwanath.saragadam}@ucr.edu, guha@rice.edu, avadathya@houstonmethodist.org

1. Experimental Results

Performance with varying number of frames. To validate that Act-INR is designed to model local motion, we conducted experiments using a downscaled version of the Bosphorus video. In these experiments, we increased the number of frames in each group of pictures (GOP) and monitored the performance. As anticipated, the performance deteriorated with a larger number of frames in the GOP as shown in Fig. 1. This observation aligns with the interpretation that our model maintains strong capacity as long as the object remains within its surrounding box.

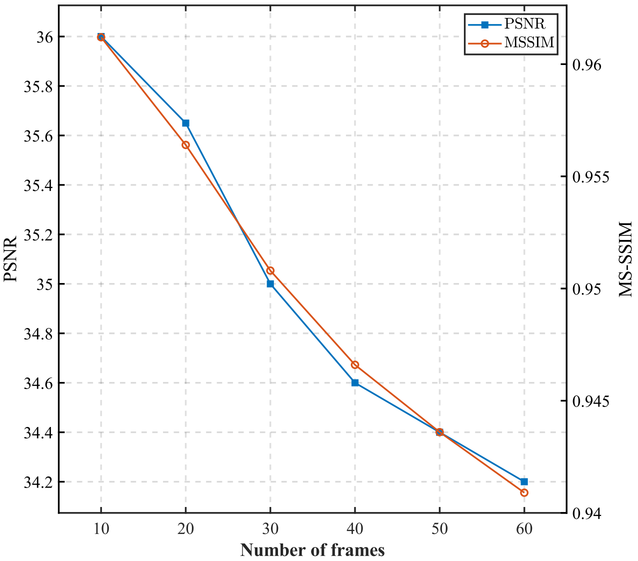

Figure 1. **Effect of varying GOP size on performance.** The plot reveals a downward trend in reconstruction performance as the number of frames in GOP increases. This finding is consistent with our argument that Act-INR is designed to model local motion.

Performance with varying patch size. In this ablation study, we investigate the effect of patch size on the local motion modeling capability of Act-INR. To ensure a fair comparison, we fix the parameter size and the number of frames in GOP while varying the patch size. This setup allows us to isolate and analyze how changes in patch size influence the model's ability to capture local motion dynamics. Neither excessively large nor excessively small patches allow the model to fully utilize its capacity, as illustrated in Fig. 2. For higher spatial resolution and more complex content, excessively large patches can overwhelm the model, while excessively small patches increase the likelihood of objects moving beyond their designated surrounding box. A moderate patch size serves as an optimal sweet spot, balancing these factors by maintaining manageable spatial resolution and ensuring that objects remain within the designated box for effective modeling

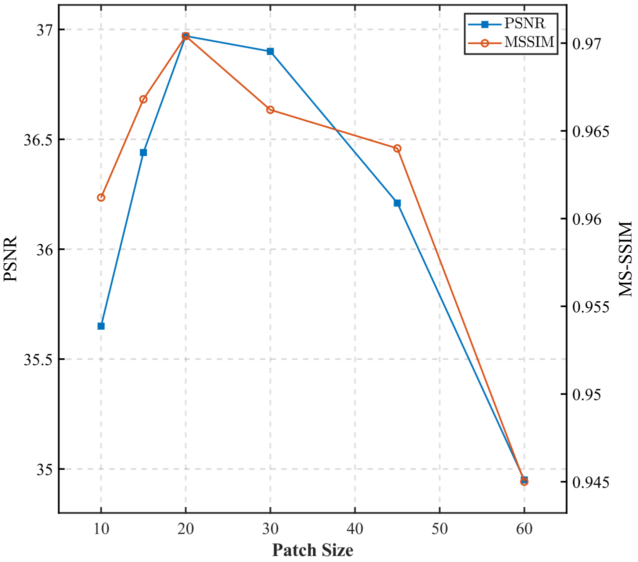

Figure 2. **Effect of varying patch sizes on performance.** The parameter size is fixed at 1.5 million, and the number of frames in GOP is kept constant at 20. The results demonstrate that a moderate-sized window yields optimal performance.

Performance with varying number of parameters. In

Dataset	Bosph	Ready	Yacht	Beauty	Jockey	Honey	Shake	Average
Ds-NeRV	35.2/-	27.1/-	29.4/-	**34.0/-**	32.9/-	**39.6/-**	35.0/-	33.3/-
H-NeRV Boost	36.1/0.96	30.4/0.91	29.3/0.90	33.8/0.90	**35.8**/0.95	39.6/0.98	**35.9**/0.96	34.4/0.94
Ours	**37.5/0.98**	**33.8/0.98**	**30.8/0.94**	33.8/**0.91**	34.9/0.95	38.4/0.98	34.6/0.96	**34.8/0.96**

Table 1. Video regression results on UVG dataset in PSNR and MS-SSIM

this ablation study, we examine how performance scales with the number of parameters. To this end, we progressively increase the feature size from 20 to 60 in increments of 10. The number of frames in GOP is fixed at 20, and the resolution is downscaled by a factor of two, consistent with previous experiments. As shown in Fig. 3, the performance of our model improves proportionally with the parameter count, highlighting its scalability.

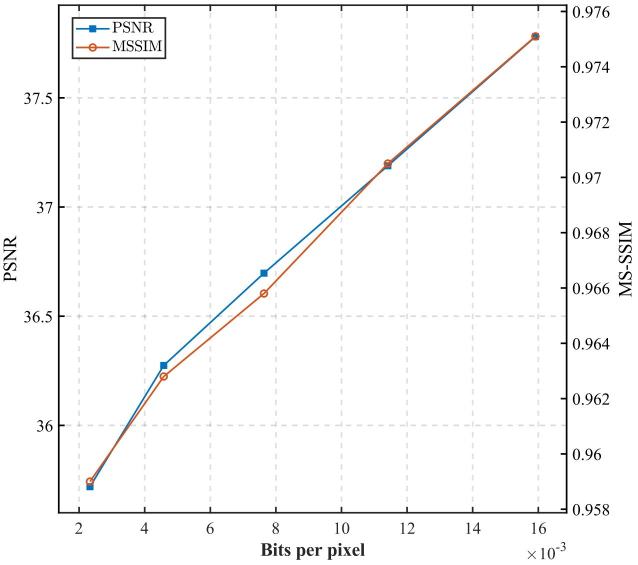

Figure 3. **Effect of number of parameters on performance.** The plot shows an increasing trend in reconstruction performance as the parameter count grows. This finding demonstrates that Act-INR scales effectively with parameter size.

Video regression. In Tab. 1, we present a comprehensive evaluation of video regression on the UVG dataset, detailing all PSNR and MS-SSIM metrics obtained. Although the Act-INR architecture is specifically designed for video processing applications, it demonstrates competitive performance compared to state-of-the-art neural representations that are explicitly tailored for video representation tasks.

Failure case. In challenging scenarios such as Jockey, we observe that transition regions across patches are not well reconstructed when objects cross patch boundaries. For example, as shown in Fig. 4, the horse's foreleg remains on the boundary between two patches during its motion. When a patch boundary splits an object into separate parts, the model may struggle to reconstruct these regions in a visually plausible manner.

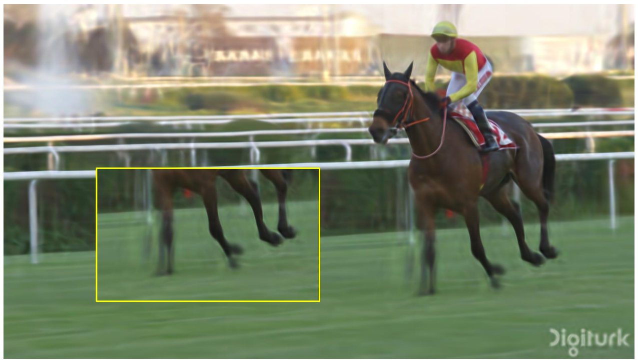

Figure 5. **Results of training with window blending for the same Video in Fig. 4.** Overlapping windows effectively eliminate blocking artifacts and improve reconstruction quality at patch transitions.

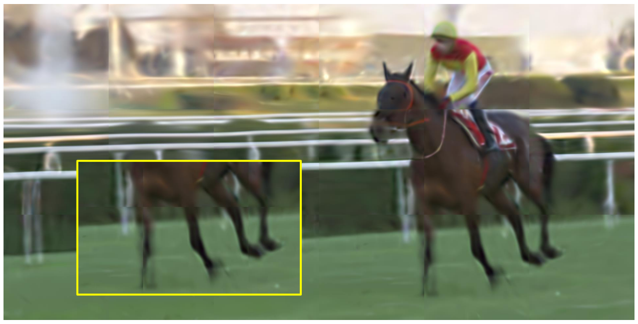

Figure 4. The above illustration highlights a failure case of Act-INR, specifically showing that patch transitions are particularly susceptible to reconstruction artifacts.

Remedy for patch transitions. To address artifacts arising at patch boundaries when objects cross over patch edges, we propose employing overlapping windows and blending them using a specific strategy, albeit at an increased computational cost. Drawing on the methodology described in Mod-SIREN [2], pixels are weighted either linearly or bilinearly depending on the number of overlapping patches. Specifically, bilinear blending is applied when a point is surrounded by four windows, whereas linear blending is used when the point is encompassed by two windows. This tech-

Video	Yacht	Ready	Jockey	Bosph
ActINR-4K	30.1	25.7	24.0	37.2
ActINR-HD	30.3	25.9	24.1	37.3

Table 2. Interpolation performance on 4K and HD resolution

Method	Encoding Time	Decoding FPS
ActINR	5h19m	59.28
FF-NeRV	6h0m	84.00
HNeRV-Boost	1h53m	13.15

Table 3. Per-video encoding speed, and decoding performance

nique effectively mitigates artifacts at patch boundaries, resulting in smoother reconstructions as shown in Fig. 5.

Resilience to higher resolution. We evaluate the resolution invariance of our interpolation method on four 4K-resolution videos from the UVG dataset. As presented in Tab. 2, our method maintains image quality even as the video resolution increases.

Further motivation for bias-motion interplay. To emphasize the relationship between biases and motion, we took video of a vibrating tuning fork, and fit our ActINR to it. We then median filtered the time series of biases across time to smoothen the motion artifacts, as shown in Fig. 6. As evident, the high frequency parts around the fork are replaced by a near-static prongs, thereby underscoring the relationship between biases and motions.

Encoding and Decoding speeds. Table. 3 tabulates encoding and decoding times for various approaches. Our encoding and decoding times are comparable to previous approaches, with a performance that is similar to FF-NeRV while decoding 4.5× faster than HNeRV-Boost and operating in real time, unlike HNeRV-Boost. We achieve speed up by evaluating all windows in a frame at a time (pytorch `bmm`). We also attribute the speed-up to initializing each group with the preceding group's weights, reducing the training epochs needed for convergence.

Further implementation details. We reproduced FF-NeRV using a 3-million-parameter model for a fair comparison, since the original FF-NeRV reported PSNR only for its 12-million-parameter version. Following the official repository, we kept all hyperparameters unchanged except for model size. Our approach consistently employs a 3-million-parameter model for both interpolation and denoising tasks. However, for inpainting, we used triple the capacity to handle missing regions more effectively.

Compression. For a fairer comparison, we evaluated the compression performance of Act-INR with simple pruning and 8-bit post-quantization. Fig. 7 shows a rate distortion curve illustrating that Act-INR performs considerably better than JPEG2000, although, slightly worse than Hybrid NeRV. These analyses demonstrate that ActINR primarily excels in solving inverse problems (interpolation, super resolution, denoising, and inpainting) while state-of-the-art techniques like Hybrid NeRV [1] are better suited for compression tasks.

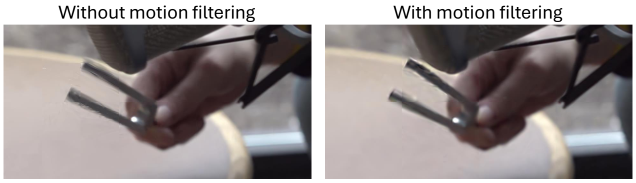

Figure 6. Video of a vibrating tuning fork. Upon applying a median filter to the biases, the pronounced shaking is significantly mitigated, as evident in the image on the right.

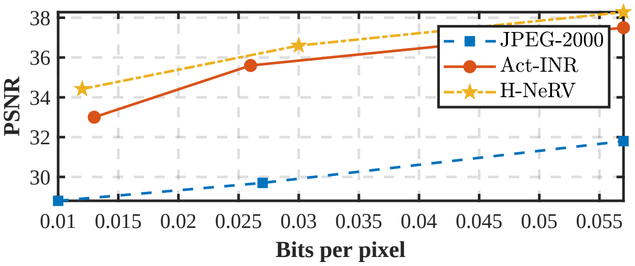

Figure 7. Rate-Distortion curve for Bosphorus

References

[1] Hao Chen, Matthew Gwilliam, Ser-Nam Lim, and Abhinav Shrivastava. Hnerv: A hybrid neural representation for videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10270–10279, 2023. 3

[2] Ishit Mehta, Michaël Gharbi, Connelly Barnes, Eli Shechtman, Ravi Ramamoorthi, and Manmohan Chandraker. Modulated periodic activations for generalizable local functional representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14214–14223, 2021. 2