RobustFormer: Noise-Robust Pre-training for Images and Videos (Supplementary)

Ashish Bastola[1,*], Nishant Luitel[2,*], Hao Wang[1], Danda Pani Paudel[2], Roshni Poudel[2], Abolfazl Razi[1]
[1]Clemson University [2]NAAMII
{abastol,hao9,arazi}@clemson.edu, {nishant.luitel,roshani.poudel,danda.paudel}@naamii.org.np

1. Overview

This supplementary includes additional results that were not included in the main paper.

- Section [2] includes the details of both image noise types [2.1] and video noise types [2.2] and the severity details [2.3.1].
- In Section [3] we show the comparison of RF-variants with the ImageMAE on several custom noise types as mentioned in [2.3.1].
- In Section [4] we show the comparison results with VideoMAE [4.1] and perform ablation experiments [4.3] with and without IDWT on UCF-101 dataset as well as all of the RF variants that additionally provide ablation experiments for using our proposed dwt-based attention module on benchmark noise types mentioned in [2]. We also show the comparison of our RF-variants on our custom noise types in 4.4 for UCF-101. We also show comprehensive robustness score comparison for all RF variants and VideoMAE in 4.2.
- In Section [5] we show the comparision results with VideoMAE [5.1]. We also show evaluation on benchmark noise types in [5.2] and custom noise types in [5.3].

2. Noise Implementation Details

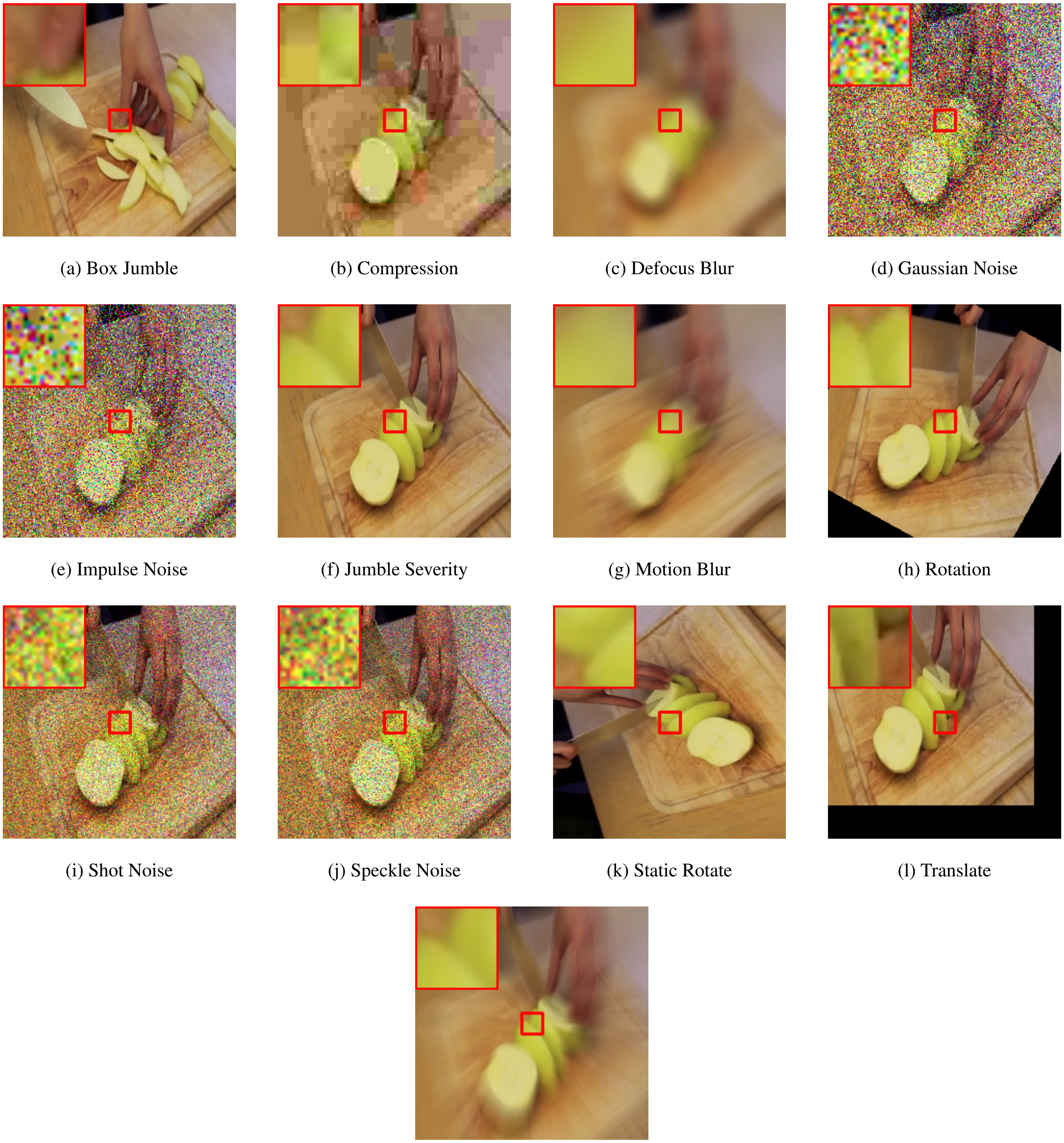
2.1. Image Noise types

The corruptions used in images were the standard corruptions types as defined by [1] in Imagenet-C and Imagenet-P. The Imagenet-C contains a total of 75 corruptions with 15 different corruption types each consisting of 5 severity levels. The corruption types include Noise(Gaussian, Shot, Impulse), Blur(Defocus, Glass, Motion, Zoom), Weather(Snow, Frost, Fog, Bright) and Digital(Contrast, Elastic, Pixel, JPEG). Similarly, the Imagenet-P contains 10 distinct type of perturbations: Noise(Gaussian, Shot), Blur(Motion, Zoom), Weather(Snow, Bright) and Digital(Translate, Rotate, Tilt and Scale); where each perturbation is used to generate more than 30 frames.

2.2. Video Noise Types

The perturbations we used in Video noises include: Defocus Blur, Motion Blur, Zoom Blur, Gaussian, Shot, Impulse, Speckle, Compression, Static Rotate, Rotate, Translate, Jumbling and Box Jumbling. An example image for each noise at maximum severity level (5) is shown in Figure 1. These perturbations represent different categories: Blur(Defocus, Motion and Zoom blur), Noise(Gauss, Shot, Impulse, Speckle), Temporal(Jumbling, Box Jumbling), Digital(Compression) and Camera(Static Rotate, Rotate, Translate). These noise types are derived from a large scale video robust benchmarking analysis [2].

[*] Equal contribution



(a) Box Jumble

(b) Compression

(c) Defocus Blur

(d) Gaussian Noise

(e) Impulse Noise

(f) Jumble Severity

(g) Motion Blur

(h) Rotation

(i) Shot Noise

(j) Speckle Noise

(k) Static Rotate

(l) Translate

(m) Zoom Blur

Figure 1. Examples of various noise types applied to video clips. Each image corresponds to a specific noise type at severity level 5.

2.3. Additional Noise types

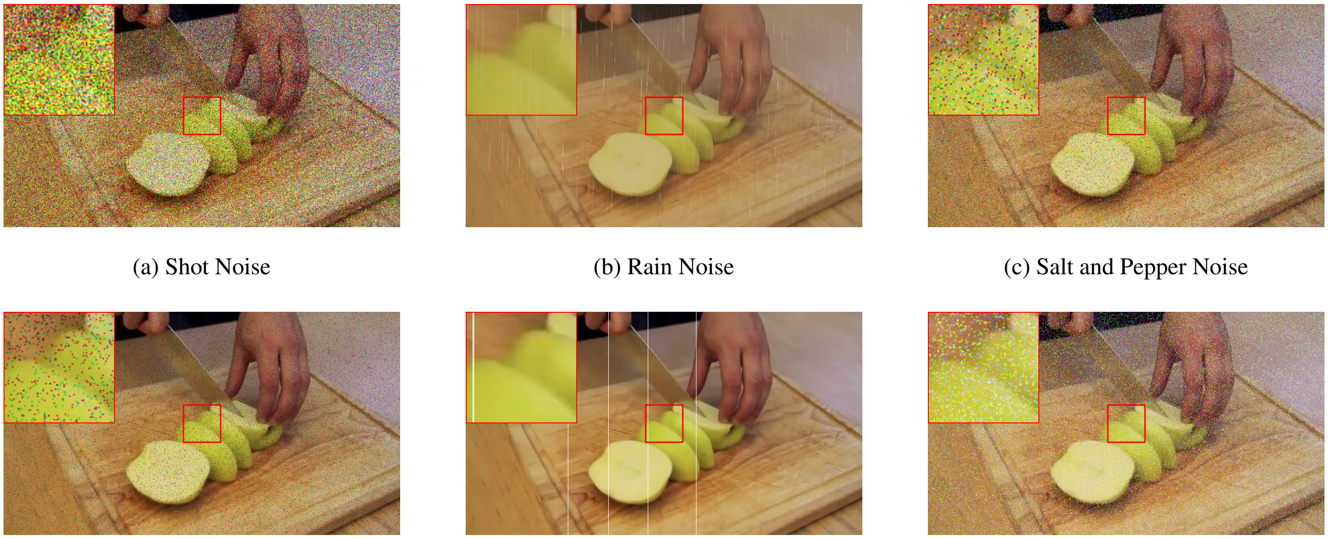
We also used 6 additional type of noises: shot, rain, salt and pepper, missing sample, packet-loss and tampering noise similar to [3]. These additional noises are shown in Figure 2. We calculated the PSNR and SSIM values for these additional noise types as shown in table 1. These noise types are different from the ones included in [2] in terms of their implementation details and intensities. Thus we consider them as a distinct noise types.

2.3.1 Noise Severity Details

Here we mention the noise severity levels of our custom noise types. These are calculated for an individual frame in a sequence.

Table 1. PSNR(SSIM) values for different noise types and severity levels for custom noise types

Noise Type	Severity 1	Severity 2	Severity 3	Severity 4	Severity 5
Gaussian	24.74(0.3042)	18.81(0.1134)	15.53(0.0597)	13.37(0.0377)	11.86(0.0267)
Shot Noise	9.90(0.0471)	11.59(0.0582)	14.74(0.0914)	17.61(0.1389)	21.23(0.2305)
Rain	29.06(0.9248)	28.57(0.8855)	25.18(0.8305)	22.75(0.7884)	20.77(0.7354)
Tampering	17.70(0.2229)	14.69(0.1003)	12.94(0.0665)	11.68(0.0512)	10.72(0.0425)
Packet Loss	29.34(0.9728)	25.32(0.9448)	23.37(0.9223)	24.60(0.9395)	23.78(0.9298)
Missing Sample	22.97(0.5561)	19.97(0.3543)	18.19(0.2566)	16.93(0.2065)	15.97(0.1771)
Salt & Pepper	18.37(0.2296)	15.34(0.0870)	13.65(0.0510)	12.46(0.0355)	11.53(0.0270)
Speckle	22.52(0.3255)	16.85(0.1685)	13.80(0.1168)	12.01(0.0939)	10.84(0.0815)



(a) Shot Noise

(b) Rain Noise

(c) Salt and Pepper Noise

(d) Missing Sample Noise

(e) Packet Loss Noise

(f) Tampering Noise

Figure 2. Examples of additional custom noise types applied to RGB video clips. (a) Shot noise results from photon fluctuations, causing random pixel intensity variations. (b) Rain noise mimics water droplets, creating streaks and distortions. (c) Salt and pepper noise features randomly scattered bright and dark pixels. (d) Missing sample noise highlights unprocessed regions, while (e) packet loss leads to visual gaps from data transmission errors. (f) Tampering noise refers to unauthorized modifications that distort the original content. Video frames from all the noise types were generated at severity level 2.

3. Evaluation on ImagenetTiny200

In this section we show that in case of small dataset like ImagenetTiny-200 the regular DWT-omit model(RF-O) significantly outperformed all of the other RF-variants. The figures 3, 4 show the Top-1 and Top-5 accuracy of this comparison across our custom noise types mentioned in 1. This shows that with very scarce data to train on, completely removing noise is the best option to maintain model robustness. In this case we consider ImageMae as our comparison baseline.

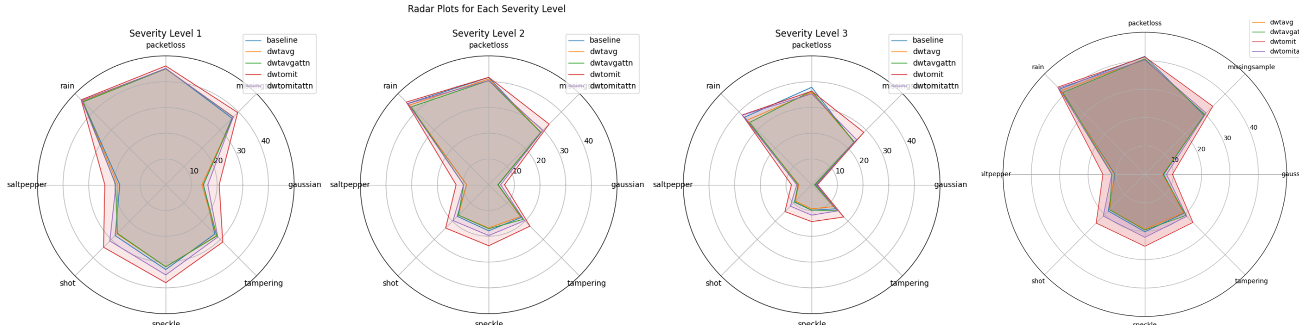


Figure 3. fig: Top-1 Accuracy comparison of first 3 severity levels across different Noisy transformations.

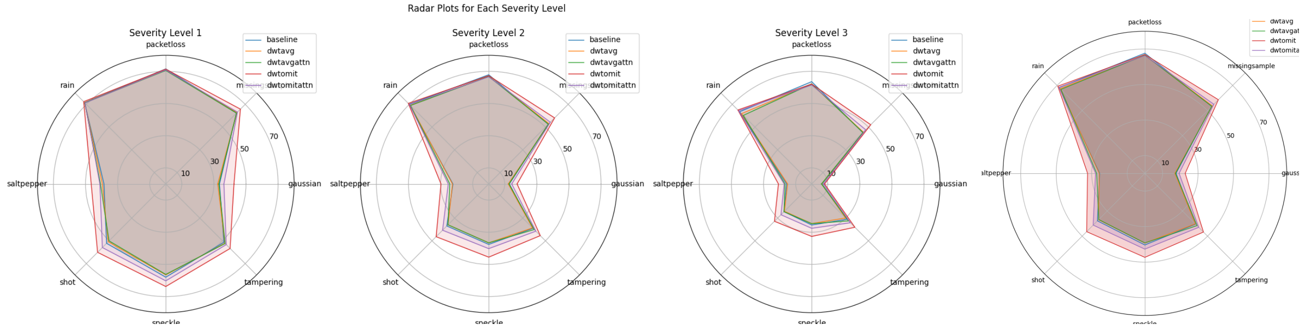


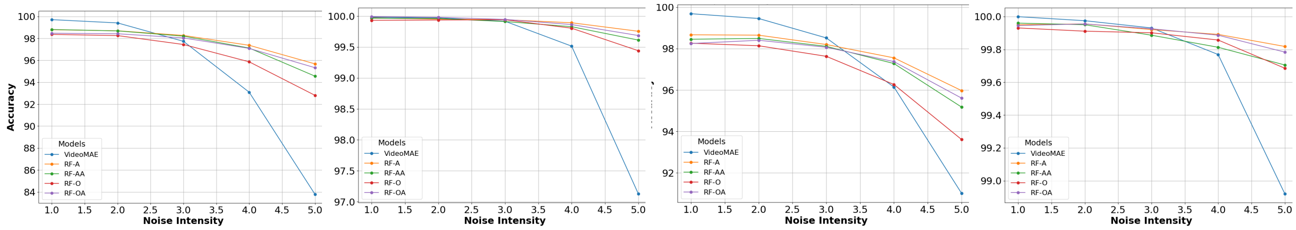
Figure 4. fig: Top-5 Accuracy comparison of first 3 severity levels across different Noisy transformations.

4. Evaluation on UCF-101 Dataset

In this section we perform ablation experiments for both IDWT and our dwt based attention module. We also show that the RF variants outperform regular VideoMAE architecture in various noise types.

4.1. Comparison with VideoMAE

In this section we compare various Robustformer (RF) variants with the Video-MAE architecture for UCF-101 datasets. Our results indicate that the RF-variants outperform the Video-MAE with a significant factor especially at high noise severity levels.



(a) Top-1 Accuracy for Defocus Blur (b) Top-5 Accuracy for Defocus Blur (c) Top-1 Accuracy for Motion Blur (d) Top-5 Accuracy for Motion Blur

Figure 5. Comparison of Top-1 and Top-5 Accuracy for different severity levels across Defocus Blur and Motion Blur for UCF-101 dataset

4.2. Absolute and Relative Robustness Scores

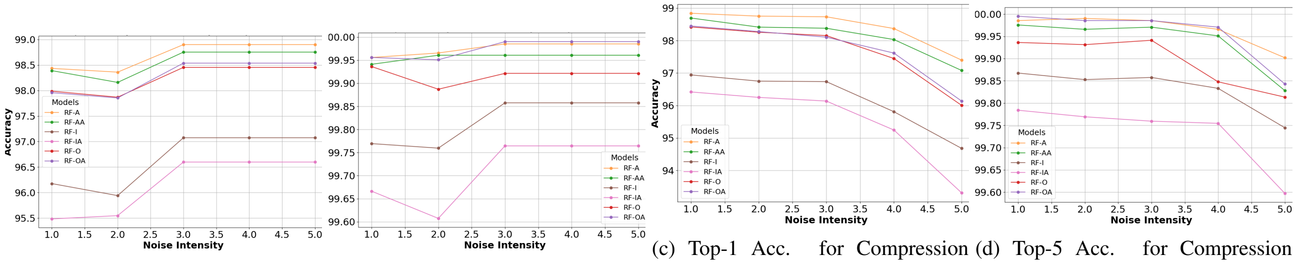
4.3. Ablation Experiments on Benchmark Noise Types

In this section we compare various RF variants on benchmark noise types presented in [2] in UCF-101 datasets. We compare methods with both IDWT and without IDWT and methods that use our DWT based attention and those that do not. On average we can see that the RF-variants with dwt attention outperform regular DWT based methods. We can also see

Table 2. Comparison of robustness of Video Classification models on 14 individual noise types and 5 different corruption categories: Noise, Blur, Temporal, Digital and Camera [2] evaluated on UCF-101P benchmark dataset. γ^a and γ^r are absolute and relative robustness scores of the models averaged across all the severity level. The best models are marked as **BOLD** for each corruption category. For both, higher is better.

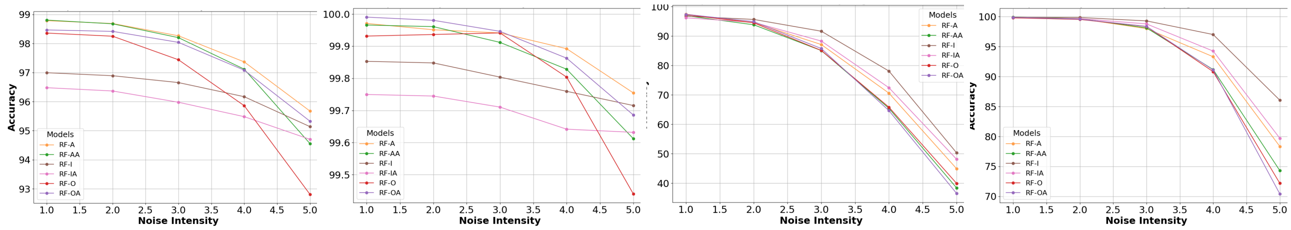
Perturbation	R3D		I3D		SF		X3D		MViT		Times		VideoMAE		RF-A		RF-AA		RF-O		RF-OA		RF-C		RF-CA	
	γ^a	γ^r	γ^a	γ^r	γ^a	γ^r	γ^a	γ^r	γ^a	γ^r	γ^a	γ^r	γ^a	γ^r	γ^a	γ^r	γ^a	γ^r	γ^a	γ^r	γ^a	γ^r	γ^a	γ^r	γ^a	γ^r
Defocus Blur	.63	.37	.74	.56	.72	.61	.84	.73	.74	.63	.79	.78	0.95	0.94	0.99	0.99	0.99	0.99	0.98	0.98	0.99	0.99	0.98	0.97	0.97	0.96
Motion Blur	.74	.56	.75	.57	.65	.51	.82	.69	.71	.58	.88	.88	0.97	0.97	0.99	0.99	0.99	0.99	0.98	0.98	0.99	0.99	0.99	0.99	0.98	0.98
Zoom Blur	.77	.61	.79	.64	.82	.74	.94	.89	.88	.82	.84	.83	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	0.99	0.99	0.99	0.99
Blur	.71	.51	.76	.59	.73	.62	.86	.76	.78	.67	.84	.83	0.97	0.97	**0.99**	**0.99**	**0.99**	**0.99**	9.98	0.98	**0.99**	**0.99**	**0.99**	0.98	0.98	0.98
Gaussian	.57	.28	.64	.39	.52	.33	.59	.30	.78	.68	.62	.61	0.81	0.65	0.80	0.64	0.77	0.54	0.78	0.57	0.77	0.52	0.80	0.65	0.77	0.51
Shot	.75	.58	.83	.71	.73	.63	.78	.62	.94	.91	.92	.92	0.98	0.98	0.97	0.96	0.97	0.96	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.96
Impulse	.55	.23	.59	.31	.57	.25	.58	.28	.76	.65	.60	.59	0.80	0.63	0.79	0.61	0.76	0.51	0.77	0.55	0.76	0.49	0.79	0.63	0.75	0.47
Speckle	.72	.52	.79	.63	.70	.57	.74	.55	.91	.87	.90	.90	0.97	0.97	0.96	0.95	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.95
Noise	.65	.40	.71	.51	.60	.45	.67	.44	.85	.78	.76	.75	**0.89**	0.81	0.88	0.79	0.87	0.74	0.87	0.76	0.87	0.74	0.88	0.80	0.86	0.77
Compression	.92	.86	.93	.89	.90	.86	.96	.93	.96	.94	.92	.91	**1.00**	**1.00**	**1.00**	**1.00**	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	**1.00**	**1.00**
Variable Rotation	.88	.79	.85	.74	.57	.39	.75	.58	.63	.45	.97	.97	0.86	0.83	0.81	0.76	0.81	0.74	0.84	0.80	0.83	0.79	0.82	0.76	0.83	0.78
Static Rotation	.75	.57	.76	.59	.63	.62	.77	.61	.80	.71	.88	.88	0.82	0.42	0.81	0.50	0.82	0.53	0.82	0.62	0.81	0.57	0.82	0.59	0.82	0.58
Translate	.92	.87	.88	.8	.67	.53	.82	.70	.65	.48	.98	.98	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00
Camera	.85	.74	.83	.71	.65	.51	.78	.63	.69	.55	**.95**	**.95**	0.89	0.75	0.87	0.75	0.87	0.75	0.88	0.80	0.88	0.78	0.88	0.78	0.88	0.79
Jumbling	.96	.94	.96	.93	.78	.70	.93	.88	.80	.70	.98	.98	0.98	0.98	0.96	0.96	0.96	0.96	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
Box Jumbling	.98	.96	.98	.95	.95	.92	.98	.94	.95	.91	.97	.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Temporal	.95	.92	.95	.92	.86	.81	.94	.90	.81	.73	.98	.98	**0.99**	**0.99**	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98

that IDWT based methods are very close to the RF-A and RF-AA variants and in many cases perform significantly poorly compared to the RF variants.



(a) Top-1 Acc. for Box Jumble noise (b) Top-5 Acc. for Box Jumble noise (c) Top-1 Acc. for Compression noise (d) Top-5 Acc. for Compression noise

Figure 6. Comparison of Top-1 and Top-5 Accuracy for different severity levels across Box jumble noise and Compression noise for UCF-101 dataset

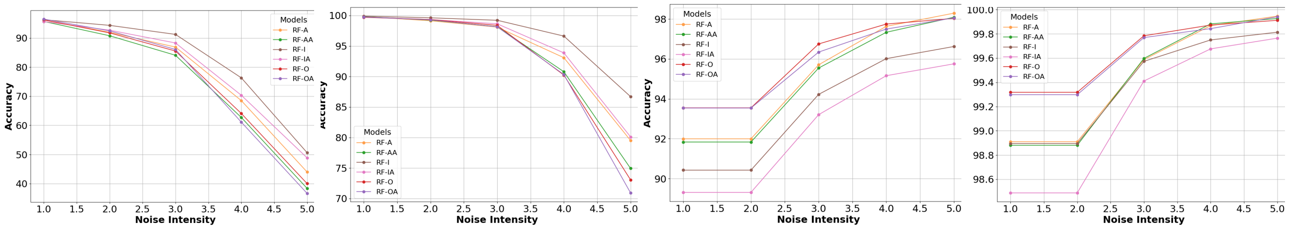


(a) Top-1 Accuracy for Defocus Blur (b) Top-5 Acc. for Defocus Blur (c) Top-1 Acc. for Gaussian noise (d) Top-5 Acc. for Gaussian noise

Figure 7. Comparison of Top-1 and Top-5 Accuracy for different severity levels across Defocus Blur and Gaussian noise for UCF-101 dataset

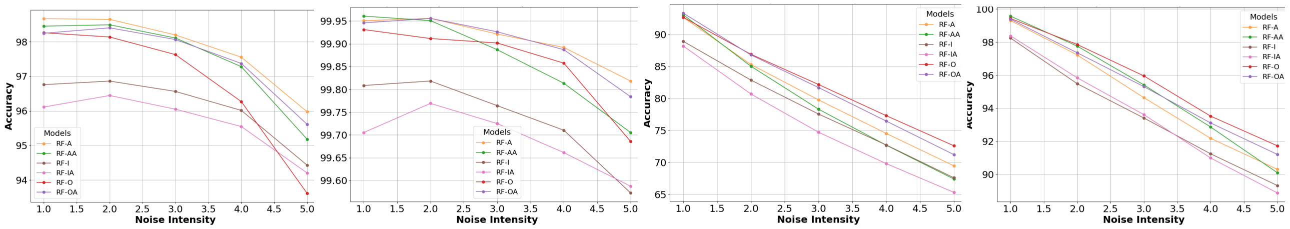
4.4. Evaluation on Custom Noise

In this section we evaluate different RF variants on our custom noise types and intensities. Many noise types in this evaluation were not included in the original benchmark noise types from [2]. For few noise types that are similar to the benchmark noise types, we vary their intensity level across those noise types as mentioned in 1 for more intense evaluations.



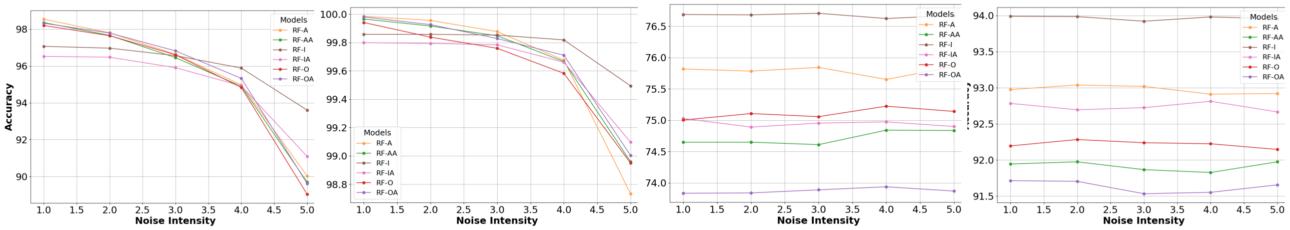
(a) Top-1 Accuracy for Impulse noise (b) Top-5 Acc. for Impulse noise (c) Top-1 Acc. for Jumble noise (d) Top-5 Acc. for Jumble noise

Figure 8. Comparison of Top-1 and Top-5 Accuracy for different severity levels across Impulse noise and Jumble noise for UCF-101 dataset



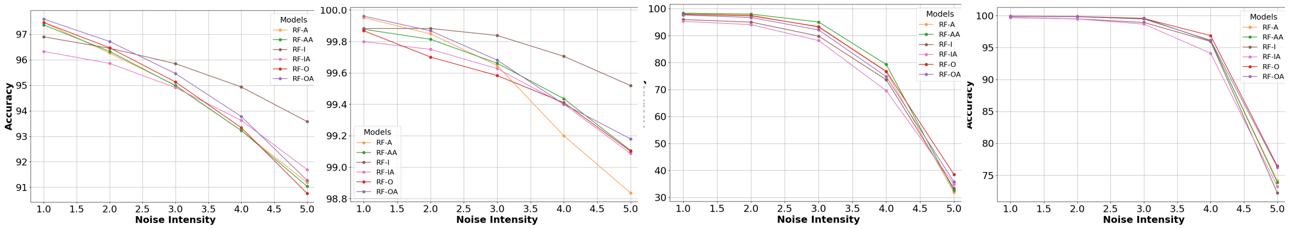
(a) Top-1 Accuracy for Motion blur (b) Top-5 Acc. for Motion blur (c) Top-1 Acc. for Rotate (d) Top-5 Acc. for Rotate

Figure 9. Comparison of Top-1 and Top-5 Accuracy for different severity levels across Motion blur and Rotate for UCF-101 dataset



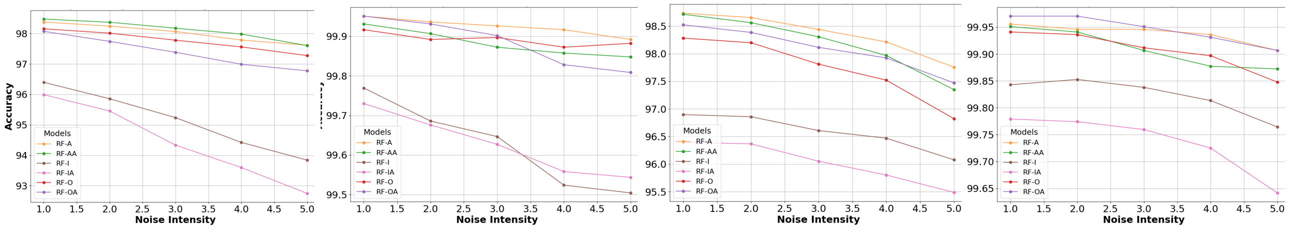
(a) Top-1 Accuracy for Shot Noise (b) Top-5 Acc. for Shot Noise (c) Top-1 Acc. for Spatial noise (d) Top-5 Acc. for Spatial noise

Figure 10. Comparison of Top-1 and Top-5 Accuracy for different severity levels across Shot noise and Spatial noise for UCF-101 dataset



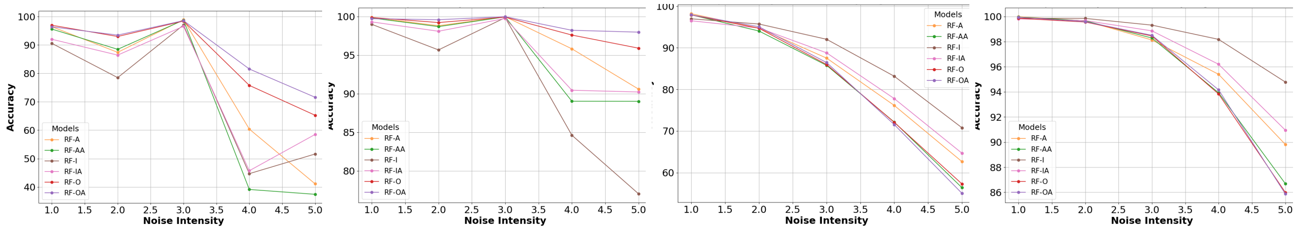
(a) Top-1 Accuracy for Speckle noise (b) Top-5 Acc. for Speckle noise (c) Top-1 Acc. for Static rotate (d) Top-5 Acc. for Static rotate

Figure 11. Comparison of Top-1 and Top-5 Accuracy for different severity levels across Speckle noise and Static Rotate for UCF-101 dataset



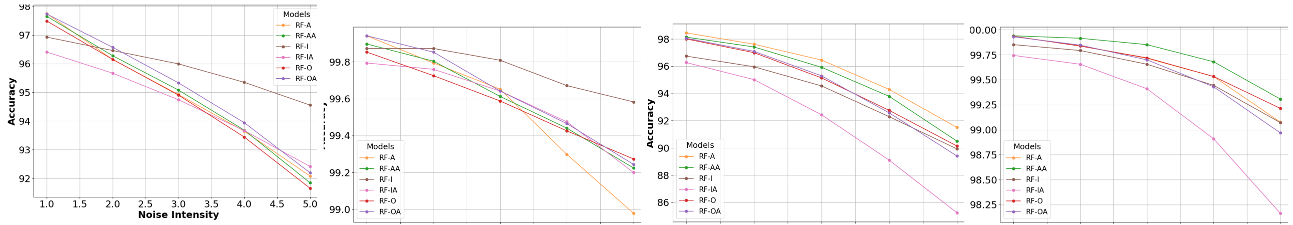
(a) Top-1 Accuracy for Translate (b) Top-5 Acc. for Zoom blur (c) Top-1 Acc. for Translate (d) Top-5 Acc. for Zoom blur

Figure 12. Comparison of Top-1 and Top-5 Accuracy for different severity levels across Translate and Zoom blur for UCF-101 dataset



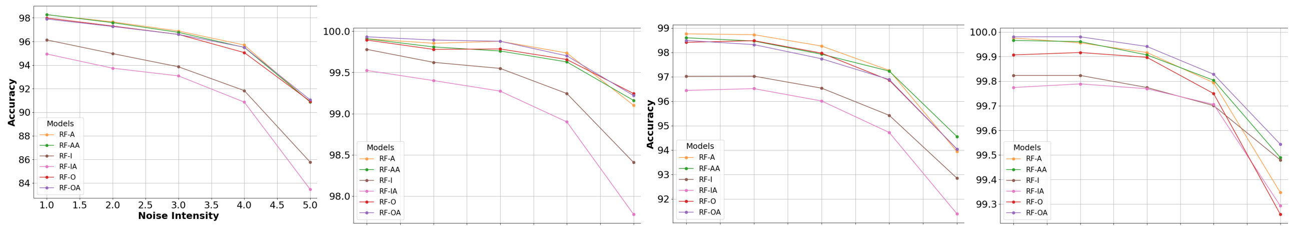
(a) Top-1 Accuracy for Flickr (b) Top-5 Acc. for Flickr (c) Top-1 Acc. for Gaussian Noise (d) Top-5 Acc. for Gaussian Noise

Figure 13. Comparison of Top-1 and Top-5 Accuracy for different severity levels across Flickr and Gaussian noise for UCF-101 dataset



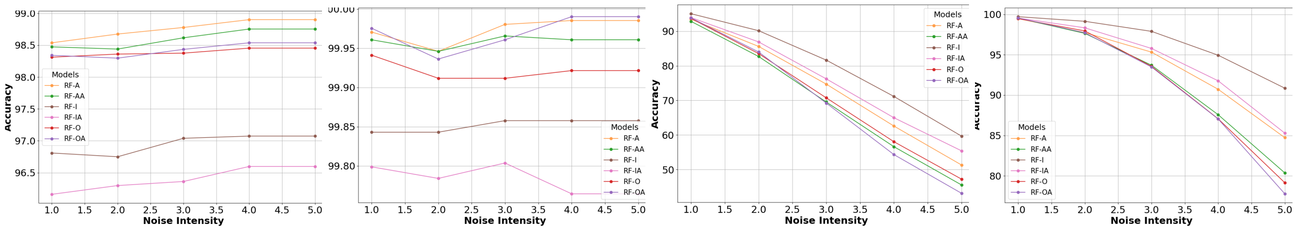
(a) Top-1 Accuracy for Missing Sample (b) Top-5 Acc. for Missing Sample (c) Top-1 Acc. for Packet Loss (d) Top-5 Acc. for Packet Loss

Figure 14. Comparison of Top-1 and Top-5 Accuracy for different severity levels across Missing sample and Packet loss for Kinetics-400 dataset



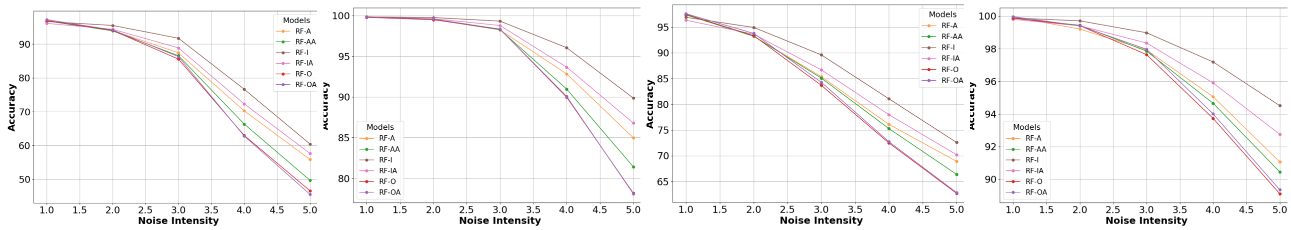
(a) Top-1 Accuracy for Quantization noise (b) Top-5 Acc. for Quantization noise (c) Top-1 Acc. for Packet Loss (d) Top-5 Acc. for Packet Loss

Figure 15. Comparison of Top-1 and Top-5 Accuracy for different severity levels across Quantization and Rain noise for Kinetics-400 dataset



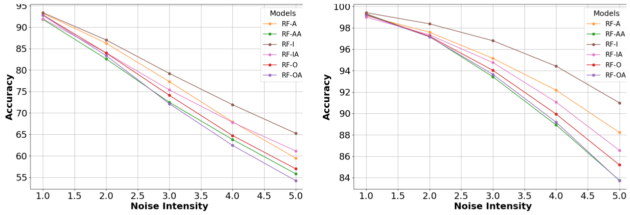
(a) Top-1 Accuracy for Random Blur (b) Top-5 Acc. for Random Blur (c) Top-1 Acc. for Packet Loss (d) Top-5 Acc. for Salt-pepper

Figure 16. Comparison of Top-1 and Top-5 Accuracy for different severity levels across Random Blur and Salt-pepper for Kinetics-400 dataset



(a) Top-1 Accuracy for Shot noise (b) Top-5 Acc. for Shot noise (c) Top-1 Acc. for Speckle noise (d) Top-5 Acc. for Speckle noise

Figure 17. Comparison of Top-1 and Top-5 Accuracy for different severity levels across Random Blur and Salt-pepper for Kinetics-400 dataset



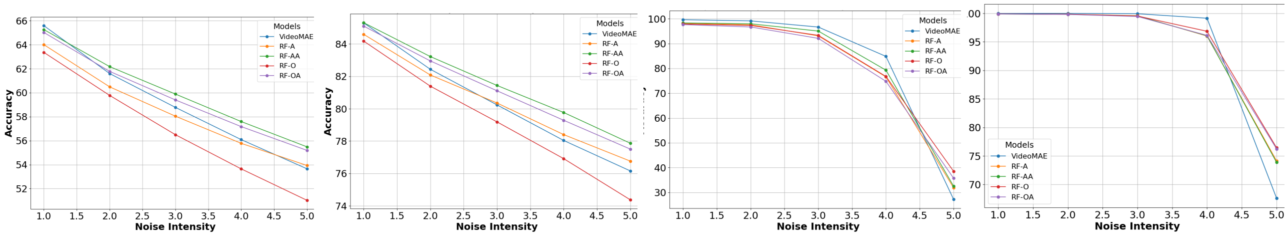
(a) Top-1 Acc. for Tampering noise (b) Top-5 Acc. for Tampering noise

Figure 18. Comparison of Top-1 and Top-5 Accuracy for different severity levels across Random Blur and Salt-pepper for Kinetics-400 dataset

5. Evaluation on Kinetics-400

5.1. Comparision with VideoMAE

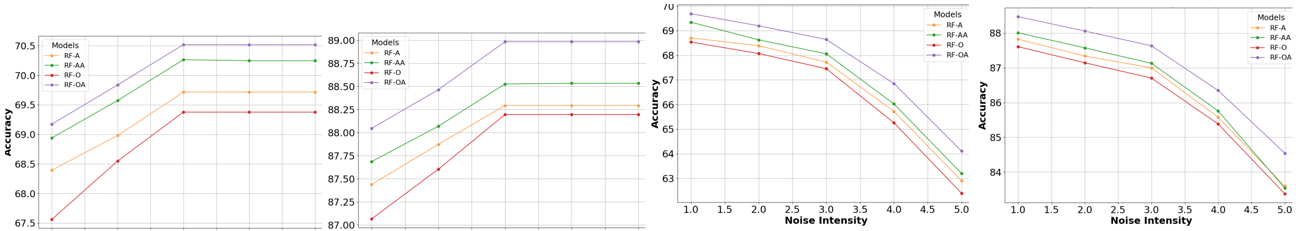
In this section we show RF variants perform very closely and in many cases outperform VideoMAE in rotate and static roate noise types.



(a) Top-1 Acc. for rotation (b) Top-1 Acc. for rotation (c) Top-1 Acc. for static rotate (d) Top-1 Acc. for static rotate

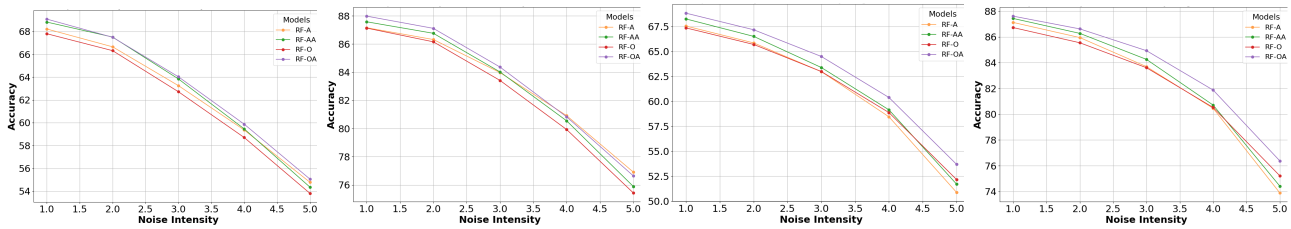
Figure 19. Comparison of Top-1 and Top-5 Accuracy for different severity levels across rain and packet loss noise for Kinetics-400 dataset

5.2. Evaluation on Benchmark noise Types



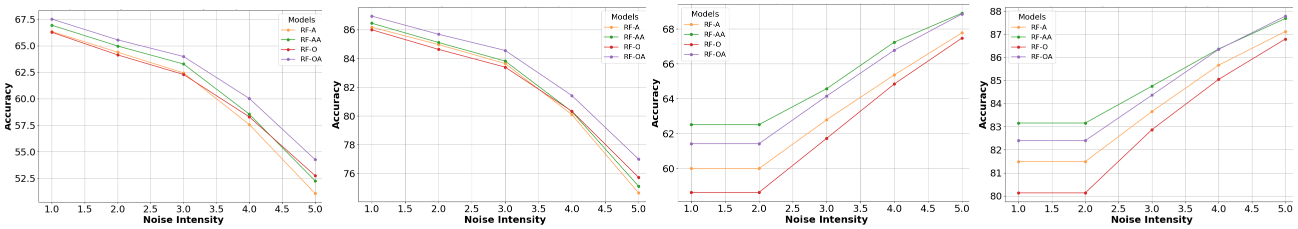
(a) Top-1 Acc. for Box Jumble noise (b) Top-5 Acc. for Box Jumble noise (c) Top-1 Acc. for Compression noise (d) Top-5 Acc. for Compression noise

Figure 20. Comparison of Top-1/Top-5 accuracy across severity levels for Box-Jumble and Compression noise on Kinetics-400.



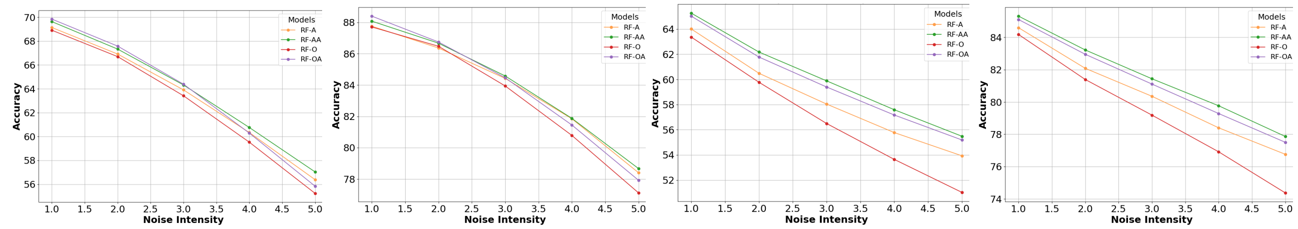
(a) Top-1 Accuracy for Defocus Blur (b) Top-5 Acc. for Defocus Blur (c) Top-1 Acc. for Gaussian noise (d) Top-5 Acc. for Gaussian noise

Figure 21. Comparison of Top-1 and Top-5 Accuracy across severity levels for Defocus Blur and Gaussian noise on Kinetics-400 dataset



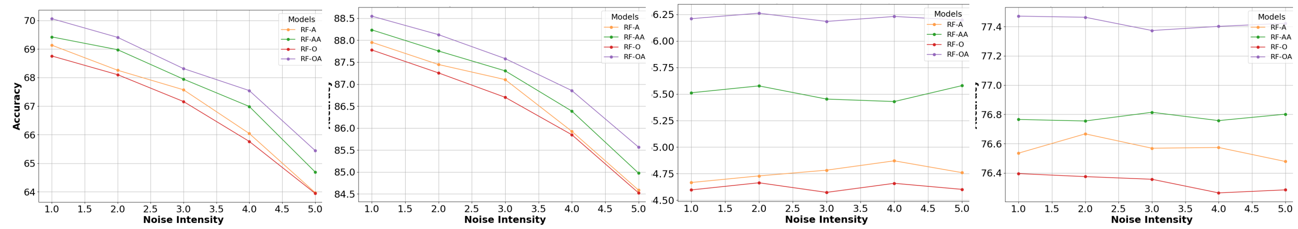
(a) Top-1 Accuracy for Impulse noise (b) Top-5 Acc. for Impulse noise (c) Top-1 Acc. for Jumble noise (d) Top-5 Acc. for Jumble noise

Figure 22. Comparison of Top-1 and Top-5 Accuracy across severity levels for Impulse noise and Jumble noise on Kinetics-400 dataset



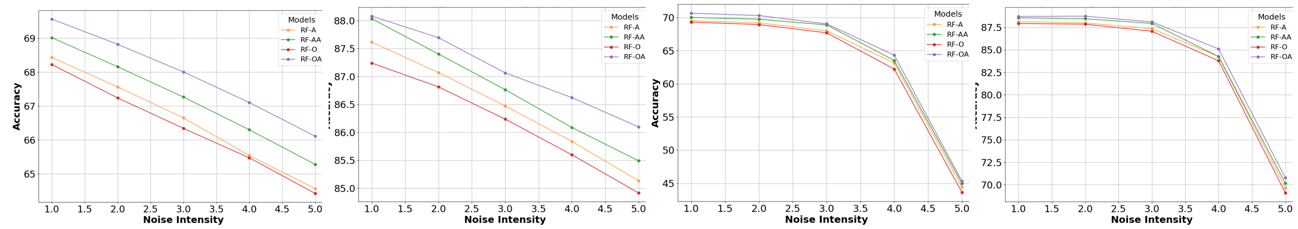
(a) Top-1 Accuracy for Motion blur (b) Top-5 Acc. for Motion blur (c) Top-1 Acc. for Rotate (d) Top-5 Acc. for Rotate

Figure 23. Comparison of Top-1 and Top-5 Accuracy for different severity levels across Motion blur and Rotate for Kinetics-400 dataset



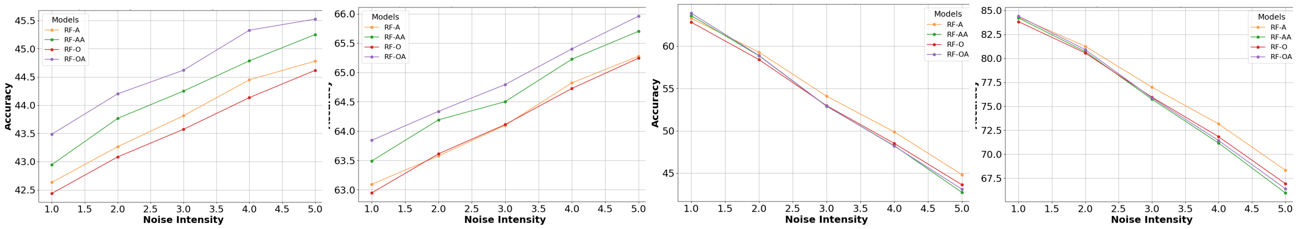
(a) Top-1 Accuracy for Shot Noise (b) Top-5 Acc. for Shot Noise (c) Top-1 Acc. for Spatial noise (d) Top-5 Acc. for Spatial noise

Figure 24. Comparison of Top-1 and Top-5 Accuracy for different severity levels across Shot noise and Spatial noise for Kinetics-40 dataset



(a) Top-1 Accuracy for Speckle noise (b) Top-5 Acc. for Speckle noise (c) Top-1 Acc. for Static rotate (d) Top-5 Acc. for Static rotate

Figure 25. Comparison of Top-1 and Top-5 Accuracy for different severity levels across Speckle noise and Static Rotate for Kinetics-400 dataset

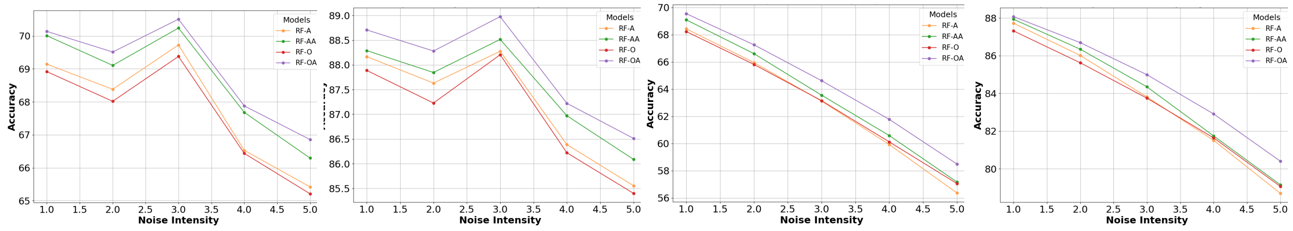


(a) Top-1 Accuracy for Translate (b) Top-5 Acc. for Zoom blur (c) Top-1 Acc. for Translate (d) Top-5 Acc. for Zoom blur

Figure 26. Comparison of Top-1 and Top-5 Accuracy for different severity levels across Translate and Zoom blur for Kinetics-400 dataset

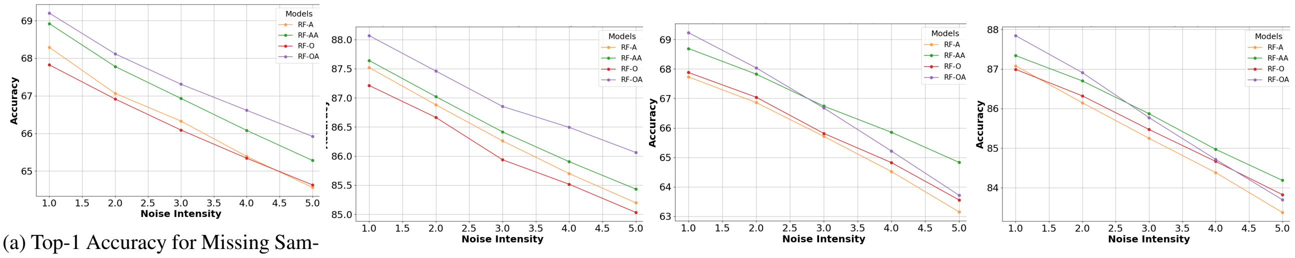
5.3. Evaluation on Custom Noise Types and Intensities

Here we present the evaluation on Kinetics-400 dataset perturbed with additional noise types mentioned in 1.



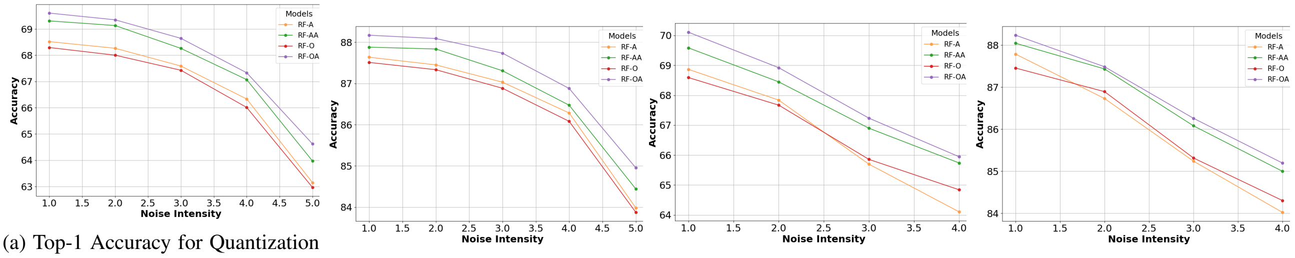
(a) Top-1 Accuracy for Flickr (b) Top-5 Acc. for Flickr (c) Top-1 Acc. for Gaussian Noise (d) Top-5 Acc. for Gaussian Noise

Figure 27. Comparison of Top-1 and Top-5 Accuracy for different severity levels across Flickr and Gaussian noise for Kinetics-400 dataset



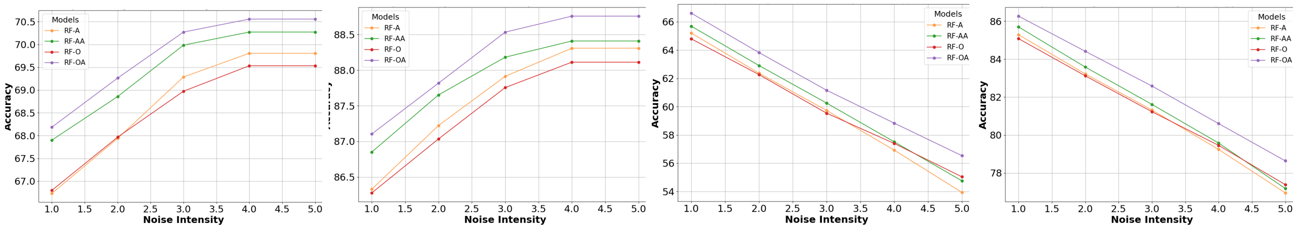
(a) Top-1 Accuracy for Missing Sample (b) Top-5 Acc. for Missing Sample (c) Top-1 Acc. for Packet Loss (d) Top-5 Acc. for Packet Loss

Figure 28. Comparison of Top-1 and Top-5 Accuracy for different severity levels across Missing sample and Packet loss for Kinetics-400 dataset



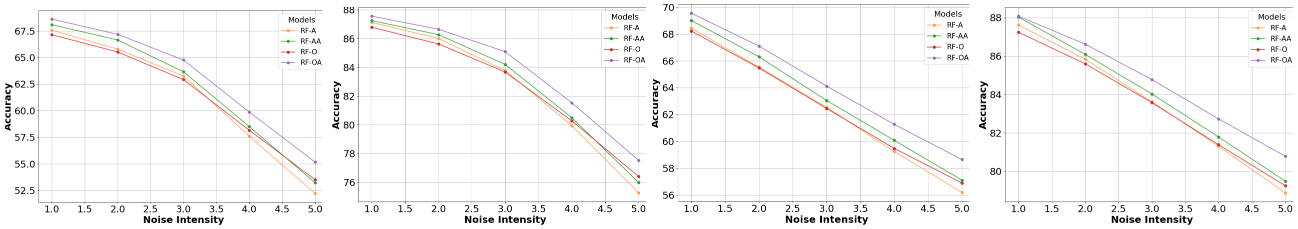
(a) Top-1 Accuracy for Quantization noise (b) Top-5 Acc. for Quantization noise (c) Top-1 Acc. for Packet Loss (d) Top-5 Acc. for Packet Loss

Figure 29. Comparison of Top-1 and Top-5 Accuracy for different severity levels across Quantization and Rain noise for Kinetics-400 dataset



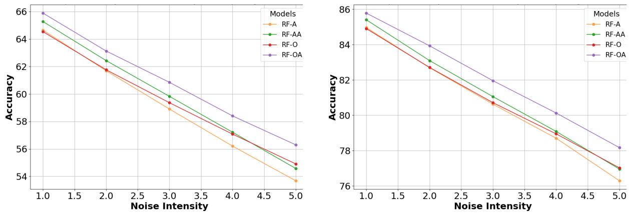
(a) Top-1 Accuracy for Random Blur (b) Top-5 Acc. for Random Blur (c) Top-1 Acc. for Packet Loss (d) Top-5 Acc. for Salt-pepper

Figure 30. Comparison of Top-1 and Top-5 Accuracy for different severity levels across Random Blur and Salt-pepper for Kinetics-400 dataset



(a) Top-1 Accuracy for Shot noise (b) Top-5 Acc. for Shot noise (c) Top-1 Acc. for Speckle noise (d) Top-5 Acc. for Speckle noise

Figure 31. Comparison of Top-1 and Top-5 Accuracy for different severity levels across Random Blur and Salt-pepper for Kinetics-400 dataset



(a) Top-1 Acc. for Tampering noise (b) Top-5 Acc. for Tampering noise

Figure 32. Comparison of Top-1 and Top-5 Accuracy for different severity levels across Random Blur and Salt-pepper for Kinetics-400 dataset

References

[1] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 1

[2] Madeline Chantry Schiappa, Naman Biyani, Prudvi Kamtam, Shruti Vyas, Hamid Palangi, Vibhav Vineet, and Yogesh S Rawat. A large-scale robustness analysis of video action recognition models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14698–14708, 2023. 1, 3, 4, 5

[3] Chenyu Yi, Siyuan Yang, Haoliang Li, Yap-peng Tan, and Alex Kot. Benchmarking the robustness of spatial-temporal models against corruptions. *arXiv preprint arXiv:2110.06513*, 2021. 2