Seeing Through Fog: Towards Fog-Invariant Action Recognition Supplementary Material

Abstract

We provide content omitted from the main paper due to space limitations. Specifically, the supplementary material is organized into three sections: i) additional details for experimental setup and dataset quality Evaluation, ii) additional ablation studies to analyze our model designs, iii) additional results.

1. Additional Experimental Setup and Dataset Quality Evaluation

Implementation Details. We summarize the optimization and data augmentation details of our method for supervised learning in Tab. 1. In addition, our model is implemented using the PyTorch framework and trained on a machine equipped with four NVIDIA RTX 4090 GPUs.

In our two-stage approach, we first apply a defogging method to the entire dataset, including both the training and testing sets, to eliminate visual distortions caused by fog. After this preprocessing step, we train and evaluate OST [1] on the defogged data. Notably, the GFLOPs and parameter counts reported for the two-stage method in Tab. 3 of the main paper represent the combined complexity of both the defogged model and OST.

When performing supervised learning with BIKE [10] on our FogAct dataset, the attribute branch was not utilized. Table 3 of the main paper shows that BIKE achieves a poor Top-1 accuracy of only 13.4% and a Top-5 accuracy of 19.5% on the FogAct dataset. The main issue lies in its algorithm design: it uses video features twice—first as a key to obtain attention weights. However, when the foggy image serves as the key and is combined with the expanded text embedding, the resulting weights become highly inconsistent. This leads to a suboptimal fusion with the image embedding and the text CLS embedding, ultimately degrading classification performance.

Datasets. We introduce our FogAct dataset, comprising 9,724 videos, which include both foggy and their corresponding clean counterparts, forming 4,862 paired video samples. For each action, we randomly allocate 80% of the pairs to the training set, resulting in 3,867 paired training videos, while the remaining 20%, approximately 995 paired videos, are set aside for testing.

Table 1. The hyperparameters of our model for supervised learning.

	Fully-supervised
Optimisation	
Optimizer	AdamW
Optimizer betas	(0.9, 0.999)
Batch size	128
Learning rate schedule	cosine
Linear warmup epochs	5
Base learning rate	5e-6
Epochs	30
Data augmentation	
RandomFlip	0.5
Resize	RandomSizedCrop
Crop size	224(Default)
GrayScale	0.2
Other regularisation	
Weight decay	0.02

We describe the simulated dataset statistics of the HMDB-51 [4], UCF-101 [6], and Kinetics-100 [3] datasets. The HMDB-51 dataset consists of 6,766 videos categorized into 51 action classes, with 3,570 videos used for training, and 1,530 videos used for testing. Each video is sourced and truncated from movies, online videos, and public databases. The UCF-101 dataset collected 9,537 training videos and 3,783 testing videos from YouTube, across 101 action categories. The Kinetics-100 [3] dataset originally consisted of 64, 12, and 24 class subsets for training, validation, and testing. In our setup, we modify this by randomly selecting 80% of the samples from each action category for the training set, resulting in 7,318 training videos, and reserving 20% for testing, leading to 1,882 testing videos.

During the simulation of foggy conditions using the ASM model, we first utilize DepthAnythingV2 [11] to obtain the relative depth of each frame in the video. Subsequently, we incorporate these depth maps $d(x)$ into the

Table 2. Quality of our FogAct dataset, showcasing clean and foggy images along with the corresponding contrast, laplacian edge, and image quality metrics, including MAE, RMSE, MI, and color distortion.

Metrics	Quality	Range	Metrics	Quality	Range
Contrast_c	59.4	$0\sim+\infty\uparrow$	Contrast_f	30.5	$0\sim+\infty\uparrow$
Edge_f	13.6	$0\sim+\infty\uparrow$	Edge_c	371.1	$0\sim+\infty\uparrow$
MAE	125.0	$0\sim255\downarrow$	MI	0.7	$0\sim1\uparrow$
RMSE	10.2	$0\sim255\downarrow$	Distortion	0.1	$-1\sim1\uparrow$

Table 3. Dynamic frame statistics of the FogAct dataset compared with three simulated datasets: HMDB-51, UCF-101, and Kinetics-100. "FR" denotes frame ratio, "PR" represents person ratio, and "PP" stands for person pixels. Bounding boxes are obtained using [7].

Dataset	Resolution	FR(%)	PR(%)	PP
HMDB-51	320 ×240	92.8	37.5	28,769
UCF-101	320 ×240	89.1	24.9	19,123
Kinetics-100	454 ×256	84.9	36.0	41,840
FogAct	1920 ×1080	99.9	9.8	201,347

ASM equation

$$I(x) = J(x)t(x) + A(1 - t(x)), \tag{1}$$

where $I(x)$ represents the foggy image, $J(x)$ is the clean image without fog, A denotes the atmospheric light, and $t(x) = e^{-\beta d(x)}$ is the transmission map, with β being the scattering coefficient. Finally, we use this equation to generate the simulated foggy video.

Dataset Evaluation. In this study, we employ two types of vision-language models to assess the realism of foggy images. Specifically, for Q-Align [9], we use the prompt `[the realism of the image is]` to evaluate fog quality. Additionally, we leverage GPT-4o to generate prompts for evaluating the realism of the FogAct dataset, which targets action recognition in foggy conditions. Thanks to its advanced natural language understanding and generation capabilities, GPT-4o effectively evaluates data realism and offers valuable support for dataset quality assurance. The final prompt adopted is: `[evaluate the realism of the fog in the given image on a scale of 0 to 5].`

Metric Details. We analyze the video quality of our FogAct dataset using contrast, edge, distortion, and similarity metrics, as shown in Tab. 2. These metrics collectively reflect the degradation in visibility, texture detail, and structural consistency caused by fog, highlighting the challenges of action recognition under such conditions.

We present the calculation formulas of the evaluation metrics for the FogAct dataset in the Tab. 2. Specifically, Edge quantifies the degree of sharpness by measuring the variance of the Laplacian operator. The Laplacian variance for a foggy image $\mathbf{I}_\text{f}$ is computed as:

$$\sigma^2 = \text{Var}(\nabla^2 \mathbf{I}_\text{f}),\ \text{Edge} = \frac{1}{\text{M}} \sum_{\text{i}=1}^{\text{M}} \sigma_\text{i}^2, \tag{2}$$

where $\nabla^2 \mathbf{I}_\text{f}$ denotes the result of applying the Laplacian filter to the foggy image $\mathbf{I}_\text{f}$, Var indicates variance and σ_i^2 represents the Laplacian variance of the i-th foggy image among a total of M images.

We present a foggy image along with the corresponding clean image for computing its color distortion

$$\text{Distortion} = \frac{\sum_{i=1}^{N} (\mathbf{H}_i^\text{f} - \mathbf{H}_\text{ave}^\text{f})(\mathbf{H}_i^\text{c} - \mathbf{H}_\text{ave}^\text{c})}{\sqrt{\sum_{i=1}^{N} (\mathbf{H}_i^\text{f} - \mathbf{H}_\text{ave}^\text{f})^2 \sum_{i=1}^{N} (\mathbf{H}_i^\text{c} - \mathbf{H}_\text{ave}^\text{c})^2}}, \tag{3}$$

where $\mathbf{H}^\text{f}$ and $\mathbf{H}^\text{c}$ denote the histograms of the hazy and clean images, respectively. Here, i indexes the histogram bins, ave denotes the mean, and N is the number of bins. MI represents mutual information, which quantifies the similarity and dependency between clean and foggy images via histogram calculations.

$$\text{MI} = \sum_{x=1}^{X} \sum_{y=1}^{Y} \text{P}(x, y) \log \left(\frac{\text{P}(x, y)}{\text{P}(x)\text{P}(y)} \right), \tag{4}$$

where $\text{P}(x, y)$ represents the joint probability distribution, while $\text{P}(x)$ and $\text{P}(y)$ are the marginal distributions of x and y, respectively. X and Y are the upper bounds of x and y.

Dynamic Analysis. We compare the FogAct dataset with existing action recognition datasets such as HMDB-51, UCF-101, and Kinetics-100 in Tab. 3 and find that FogAct has the highest resolution among them. Additionally, we randomly sample 10 videos per action from each dataset and use YOLO [7] to detect the bounding box of each frame in the videos. We observe that 99.9% of the frames in FogAct contain human figures, significantly surpassing the 84.9% in Kinetics-100. Moreover, the average bounding box size in FogAct reaches 200,000 pixels, which is five times larger than those in other datasets, further demonstrating the high visual quality of FogAct.

Table 4. Ablation experiments on different text prompt templates.

Template	Top-1(%)	Top-5(%)
`A video of action { }`	88.6	99.4
`A picture of action { }`	88.6	99.4
`This is a video about { }`	88.6	99.4
`A video of { }`	88.7	99.4
`This is a video of { }`	88.7	99.4
`This is an example of { }`	88.7	99.4
`This clip contains { }`	88.7	99.4

Table 5. Baseline performance comparison for different pre-training datasets.

Train	Test	Top-1(%)	Top-5(%)
foggy	foggy	75.0 (73.2)	95.7 (95.6)
clean	clean	86.7 (79.5)	98.9 (97.6)
clean	foggy	59.8 (51.7)	85.9 (79.9)
defogged	defogged	75.7 (64.4)	96.3 (93.0)
clean+foggy	foggy	79.7 (78.5)	97.8 (97.8)
Ours		**88.7**	**99.4**

Human Evaluation. To intuitively assess the quality of our FogAct dataset, we conduct a human evaluation. We enlist 21 experts, all with at least a graduate-level education, and randomly assign each expert 30 pairs of images. Each pair consists of one image from our dataset and one from the fog simulation generated by ASM. The evaluation is conducted in a blind setting, where participants are unaware of which image belonged to our dataset. Participants are asked to determine which image exhibited better quality by assigning a score on a scale from 1 to 5, where 1 represents the lowest quality and 5 represents the highest quality. An example is shown in Fig. 7.

The results of the human evaluation indicate a strong preference for our FogAct dataset among experts. With a maximum score of 5, our dataset achieves an average score of 3.6, while the simulated dataset scores 2.5. This further validates the realism of the FogAct dataset in real-world foggy conditions.

2. Additional Ablation Study and Discussion

Different prompt template. In the evaluation process, we design various text prompt templates to analyze the alignment between fog-invariant features and different textual characteristics. The results, as shown in Tab. 4, demonstrate that regardless of template variations, our model consistently achieves stable recognition accuracy, with a Top-1 accuracy of 88.7% and a Top-5 accuracy of 99.4%.

Number of ME. We analyze the impact of the number of ME on recognition accuracy. Using 1 ME achieves 86.9%, which improves to 87.9% with 2 MEs. However, increasing to 3 MEs slightly reduces accuracy to 87.6%. Considering both performance and efficiency, we use 2 MEs in our experiments.

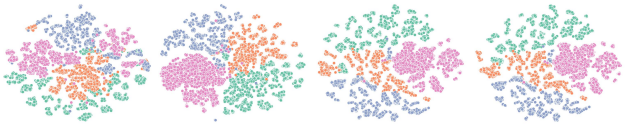
Pre-training Dataset. In Tab. 5, we evaluate baseline pre-trained on Kinetics-400 (green) and find that WIT-400M (black) pre-training yields better performance, leading us to adopt it for all experiments.

Generalization to Real-World Foggy Data. As shown in Tab. 6, we evaluate the Top-2 action recognition methods and the baseline model on real-world foggy data. All models are trained on our FogAct dataset and directly tested on real-world foggy videos in a zero-shot manner. Our method achieves an accuracy of 70.0%, outperforming the second-best approach by a notable margin of 10%, which demonstrates its strong generalization capability under real-world foggy conditions.

Figure 1. Visualization of real-world action video examples.

Table 6. Generalization experiments to real-world foggy data.

Method	Top-1(%)
Baseline [8]	50.0
ViFi-CLIP [5]	40.0
OST [1]	60.0
Ours	**70.0**



(a) Baseline (b) +FAS (c) +ME (d) Ours

Figure 2. T-SNE visualization comparing the baseline and our method with added components.

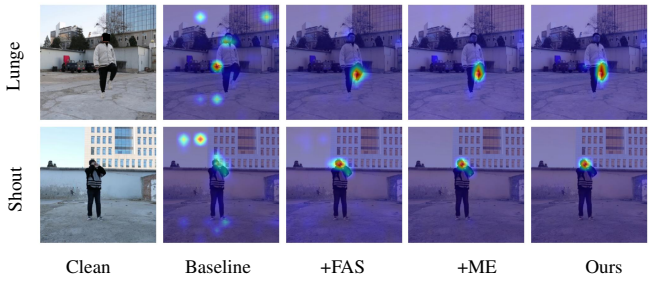


Figure 3. Heatmap visualization of components on clean counterparts.

3. Additional Results

Visualization. Fig. 1 presents real-world foggy action videos collected from publicly accessible platforms, covering a diverse range of actions and conditions. These videos highlight the challenges of action recognition under foggy conditions. The dataset is obtained in compliance with platform terms of service and fair use principles, with no copyrighted material directly reproduced. Additionally, all per-

Figure 4. Comparison of PTTD [2] zero-shot performance after training on DIV&Flickr and its fine-tuned performance on our FogAct dataset.

sonal identifiers have been removed to ensure anonymity, and the data is solely used for academic research, adhering to ethical and legal standards.

Fig. 2 presents a t-SNE visualization illustrating the impact of the added components on action classification in foggy conditions. As additional components are integrated, the feature distributions across action categories become more concentrated, resulting in improved recognition performance under fog conditions.

Fig. 3 presents heatmaps generated across components of our framework during execution. As the heatmaps progressively focus on key regions, we observe that foggy videos, despite the visual distortions caused by fog, contain dynamic features absent in the clean videos, providing additional effective information at the feature level.

Fig. 4 shows the visualization of PTTD [2], including its zero-shot defogging performance on our FogAct dataset and the performance after fine-tuning on FogAct. It is evident that fine-tuning on our FogAct dataset significantly improves the defogging results, thereby enhancing the foggy action recognition accuracy.

Fig. 5 presents the detailed class-wise sample statistics of the FogAct dataset. The mean and standard deviation of video counts per class are 88.4 and 7.6, respectively. For instance, the action classes 'nod', 'push', 'look back', 'applaud', and 'walk' contain 90, 86, 89, 87, and 88 videos, respectively.

Fig. 6 illustrates the distribution of volunteer counts across action categories in the FogAct dataset. Each action is performed by 11 to 15 volunteers, with most actions involving 12 to 14.

Fig. 8 provides an in-depth visualization of the FogAct dataset, highlighting three representative actions: *'Rub eyes'*, *'Sweep floor'*, and *'Wave to stop'*. To account for viewpoint variations, each action is captured from four distinct perspectives: front, back, left, and right. To better illustrate the impact of fog on action recognition, we include both clean and foggy sequences, where the clean frames depict the temporal progression of each action, and their corresponding foggy versions demonstrate how visual degradation affects motion perception.

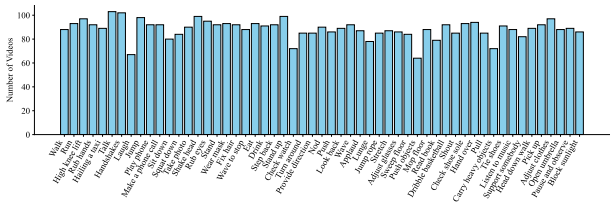


Figure 5. Distribution of the number of videos across different action categories in the FogAct dataset.

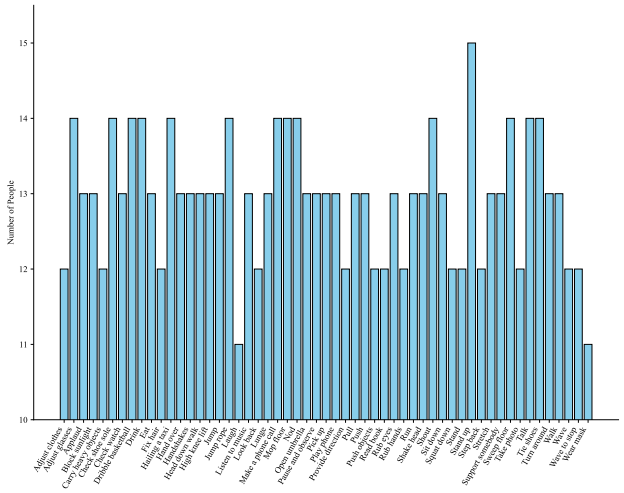


Figure 6. Distribution of the number of volunteers across different action categories in the FogAct dataset.

Future Work. FogAct is the first real-world foggy action dataset with paired clean and foggy videos captured by a stereo system, enabling extraction of fog-invariant features. Next, we plan to use neural networks to map left-view clean images to the right view, generating pixel-level aligned video pairs to further advance this field.

Figure 7. Example image pairs used in the human evaluation study. Each pair consists of an image from our FogAct dataset and a corresponding image generated by ASM-based fog simulation. Experts are asked to assess the quality of the images in a blind setting and assign a score from 1 to 5. The scores displayed below the images represent their average ratings across all experts, reflecting the high real-world quality of our FogAct dataset.

Figure 8. Visualization of the FogAct dataset on three actions: *'Rub eyes'*, *'Sweep floor'*, and *'Wave to stop'*. For each action, we present four perspectives—front, back, left, and right. The left side displays four sequential frames over time, while the right side shows their corresponding foggy versions.

References

[1] Tongjia Chen, Hongshan Yu, Zhengeng Yang, Zechuan Li, Wei Sun, and Chen Chen. Ost: Refining text knowledge with optimal spatio-temporal descriptor for general video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18888–18898, 2024. 1, 3

[2] Zixuan Chen, Zewei He, Ziqian Lu, Xuecheng Sun, and Zhe-Ming Lu. Prompt-based test-time real image dehazing: a novel pipeline. In *European Conference on Computer Vision*, pages 432–449. Springer, 2024. 4

[3] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[4] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 1

[5] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6545–6554, 2023. 3

[6] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision*, 2(11):1–7, 2012. 1

[7] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, et al. Yolov10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems*, 37:107984–108011, 2024. 2

[8] Mengmeng Wang, Jiazheng Xing, Jianbiao Mei, Yong Liu, and Yunliang Jiang. Actionclip: Adapting language-image pretrained models for video action recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 3

[9] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 2

[10] Wenhao Wu, Xiaohan Wang, Haipeng Luo, Jingdong Wang, Yi Yang, and Wanli Ouyang. Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6620–6630, 2023. 1

[11] Honghui Yang, Di Huang, Wei Yin, Chunhua Shen, Haifeng Liu, Xiaofei He, Binbin Lin, Wanli Ouyang, and Tong He. Depth any video with scalable synthetic data. *arXiv preprint arXiv:2410.10815*, 2024. 1