Harnessing Temporal Causality for Advanced Egocentric Video Understanding

Shuming Liu[1] Lin Sui[2] Chen-Lin Zhang[3]
Fangzhou Mu[4] Chen Zhao[1] Bernard Ghanem[1]

[1]King Abdullah University of Science and Technology (KAUST)
[2]4Paradigm Inc [3]Moonshot AI [4]NVIDIA

Abstract

This report describes our submission to the EPIC-Kitchens Challenge 2024, including Action Recognition, Action Detection, and Audio-Based Interaction Detection. Our key findings are: (1) Ensembling different models improves the action recognition task, (2) Hybrid temporal causal modeling is important for egocentric action detection, and (3) A one-stage action detection framework can provide a strong baseline for audio-based interaction detection. We achieved 1st place in all three tasks. Our code is available at `https://github.com/sming256/OpenTAD`.

1. Introduction

Action Recognition, Action Detection, and Audio-Based Interaction Detection are three critically important tasks essential for automatic video or audio processing and even multi-modal learning. Action recognition aims to classify videos into given action categories. Action detection seeks to localize action instances and then recognize corresponding categories in untrimmed videos. Audio-Based Interaction Detection, however, aims to map audio inputs to the corresponding action labels. Many previous works have explored these three areas, but few have focused on the egocentric setting. Hence, the EPIC-Kitchens dataset [4], which has diverse verb, noun, and action categories, remains a significant challenge.

For the Action Recognition task, our solution builds on InternVideo2 [9] and LAVILA [11]. We explore the training process and ensemble policy to obtain strong top-1 accuracy. We first fine-tune the pretrained InternVideo2 model separately on the verb and noun subsets, resulting in two strong action recognition models. To further improve the action accuracy, we explore different ensemble strategies. Finally, we use a simple but effective ensemble method that aimed to increase the diversity of model predictions. We ensemble models with different training targets, *i.e.*, models trained to predict verb and noun labels or directly predict the action label. We also ensemble models with different architectures (InternVideo2 and LAVILA). Our ensemble policy ultimately achieves a top-1 action accuracy of 57.9% on the test set, securing first place in the competition, which is 1.1% ahead of the second-place result.

For Action Detection, our approach is built on ActionFormer [10] and implemented under the OpenTAD [6] framework. We train the InternVideo2 separately on verb and noun subsets and use them to extract noun and verb features. To capture long-range temporal relationships, we further propose the hybrid causal block to aggregate the forward and backward information. Our approach results in an average mAP of 31.97% on action task, which is 5.75% higher than the second place.

We also select ActionFormer [10] as the base method for the Audio-Based Interaction Detection track. We achieve first place with an average mAP of 14.82%, surpassing the second place by 3.42% average mAP.

2. Action Recognition

A robust action recognition model is crucial for downstream tasks, such as action detection. Therefore, we devote considerable effort to obtaining a strong recognition model. Our solution first trains the recent InternVideo2 [9] model on the verb and noun subsets and combines the predictions to get the final action scores. Subsequently, to improve action recognition accuracy, we explore the ensemble strategy and achieve the first place with a 57.9% top-1 accuracy.

2.1. Method

To obtain a strong action recognition model, we use two recently proposed action recognition methods, InternVideo2 [9] and LAVILA [11]. InternVideo2 proposes a progressive learning paradigm, which includes unmasked video token reconstruction, multimodal contrastive learning, and next token prediction training stages. LAVILA, on the other hand, proposes enhancing representations by leveraging pre-trained large language models. We select InternVideo2 as our base model for fine-tuning. The LAVILA model is only used to obtain prediction results to enrich the prediction diversity for the model ensemble process. InternVideo2 is pretrained on a hybrid dataset, K-Mash 1.1M, which contains 1.1M video clips. The model is then fine-tuned on the Kinetics700 [2] dataset. We further fine-tune the model on the EPIC-Kitchens 100 [4].

Fine-tuning InternVideo2 on EPIC-Kitchens Action Recognition. We fine-tune the action recognition model with the pretrained InternVideo2$_{\text{s1}}$-1B. Following previous experience in the action detection area, we first fine-tune the model on the verb and noun subsets individually, resulting in two individual action recognition models. Action scores are obtained by combining the prediction results of the verb and noun models. To enhance the model prediction diversity of the ensemble process, we also fine-tune another InternVideo2 model to predict action labels directly. For simplicity, the models will be referred to as InternVideo2$_{\text{s1}}$-1B$_{\text{sep}}$ and InternVideo2$_{\text{s1}}$-1B$_{\text{act}}$, respectively.

We select the training hyper-parameters on the verb and noun training set, and perform fine-tuning directly on the training plus validation set of EPIC-Kitchens 100 with the searched hyper-parameters. For each video, we use sparse sampling to sample 16 frames, and the short edge of sampled videos is resized to 288. Then, we perform the center crop on these resized videos. The model is fine-tuned for 10 epochs with a batch size of 128. We use AdamW as the optimizer with a 0.05 weight decay, an initial learning rate of 1e-4, and we use the cosine learning rate decay strategy. We also warm up the model for 3 epochs. During the inference, we sample 4 segments and 3 crops for each video.

Model Ensemble. We use a simple but effective ensemble method to enhance model accuracy. For models trained to predict action labels directly, *i.e.*, the LAVILA model and InternVideo2$_{\text{s1}}$-1B$_{\text{act}}$ model, we directly use the 3806-dim predictions. As for InternVideo2$_{\text{s1}}$-1B$_{\text{sep}}$, we first select the scores of the possible 3806 actions from the 97$\times$300-dim action probability matrix and then normalize the scores. Finally, we obtain the prediction results by weighted summation of these predictions from different models.

2.2. Results

Model accuracy on the validation set. Firstly, we show the verb & noun top-1 accuracy of our used models on the validation set in Table 1. The LAVILA-L model, which is obtained from the official GitHub repo, could achieve 65.4% top-1 noun accuracy and 73.0% top-1 verb accuracy. Our finetuned InternVideo2$_{\text{s1}}$-1B$_{\text{sep}}$ can achieve 70.5 % top-1 noun accuracy and 77.6 top-1% verb accuracy, surpassing the previous SOTA by a large margin.

Model accuracy on the test set. Table 2 summarizes the accuracy of our models on the EPIC-Kitchens test set. Our fine-tuned InternVideo2$_{\text{s1}}$-1B$_{\text{sep}}$ achieves top-1 noun accuracy of 68.4%, top-1 verb accuracy of 73.5%, and top-1 action accuracy of 55.3%. By employing model ensembling, we secure first place with a top-1 action accuracy of 57.9%.

Table 1. **Results of Top-1 accuracy on the validation set of EPIC-Kitchens 100 action recognition task.**

Model	Noun	Verb
LAVILA-L	65.4	73.0
InternVideo2$_{\text{s1}}$-1B$_{\text{sep}}$	**70.5**	**77.6**

In Table 2, Ensemble 1 refers to the combination of InternVideo2$_{\text{s1}}$-1B$_{\text{sep}}$, InternVideo2$_{\text{s1}}$-1B$_{\text{act}}$, and LAVILA-L models, each contributing equally with a weight ratio of 1:1:1. Additionally, we observe that ten epochs were insufficient for the InternVideo2$_{\text{s1}}$-1B$_{\text{act}}$ model, prompting us to fine-tune another model of the same type for 20 epochs. Ensemble 2 and 3 incorporate these four action recognition models, differing only in their weighting ratios. Ensemble 2 utilizes an equal weighting of 1:1:1:1. For Ensemble 3, we adjust the weights by reducing that of InternVideo2$_{\text{s1}}$-1B$_{\text{act}}$ and increasing the weight of InternVideo2$_{\text{s1}}$-1B$_{\text{sep}}$, resulting in a final ratio of 0.3:0.225:0.225:0.25.

Table 2. **Results of Top-1 accuracy on the test set of EPIC-Kitchens 100 action recognition task.**

Model	Action	Noun	Verb
InternVideo2$_{\text{s1}}$-1B$_{\text{sep}}$	55.3	68.4	73.5
Ensemble 1	57.6	-	-
Ensemble 2	57.7	-	-
Ensemble 3	**57.9**	-	-

3. Action Detection

3.1. Method

Temporal Action Detection (TAD) is a fundamental task in understanding long-form videos, aimed at localizing candidate actions in untrimmed videos and predicting their start times, end times, and categories. In our submission, we utilize a feature-based TAD pipeline that encompasses feature extraction and action detection.

Feature Extraction. We employ the VideoMAE-L model and InternVideo2$_{s1}$-1B as the video encoding backbones. Both models undergo self-supervised pretraining followed by supervised fine-tuning on the Kinetics-700 dataset. Given the substantial domain difference between the third-person perspective videos used in pretraining and the ego-centric videos in our tasks, we further fine-tune the models on the EPIC-Kitchens action recognition tasks, detailed in Section 2. The noun and verb subsets are fine-tuned separately, and then we use a sliding window approach to extract

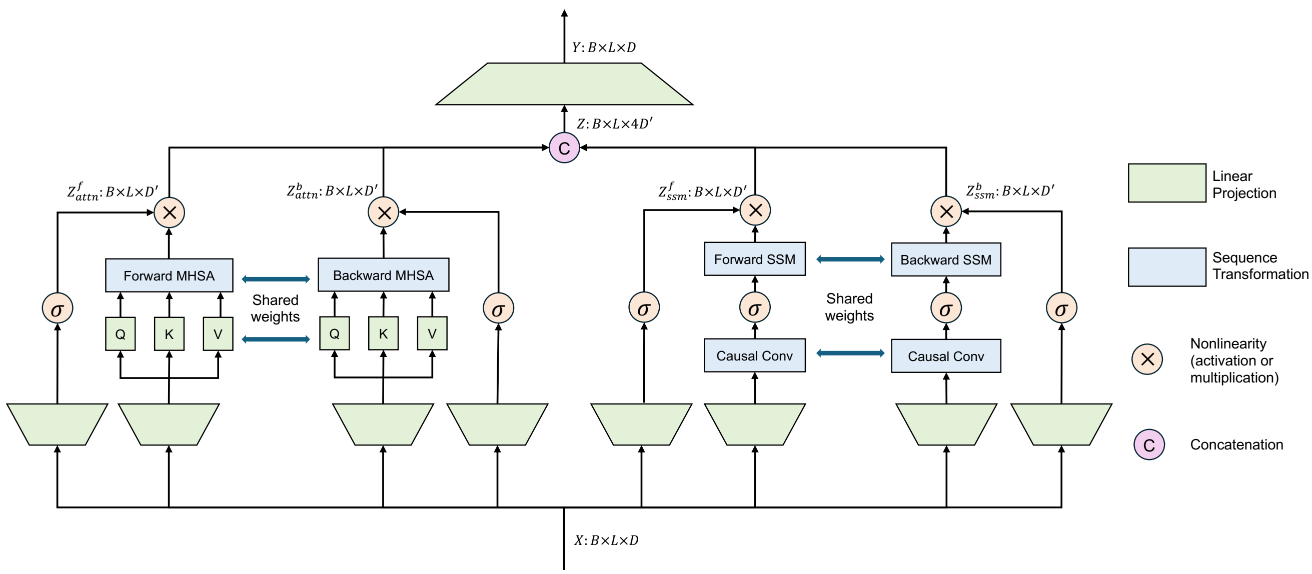

Figure 1. **Hybrid Causal Block.**

snippet features, with each snippet spanning 16 frames and a stride of 8 frames between consecutive snippets.

Improved Detection Baseline. Our detection model is based on ActionFormer, a simple yet effective one-stage action detection framework. We optimize the model's hyper-parameters, including the number of feature pyramid levels, the regression loss weight, the probability of input channel dropout, and the number of training epochs, thereby establishing a stronger baseline for action detection.

Bidirectional Temporal Causal Modeling. In ActionFormer, a local transformer is used to integrate temporal context. Drawing inspiration from recent advancements such as VideoMambaSuite [3], which recommends using Mamba for video understanding tasks, we introduce a hybrid causal block, depicted in Figure 1. This block consists of both a self-attention module and an SSM (State-Space Model) module, facilitating causal modeling from both forward and backward directions. In the hybrid causal block, input projectors are learned separately in each direction, while the query, key, and projection layers and SSM modules share parameters for two directions. Hybrid causal block not only preserves the benefits of SSM and self-attention but also enhances the capability to capture long-range temporal relationships, providing a significant improvement over both ActionFormer and VideoMambaSuite.

3.2. Results

Our results on the validation subset are summarized in Table 3. It is important to note that the models for noun and verb recognition are trained separately. Initially, by optimizing the hyper-parameters of the detection model, we achieve an increase in mAP for both noun and verb tasks. Replacing the local transformer with either Mamba or our proposed hybrid causal block further improves the detection performance, affirming the efficacy of bidirectional temporal causal modeling for temporal aggregation in TAD tasks.

Table 3. **Results on the validation set of EPIC-Kitchens 100 action detection task.** The noun and verb models are trained separately, and we report the average mAP.

Method	Feature	Noun	Verb
ActionFormer	VideoMAE	28.73	28.61
Improved Baseline	VideoMAE	30.21	30.08
Mamba	VideoMAE	30.51	30.35
Hybrid Causal Block	VideoMAE	**30.96**	**30.83**
Hybrid Causal Block	InternVideo2	**37.02**	**33.08**

he conclusive results for noun, verb, and action tasks on the validation and test subsets are reported in Table 4. To construct the action labels, we select the top-10 noun classes and top-10 verb classes for each timestamp and calculate their product to establish candidate action probabilities. And Soft-NMS [1] is applied to eliminate redundant proposals. Using the above action labels, the detection performance for the noun and verb tasks decreases by approximately 2-3% mAP, which indicates a potential conflict between the two tasks. For example, the mAP for the noun task decreased from 37.02% to 34.44% on the val subset, and for the verb task, it dropped from 33.08% to 27.67%. Ultimately, we achieve an average mAP of 31.97% on the action task in the test set.

Table 4. **Results on the EPIC-Kitchens 100 action detection task.** All results on the test set are evaluated on the test server.

Split	Method	Feature	Task	mAP@tIoU					
				0.1	**0.2**	**0.3**	**0.4**	**0.5**	**Avg.**
Val	Ours 2023 Submission [10]	VideoMAE [7, 8]	Verb	32.73	31.60	29.13	26.74	23.67	28.77
			Noun	31.32	29.70	27.25	25.32	21.33	26.98
			Action	25.73	24.98	23.72	22.46	19.11	23.20
	Ours 2024 Submission [10]	InternVideo2-1B	Verb	31.05	29.96	28.02	25.75	23.60	**27.67**
			Noun	39.36	37.78	35.53	32.12	27.42	**34.44**
			Action	33.01	32.03	30.28	27.86	24.98	**29.63**
Test	Ours 2023 Submission [10]	VideoMAE [7, 8]	Verb	31.01	30.04	28.01	25.44	22.32	27.36
			Noun	30.32	28.76	27.20	24.28	20.74	26.26
			Action	25.54	24.54	23.16	21.04	18.35	22.52
	Ours 2024 Submission [10]	InternVideo2-1B	Verb	35.79	34.10	30.48	28.02	24.73	**30.02**
			Noun	40.66	38.62	36.31	32.54	27.98	**35.22**
			Action	36.09	34.69	32.67	29.91	26.50	**31.97**

4. Audio-Based Interaction Detection

4.1. Method

Audio-based interaction detection aims to localize candidate actions within untrimmed videos, emphasizing audio cues as the primary indicators of target actions. Following the detection methodology outlined in Section 3, we use ActionFormer as the baseline model and incorporate the hybrid causal block for enhanced temporal modeling. Uniquely, rather than relying on the InternVideo2 features, we prioritize audio cues, utilizing the Audio-SlowFast model [5] for feature extraction. Audio-SlowFast is a dual-stream convolutional network designed for audio recognition, processing time-frequency spectrogram inputs, and pretrained on the EPIC-Kitchens action recognition task.

4.2. Results

Our implementation is based on OpenTAD framework [6]. We present the results of our audio-based interaction detection in Table 5. Compared to the baseline, which uses the ActionFormer, we improve the average mAP from 7.35% to 14.82%. This significant enhancement confirms the effectiveness of our detection model, establishing a solid baseline for audio-based interaction detection. With stronger audio features and additional visual features, we expect higher detection performance.

5. Conclusion

In this report, we present our solution for action recognition, temporal action detection, and audio-based interaction detection in egocentric videos. By harnessing the stronger video backbones and capturing long-range temporal causal relationships, we have successfully established a new state-of-the-art, achieving first place in the three respective tracks of the EPIC-Kitchens 2024 challenges. We hope that our approaches and findings can shed light on the field of long-form egocentric video understanding.

Table 5. **Results on the EPIC-Kitchens 100 audio-based interaction detection task.** All the methods use the Audio-SlowFast feature for fair comparison. [5]

Method	Subset	0.1	0.2	0.3	0.4	0.5	Avg.
Ours	Val	16.85	15.64	14.60	12.99	11.22	14.26
Baseline	Test	9.57	8.51	7.38	6.22	5.05	7.35
Ours	Test	**19.81**	**17.24**	**14.82**	**12.48**	**9.74**	**14.82**

References

[1] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-NMS–improving object detection with one line of code. In *Int. Conf. Comput. Vis.*, pages 5561–5569, 2017. 3

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the Kinetics dataset. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4724–4733, 2017. 2

[3] Guo Chen, Yifei Huang, Jilan Xu, Baoqi Pei, Zhe Chen, Zhiqi Li, Jiahao Wang, Kunchang Li, Tong Lu, and Limin Wang. Video mamba suite: State space model as a versatile alternative for video understanding. *arXiv preprint arXiv:2403.09626*, 2024. 3

[4] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision. *arXiv preprint arXiv:2006.13256*, 2020. 1, 2

[5] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Slow-fast auditory streams for audio recognition. *CoRR*, abs/2103.03516, 2021. 4

[6] Shuming Liu, Chen Zhao, Fatimah Zohra, Mattia Soldan, Carlos Hinojosa, Alejandro Pardo, Anthony Cioppa,

Lama Alssum, Mengmeng Xu, Merey Ramazanova, Juan León Alcázar, Silvio Giancola, and Bernard Ghanem. Opentad: An open-source toolbox for temporal action detection. 2024. 1, 4

[7] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022. 4

[8] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 4

[9] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024. 1

[10] Chen-Lin Zhang, Jianxin Wu, and Yin Li. ActionFormer: Localizing moments of actions with transformers. In *Eur. Conf. Comput. Vis.*, pages 492–510, 2022. 1, 4

[11] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *CVPR*, 2023. 1