Can't make an Omelette without Breaking some Eggs: Plausible Action Anticipation using Large Video-Language Models (Supplementary Material)

Himangi Mittal[1,2*] Nakul Agarwal[1] Shao-Yuan Lo[1] Kwonjoon Lee[1]
[1]Honda Research Institute, USA [2]Carnegie Mellon University
hmittal@andrew.cmu.edu {nakul_agarwal, shao-yuan_lo, kwonjoon_lee}@honda-ri.com

1. Implementation Details

We train our method end-to-end with a batch size of 2 for Ego4D and 4 for EPIC-Kitchens-100, linear warmup cosine as learning rate scheduler, along with the pre-trained weights of Video-LLaMA [10] on 2 A6000 GPUs for 2.5 days.

1.1. Metrics

Edit Distance (ED@(Z=20)) [5]: This metric is computed over a sequence of verb and noun predictions using the Damerau-Levenshtein distance [3, 6] and takes into account the sequential nature of the action anticipation task. A prediction is considered correct if it matches the ground truth at a specific time step using the edit distance operations - insertion, deletion, substitution, and transposition. A total of K predictions are evaluated and the smallest edit distance between a prediction and ground truth is reported [5]. We consider the value of $Z = 20$ and $K = 5$ which is the same as Ego4D [5].

Class-mean Top-5 Recall (%) [2]: This metric evaluates if the ground truth class is within the top-5 predictions and averages the per-class performance to equally weight all the classes. The top-k criterion takes into account the uncertainty/multi-modality in the future action prediction and class-mean is helpful for balancing the long-tail distribution.

2. Quantitative Analysis

Generalization and robustness to long-tail: We evaluate our method on the unseen participants and tail classes of EPIC-Kitchens-100 [2] and present the results in Table 1. Unseen participants consists of those participants that are not present in the train set and tail classes are defined to be the smallest classes whose instances are around 20% of the total number of instances in the train set. We observe that a better performance of our approach on the unseen participants as compared to the other baselines shows the generalizability of our model across unseen data. Similarly, a better performance on the tail classes shows that our model is robust to the long-tail distribution of the EPIC-Kitchens-100 dataset.

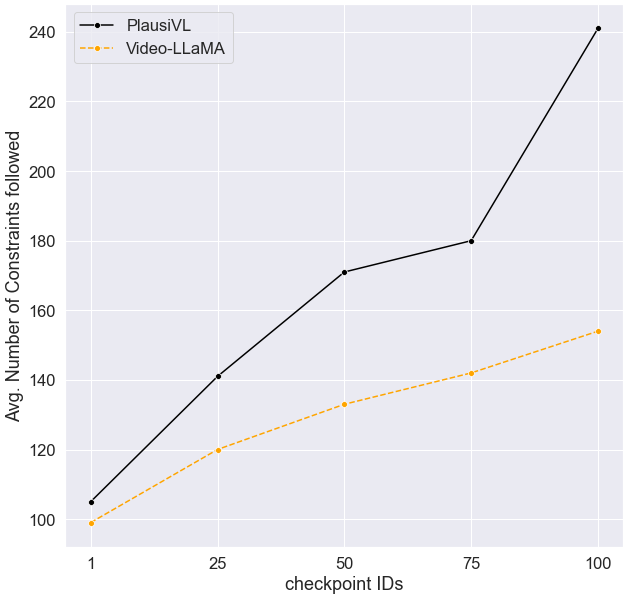

Figure 1. Analysis of plausibility in generated action sequence: Black line represents our method and orange is the baseline, Video-LLaMA. Comparing the two line plots, we can observe that PlausiVL follows more number of temporal and action constraints over training than Video-LLaMA indicating that the objective functions $\mathcal{L}_{\texttt{plau}}$ and $\mathcal{L}_{\texttt{rep}}$ are helping the model to learn temporal cues needed to generate plausible action sequences for action anticipation.

Analysis of plausibility in generated action sequence: To evaluate if the generated text is a plausible action sequence and additionally, the efficacy of the $\mathcal{L}_{\texttt{plau}}$ and $\mathcal{L}_{\texttt{rep}}$ objective functions, we calculate the average

* This work was done as Himangi Mittal's internship project at Honda Research Institute, USA.

Method	Unseen ↑			Tail ↑		
	Verb	**Noun**	**Action**	**Verb**	**Noun**	**Action**
RU-LSTM [2]	28.78	27.22	14.15	19.77	22.02	11.14
Temporal Aggregation [8]	28.80	27.20	14.20	19.80	22.00	11.10
Video LLM [1]	-	-	12.60	-	-	12.00
AFFT [11]	24.80	26.40	15.50	15.00	27.70	16.20
AVT [4]	29.50	23.90	11.90	21.10	25.80	14.10
MeMViT [9]	28.60	27.40	15.20	25.30	31.00	15.50
InAViT [7]	46.45	51.30	25.33	45.34	39.21	20.22
Video LLaMA [10]	46.87	51.47	25.40	45.71	39.32	20.35
PlausiVL	**49.50**	**53.90**	**27.01**	**48.44**	**41.29**	**22.10**

Table 1. Performance of action anticipation on EPIC-Kitchens-100 Unseen Participants and Tail Classes on class-mean Top-5 recall (%) ↑): Higher is better. Our method is able to outperform all the previous baselines.

Method	ED@(Z=20) ↓	
	Verb	**Noun**
PlausiVL (w/ DNR)	0.689	0.695
PlausiVL	0.679	0.681

Table 2. (Ego4D) Performance of PlausiVL with and without "DNR: Do NOT repeat actions" in the prompt. We can observe that having DNR in the prompt does not give much improvement in the performance as compared to training the model with long-horizon action repetition loss ($\mathcal{L}_{\texttt{rep}}$) as objective function.

Method	Class-mean Top-5 recall (%) ↑		
	Verb	**Noun**	**Action**
PlausiVL (w/ DNR)	54.30	53.20	26.63
PlausiVL	55.62	54.23	27.60

Table 3. (EPIC-Kitchens-100) Performance of PlausiVL with and without "DNR: Do NOT repeat actions" in the prompt. Having DNR in the prompt is less effective than training the model with long-horizon action repetition loss ($\mathcal{L}_{\texttt{rep}}$) as objective function.

Method	n_rep=2		n_rep=3		n_rep=4	
	Verb	**Noun**	**Verb**	**Noun**	**Verb**	**Noun**
Video-LLaMA	0.703	0.721	0.704	0.724	0.704	0.726
PLausiVL	0.680	0.681	0.679	0.681	0.680	0.683

Table 4. Results on different n_rep for Ego4D on ED@(Z=20) ↓

Method	Verb	Noun
CLR Paradigm	0.726	0.766
PlausiVL w/ $\mathcal{L}_{\texttt{plau}}$	0.686	0.698
PlausiVL	**0.679**	**0.681**

Table 5. Contrastive Loss with negative sample from other videos (CLR Paradigm) for Ego4D on ED@(Z=20) ↓

number of temporal and action constraints followed in the generated text. We compare the average number of constraints followed by PlausiVL versus the baseline Video-LLaMA [10] and present the graph visualization in Figure 1. We report the average number of constraints followed over the training and show the number over the checkpoints from beginning till the end of training. From the figure, we can observe that as the training of the model with $\mathcal{L}_{\texttt{plau}}$ and $\mathcal{L}_{\texttt{rep}}$ losses progresses, the average number of constraints followed increases in the generated text. Morever, the average number of PlausiVL is higher than that of Video-LLaMA. This indicates that by training the model with $\mathcal{L}_{\texttt{plau}}$ and $\mathcal{L}_{\texttt{rep}}$ objective functions, the model can generate more plausible action sequences and they help the model learn the implicit temporal information needed for plausible action anticipation.

Training with $\mathcal{L}_{\texttt{rep}}$ loss vs prompt tuning: We perform an analysis where instead of training the model with $\mathcal{L}_{\texttt{rep}}$ objective function, we simply prompt the model with the phrase: "Do NOT repeat actions" (DNR). We compare PlausiVL trained with $\mathcal{L}_{\texttt{plau}}$ and $\mathcal{L}_{\texttt{rep}}$ losses (row 2) and PlausiVL trained with $\mathcal{L}_{\texttt{plau}}$ and DNR prompt (row 1) and present the results of this analysis for Ego4D in Table 2 and for EPIC-Kitchens-100 in Table 3. We can observe that simply prompting the model with DNR in the prompt does not give much improvement in the performance as compared to training the model with long-horizon action repetition loss ($\mathcal{L}_{\texttt{rep}}$) as objective function. Training the model $\mathcal{L}_{\texttt{rep}}$ penalizes the model for repeating the actions and makes the model learn to generate more diverse actions. This penalty is helpful in reducing repetition of the actions over a long-horizon. Simply stating DNR in the prompt only gives an instruction/command to the model, whereas, training the model with $\mathcal{L}_{\texttt{rep}}$ loss influences the learning of the model which is needed for the task of action anticipation.

$\mathcal{L}_{\texttt{rep}}$ loss is dataset independent: We perform an analysis to highlight that repetition loss is independent of the dataset. In other words, the performance of the repetition loss does not depend on the number of repeated actions in a dataset. We present this analysis in Table 4. We observe

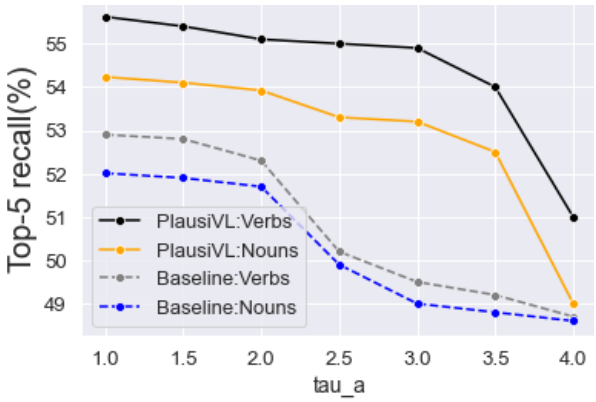

Figure 2. Analysis of τ_a vs. verb-noun class-mean Top-5 recall (%) accuracy on EK100 ↑.

no strong correlation between n_rep and performance, showing data-independency and also show that PlausiVL w/ repetition loss reduces repetition and outperforms the baseline.

Different videos as negative samples for $\mathcal{L}_{\texttt{plau}}$ loss: For the $\mathcal{L}_{\texttt{plau}}$ loss, we use an implausible action sequence as a negative sample. We perform an analysis of using negative samples from other videos and show the results in Table 5. This setting performs worse than Row 2,3 as it gives a weaker signal of counterfactual temporal plausibility than the signal of an implausible action sequence, since sequences from other videos are also temporally plausible.

Anticipation time τ_a vs Accuracy: τ_a is the anticipation time between the end time of observed video and the starting time of the first action to be anticipated. The video during the anticipation period τ_a is unobserved. For EK100, τ_a=1s and for Ego4D, τ_a=2.20s on an average. We analyze changing τ_a versus accuracy on EK100 in Figure 2. We can observe that the method is quite robust till τ_a=3.5s whereas Video-LLaMA is only robust till τ_a=2.0s for EK100. This shows that the model can predict future actions even with a far anticipation time.

3. Qualitative Analysis

In this section, we present more qualitative results of our method. Given a video, the top blue box shows the prediction from PlausiVL and the green box contains the ground truth action sequence for reference. We can observe that our method is able to understand the activity happening in the video and then, generate action sequences accordingly. Additionally, PlausiVL is able to generate action sequences that satisfy the temporal logic constraints and are diverse with less repetitions. The predicted action sequence is also closer to the ground truth action sequence.

Figure 3. Qualitative Results over videos of diverse environments like kitchen, construction sites, etc. and their respective anticipated actions from our method. Given a video, the top blue box shows the prediction from PlausiVL and the green box contains the ground truth action sequence for reference. The model is able to generate plausible action sequences.

References

[1] Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, et al. Videollm: Modeling video sequence with large language models. *arXiv preprint arXiv:2305.13292*, 2023. 2

[2] 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

[3] Fred J Damerau. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176, 1964. 1

[4] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13505–13515, 2021. 2

[5] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 1

[6] Vladimir I Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, pages 707–710. Soviet Union, 1966. 1

[7] Debaditya Roy, Ramanathan Rajendiran, and Basura Fernando. Interaction visual transformer for egocentric action anticipation. *arXiv preprint arXiv:2211.14154*, 2022. 2

[8] Fadime Sener, Dipika Singhania, and Angela Yao. Temporal aggregate representations for long-range video understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 154–171. Springer, 2020. 2

[9] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13587–13597, 2022. 2

[10] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 1, 2

[11] Zeyun Zhong, David Schneider, Michael Voit, Rainer Stiefelhagen, and Jürgen Beyerer. Anticipative feature fusion transformer for multi-modal action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6068–6077, 2023. 2