Intention-Conditioned Long-Term Human Egocentric Action Forecasting @ EGO4D Challenge 2022

Esteve Valls Mascaro∗ 1, Hyemin Ahn2, and Dongheui Lee1, 3

1Autonomous Systems, Technische Universität Wien (TU Wien), Austria
2Ulsan National Institute of Science and Technology (UNIST), South Korea
3Institute of Robotics and Mechatronics, German Aerospace Center (DLR), Germany

Abstract—To anticipate how a human would act in the future, it is essential to understand the human intention since it guides the human towards a certain goal. In this paper, we propose a hierarchical architecture which assumes a sequence of human action (low-level) can be driven from the human intention (high-level). Based on this, we deal with Long-Term Action Anticipation task in egocentric videos. Our framework first extracts two level of human information over the N observed videos human actions through a Hierarchical Multi-task MLP Mixer (H3M). Then, we condition the uncertainty of the future through an Intention-Conditioned Variational Auto-Encoder (I-CVAE) that generates K stable predictions of the next Z = 20 actions that the observed human might perform. By leveraging human intention as high-level information, we claim that our model is able to anticipate more time-consistent actions in the long-term, thus improving the results over baseline methods in EGO4D Challenge. This work ranked first in the EGO4D LTA Challenge by providing more plausible anticipated sequences, improving the anticipation of nouns and overall actions. The code is available at https://github.com/Evm7/ego4dlta-icvae

I. INTRODUCTION

Reasoning about the next actions to take for performing a certain task is essential in the humans everyday life and has direct applications in task planning or in robot collaborative scenarios. The fundamental challenges of anticipating human actions is the inherent uncertainty of the future, which results in high variability in the task execution. However, the future often has only a limited number of plausible predictions. We aim then to reduce the arbitrariness of events by means of conditions. These conditions are hypotheses extracted from past observations of the human that allow to guide the human in its actions: the intention. For instance, if we observe a human cutting a tomato and we know that the intention is to cook a hamburger, our method aims at capturing the human intention, as the highest level concept that defines the objective pursued by the human, to narrow the inherent uncertainty of the future. Moreover, low-level actions are also predicted, which each define exactly the concrete steps followed by the human in this determined task.

II. METHODOLOGY

Long-Term Human Action Anticipation needs to exploit temporal dependencies among observed actions to generate plausible human action sequences in the long-term. Our two-step approach aims at first understanding the observed actions through a Hierarchical MultiTask MLP Mixer (H3M), described in section II-B. Then, an Encoder-Decoder structure inherited by Transformer is proposed for the Intention-Conditioned Variational AutoEncoder (I-CVAE) in section II-C inspired by [2]. Figure 1 shows the overview of the proposed framework.

A. Problem Formulation

Let \( a_{past} = [a_{p,1}, \cdots, a_{p,N}] \in \mathbb{R}^{N,2} \) denote the past N observed actions (composed of verb-noun pair) of the camera wearer from a given untrimmed video, the aim is to anticipate the future sequence of Z actions \( a_{fut} = [a_{f,1}, \cdots, a_{f,Z}] \in \mathbb{R}^{Z,2} \) by generating K possible sequences.

Given \( v_t \) where \( t = [1, \cdots, N] \) as the sequence of frames of a given \( a_{p,t} \), this paper makes use of the pre-extracted features provided by the Ego4d benchmark. For every 1 second of frames (30 fps) a Video-Recognition SlowFast 8x8 (R101) Model pre-trained in Kinetics dataset is used that extracts a feature vector \( f \in \mathbb{R}^{2304} \). By synchronization of videos with timestamp features, zero-padding approach is applied to obtain for every \( v_t \) a global \( f_t \in \mathbb{R}^{14,2304} \). This strategy is validated as average \( v_t \) has 13.04 second length
with 14 sec as the longest clip for a given action \( a_t \).

Therefore a \( N \)-sequence of pre-extracted visual features \( f_{past} = [f_{p,1}, \cdots , f_{p,N}] \) is used as input for the H3M. 2-level outputs are obtained that compose the input structure for the I-CVAE. Intention prediction \( I \) is denoted as a unique label that classifies the overall goal of the camera wearer for the whole past-future actions. Low-level action pair predictions are defined as \( a_{past} = [a_{p,1}, \cdots , a_{p,N}] \) being \( a_{p,t} \) a (verb, noun) pair label. Finally, both \( I \) and \( a_{past} \) are used for generating \( K \) possible future action sequences \( A_{fut} = [a_{fut,1}, \cdots , a_{fut,K}] \in \mathbb{R}^{K \times 2} \).

**B. Hierarchical Multitask MLP Mixer (H3M)**

As shown in Fig. 2, the pre-extracted features \( f_{past} = [f_1, \cdots , f_N] \) are passed to a Hierarchical Multitask MLP-Mixer (H3M) architecture. The Action Mixer models in parallel each \( f_t \) to encode the visual information that define a given (verb, noun) action pair. Let \( f_t = [x_1, \cdots , x_{14}] \in \mathbb{R}^{14,2304} \), composed by 14 feature patches \( x_t \). Mixer layer is defined in [1] and is composed of 2 MLP blocks encapsulated between two matrix transposition operations that captor the global context of \( x_t \). First MLP block identifies strong temporal dependencies and mixes the data among each tokenized patched, while second MLP block leverages spatial same-patch features. Each MLP block consists on two fully-connected layers with a GELU activation function in-between. Finally, global average pooling is applied in the 14 number of patches dimension to project \( R^{14,2304} \) to \( R^{N,2304} \), encoding the action representation for each \( N \) observed actions \( f_{act} = [f_{act,1}, \cdots , f_{act,N}] \). The action head of the H3M projects and classifies each \( f_{act,t} \) as verb and noun through a Fully Connected layer each, obtaining \( N \) (verb, noun) pairs \( a_{past} \). Finally, the attention head applies a second MLP Mixer to leverage the global context of the sequence of action representation features \( f_{act} \) and classifies the human intention \( I \) in the observed video. The overall hierarchical multitask classifier obtains the top-low level information from past observation.

**C. Intention-Conditioned Variational Autoencoder (I-CVAE)**

Conditional Variational Autoencoder architecture is shown in fig. 3. Encoder-Decoder Transformer structure is needed to leverage temporal dependencies between past actions to anticipate future sequence of human (verb, noun) pairs. During training I-CVAE learns the conditional probabilistic distribution in the encoder through exploiting both \( a_{past} \) and \( a_{fut} \) labels. In testing, only the decoder block is used without any provided information of its future.

First of all, learnable embedding matrices are used to project independently each verb and noun pair of the observed sequence \( a_{past} \) into a \( d \)-dimensional embedding vector. Therefore, \( a_{past,emb,v} = [a_{p,emb,v,1}, \cdots , a_{p,emb,v,N}] \in \mathbb{R}^d \) represents the embedded \( N \) observed verbs of the human (same procedure with nouns). For the encoder block, forecasting \( a_{fut,emb,v/n} = [a_{fut,emb,v/n,1}, \cdots , a_{fut,emb,v/n,2304}] \) are also extracted using same embedding weights as past actions. Due to the (verb,noun) pair composition of an action, \( a_{past,emb,v/n} \) and \( a_{past,emb,n} \) are concatenated to \( a_{past,emb} \in \mathbb{R}^{N,2304} \) in the Action Embedder. Intuitively, then, our Transformer Encoder-Decoder aims to reconstruct the \( a_{fut,emb,v/n} \) in the decoder given only the \( a_{past,emb,v/n} \) and the intention label \( I \) as conditions.

Inspired by [2], condition label \( I \) is projected into extra-learnable \( \mu_{token,\text{intention}} \) and \( \Sigma_{token,\text{intention}} \) to inherit the action sequence representation after the Transformer Encoder. Together with embedded observed and forecasting actions, this so-called distribution parameter tokens are summed to a sinusoidal Positional Encoder (PE) and fed into the Transformer encoder, as can be seen in the top diagram of fig. 3. Output \( \mu \) and \( \Sigma \) are then extracted to perform the define a latent distribution from which we will sample in the decoder (\( z \in \mathbb{M} \) where \( \mathbb{M} \in \mathbb{R}^{2304} \)).

During the decoder, as illustrated in the bottom diagram of fig. 3 only the \( N \) observed actions \( a_{past} \) and intention label \( I \) are used to condition the future generation of next \( Z \)
actions $a_{fut}$. A learnable bias token $b_{\text{token}}$ is used to shift the latent representation vector $z$ to an intention-dependent space. $N$ observed actions embeddings $a_{\text{past,emb}}$ are fed into the Transformer Decoder together with the shift latent space and a $O - vector$ for each $Z$ actions to predict, all summed with a sinusoidal Positional Encoder (PE), to anticipate a future sequence of actions representations $\hat{a}_{fut,emb}$. Finally, for each $Z$ actions predicted $\hat{a}_{fut,emb,t}$, a Multi-Head Decoder applies 2 fully-connected layer for classifying verb and noun respectively.

To guide I-CVAE to decode plausible actions representations, an L2 reconstruction loss is applied between the Transformer Encoder and Decoder output: $\hat{a}_{fut,emb}$ and $a_{fut,emb}$. Finally, to ensure the correct classification of verb-noun pairs, a weighted cross-entropy loss is used between the generated results and the ground-truth action sequence to anticipate $a_{fut}$.

### III. Experiments

#### A. Evaluation Metrics

**Metrics.** Following the proposed evaluation protocols in Ego4D LTA Challenge, we report the Edit Distance (ED) metric computed as the Damerau-Levenshtein distance over sequences of predictions of verbs, nouns and actions.

**B. Quantitative Results**

We report our submitted results for the challenge in table I based on the test set of the Ego4D LTA dataset. In this experiment, our framework predicted the $N = 8$ observed actions and the overall intention, to anticipate $Z = 20$ future actions by generating $K = 5$. Our framework slightly performs worse when anticipating verbs, but conditioning the generative model through the intention reduces the ED for nouns, thus performing better in the overall action anticipation possibilities.

Due to our limited computational resources, training was performed independently for each module. Next we will describe the quantitative evaluation first for the H3M module and after for I-CVAE.

**H3M.** Our model is able to classify verb-noun pairs with similar performance than provided baseline model, as reported in II but in our case only making use of the pre-extracted features. Due to the lower dimensionality of the features, several techniques were applied to avoid overfitting.

We applied gaussian noise injection (N) in the pre-extracted input features to improve robustness of the classifier. By adding multi-task approach and sharing inner layers among the tasks (M) we inherited implicit data augmentation: as there was a need of modelling a representation of three tasks (classifying intention, verb and noun) the model had to improve pattern extraction in the inner shared layers to optimize results for each task. Finally, focal loss modulation factor $(\text{beta}=0.99)$ for the cross-entropy loss was applied to address the class imbalance (I) of verbs, nouns and intentions.

In table II it is shown that applying M+I+N obtains the best Top1 accuracy results of our framework, but our model slightly decreases its performance regarding Top5 predictions. We claim that by dealing with imbalance of the dataset, most-frequent classes lose importance in the training loss and therefore are worst classified (with impact in accuracy). Finally, our model is limited by the low-dimensionality input data used (visual features pre-extracted) but is still able to approximate to the baseline.

We have also compared the influence of predicting intention correctly to action classification results, illustrated in table III Results show a significant direct relationship between nouns and intention, which intuitively indicates that by conditioning the framework through the intention, lower uncertainty can be obtained in nouns representations.

**I-CVAE.** We evaluate the performance of our stand-alone model based on the ground-truth action and intention labels provided by Ego4d dataset. We report ED@Z=20 as our evaluation metric for both nouns and verbs.

Next we will describe the influence of different strategies applied for the training, Action Embedder (AE) and Multi-Head Decoder (MD). The results, detailed in table IV are based on the validation set when $N = 8$ and $Z = 20$. In first experiments, embedding matrices from AE were initialized based on pre-trained S-BERT language model to capture the semantic representation of nouns and verbs. Experiments showed that training these layers from scratch resulted in better generalization capacity. Second experiment attempt to unify nouns and verbs inside the Transformer Encoder-Decoder structure. AE concatenates $a_{\text{past,emb,v}} \in \mathbb{R}^d$ and $a_{\text{past,emb,n}} \in \mathbb{R}^d$, to create embedded representations of the observed actions $a_{\text{past,emb}}$ and feed it into the Transformer Decoder. Separated nouns and verbs imply that the the predicted forecasting embedding $\hat{a}_{fut,emb} \in \mathbb{R}^d$ is split in

### TABLE I

|          | VERB | NOUN |
|----------|------|------|
| Baseline | 0.739 | 0.780 |
| Ours     | 0.752 (±0.013) | 0.748 (±0.032) |

|          | ACTION |
|----------|--------|
| Baseline | 0.930 (±0.013) |

**TABLE II**

|          | VERB | NOUN |
|----------|------|------|
| Ours (M+I+N) | 77.39 | 93.16 |
| Ours+I | 76.13 | 90.06 |
| Ours | 74.92 | 94.32 |
| Baseline | 72.08 | 88.70 |

**TABLE III**

|          | VERB | NOUN |
|----------|------|------|
| Correct | 20.13 | 54.32 |
| Error | 21.48 | 57.46 |

**TABLE IV**

|          | VERB | NOUN |
|----------|------|------|
| Correct | 19.48 | 57.46 |
| Error | 22.06 | 56.90 |

**TABLE V**

|          | VERB | NOUN |
|----------|------|------|
| Correct | 41.58 | 41.58 |
| Error | 38.79 | 33.28 |
verbs and nouns (inverse operation of AE concatenation), and each part is then projected to a verb/noun classification each through a fully connected layer. We observe that combining verb and noun as action representations allowed the model to avoid ‘out of sense’ pairs (some verbs-nouns do not have any semantic meaning together). Finally, two training tricks were used that improve results: addressing class imbalance through focal loss, as in H3M classification; and adding random noise to the ‘observed actions’ to create more robust forecasting even if some errors were found in the initial sequence. This second approach helped to improve the performance in end-to-end evaluation, when I-CVAE used H3M outputs to anticipate the next sequence of actions.

Results from table [IV] show the effectiveness of our model when we input \( N = 8 \) observed actions. In fig. 4 we illustrate the influence of reducing the observed number of actions. By only leveraging \( N = 4 \) past actions, I-CVAE is able to forecast with even higher performance than with longer past sequences. We claim that taking too many past information causes the model to focus on less-relevant cues, thus collapsing in less plausible action sequences. These behaviour is also seen in Fig. 5, where the ED of our model is plot for different number of observed actions \( N \) at each time-horizon of \( Z = 20 \). We can see the difficulty of anticipating noun changes correctly, which is very influenced by the time-horizon to forecast.

From Fig. 5 we can also observe that looking more than \( N = 4 \) actions in the past provides worse results in the short-term and converges in similar results in long-term. Our initial claim was to use intention to provide long-term guidance to the model. From Fig. 6 we can determine that conditioning the CVAE with the intention label allows to obtain better results in long-term future horizons, as it guides properly the model through the sequence. One limitation of our approach, however, is that when \( N = 8 \), the model performs slightly better without the condition from intention. We claim that the Transformer Encoder-Decoder is able to capture the intention cue and therefore does not need the label.

### IV. Conclusion

In this work, we propose two-module framework to efficiently exploit human intention for human action sequence anticipation. Our H3M module leverages multitask approaches to detect with adequate results the observed human actions and intentions, which then condition the future through our I-CVAE. Our framework is able to outperform state-of-the-art baseline.

### References

[1] Tolstikhin, Ilya O., et al. "Mlp-mixer: An all-mlp architecture for vision." Advances in Neural Information Processing Systems 34 (2021).

[2] P. Mathis, M J. Black, and G Varol. "Action-conditioned 3d human motion synthesis with transformer vae." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.