Asymmetric multi-task learning for interpretable gaze-driven grasping action forecasting

Iván González-Diaz, Miguel Molina-Moreno, Jenny Benois-Pineau, Aymar de Rugy

Asymmetric multi-task learning for interpretable gaze-driven grasping action forecasting

Iván González-Díaz, *Member, IEEE*, Miguel Molina-Moreno, Jenny Benois-Pineau, *Member, IEEE*, Aymar de Rugy

***Abstract*— This work tackles the problem of automatically predicting the grasping intention of humans observing their environment, with eye-tracker glasses and video cameras recording the scene view. Our target application is the assistance to people with motor disabilities and potential cognitive impairments, using assistive robotics. Our proposal leverages the analysis of human attention captured in the form of gaze fixations recorded by an eye-tracker on the first person video, as the anticipation of prehension actions is a well studied and well known phenomenon. We propose a multi-task system that simultaneously addresses the prediction of human attention in the near future, and the anticipation of grasping actions. In our model, visual attention is modeled as a competitive process between a discrete set of states, each one associated to a well-known gaze movement pattern from visual psychology. We additionally consider an asymmetric multi-task problem, where attention modeling is an auxiliary task that helps to regularize the learning process of the main action prediction task, and propose a constrained multi-task loss that naturally deals with this asymmetry. Our model shows superior performance than other losses for dynamic multi-task learning, current dominant deep architectures for general action forecasting and particularly-tailored models for predicting grasping intention. In particular, it provides state-of-the-art performance in three datasets for egocentric action anticipation, with an average precision of 0.569 and 0.524 in GITW and Sharon datasets, respectively, and an accuracy of 89.2% and a success rate of 51.7% in Invisible dataset.**

***Index Terms*— Grasping action forecasting, multi-task learning, interpretable attention prediction, constrained loss**

I. Introduction

INTENTION prediction has become a relevant task in many applications, especially in those that rely on human-robot and human-machine interaction as assistive robotics [1], shared control systems as neuroprostheses [2]–[4] or Advanced Driving Assistance Systems (ADAS). The reason is the strong coupling between humans' intention and their physical actions, since the former represent the human internal mental state that prompts and coordinates the execution of the later, initiating, guiding and controlling actions up to their completion [5].

Human intention can be predicted not only before the action actually begins but also during its execution. In fact, following the characterization proposed by Pacherie [6], which models temporal distances and relations between the intention and the corresponding action, intentions can be classified into distal (seconds to minutes), proximal (seconds) and motor intentions (hundreds of milliseconds). The first two are clearly anticipative at long/medium- and short-term, respectively, whereas motor intention starts just before and spans along full action execution. This division has an important impact on the temporal scale of the analysis, known by some authors as tempo [7], requiring an adjustment of the granularity based on the prediction horizon and the particular goals of the task.

In both proximal and motor scenarios, gaze is a paramount cue to predict human intention. The rationale behind this hypothesis is related to the notion of active perception [8] and intentionality [9], which can be defined as a person's commitment to perform a particular action [10], and requires skills such as foresight and planning. In particular, visual searchers can be considered as perceptual experiments to generate sensory data necessary to plan and execute physical actions in the very near-future. In this sense, Ballard et al. [11] proposed a 'just-in-time' fixation strategy stating that fixations that provide information for a particular action immediately precede that action and are crucial for a fast and economical execution of the task.

Furthermore, in the last few years, gaze sensors and, especially, wearable eye-tracking glasses that jointly capture egocentric visual field, gaze information and other physiological signals as egomotion, are coming to the market at very competitive prices, allowing the intensive study of human behavior and gaze dynamics in an ecological way, with subjects performing their daily activities [12].

Among the full repertoire of human actions, gaze is specially relevant for manipulation tasks, as they exploit the process eye-hand coordination [13]. In the light of this hypothesis, many approaches can be found in the literature [2], [14]–[17] that rely on the analysis of gaze dynamics to infer human intention and predict manipulation actions (anticipatedly or during their execution). However, most methods follow a bottom-up strategy: gaze is used as a sensory signal to compute features modeling the dynamics of visual attention, on top of which models for action prediction are developed.

This work, instead, further exploits the existent synergies between future gaze prediction and action forecasting. We

I. González-Díaz is with the Department of Signal Theory and Communications, Universidad Carlos III de Madrid, 28045, Spain, e-mail: igonzalez@tsc.uc3m.es.

Miguel Molina Moreno is with the Department of Immunobiology, Yale School of Medicine, USA. e-mail: migmolin@ing.uc3m.es

J. Benois-Pineau is with the Laboratoire Bordelais de Recherche en Informatique UMR 5800, 33400, University of Bordeaux, Talence, France. e-mail: jenny.benois@labri.fr

Aymar de Rugy is with the Institut de Neurosciences Cognitives et Intégratives d' Aquitaine, UMR 5287, CNRS and University of Bordeaux, 33076 Bordeaux, France. e-mail: aymar.derugy@u-bordeaux.fr

propose AMT-GAF: Asymmetric Multi-Task system for Gaze-driven grasping Action Forecasting. AMT-GAF is a multi-task model that simultaneously predicts future human visual attention and grasping intention. Our hypothesis is that subjects drive their attention based on the task being performed, so both concepts (visual attention and intentionality) are closely linked. In consequence, by addressing them simultaneously, we pursue a twofold goal: first, we aim to induce regularization over the shared representations and improve the performance of action forecasting; and second, we aim to enhance model interpretablity by establishing links between gaze dynamics and future actions. Regarding the second goal, we propose an interpretable system for future visual attention, in which attention is modeled through a discrete state model, in which states are associated to well-known eye-movement patterns from visual psychology [18].

Next we summarize the main scientific contributions of our work:

1) We introduce an interpretable model for task-based visual attention prediction. Attention is modeled as a competitive process between a discrete set of states, each one associated to a well-known gaze movement pattern from visual psychology. Future gaze is then computed from a combination of several spatial attention maps, each one modeling future attention under one of the considered states. In addition, we introduce the use of an additional cognitive ingredient: a short-term Visual Working Memory that stores information about objects that have attracted human attention in the near past.
2) We propose an asymmetric multi-task model for the simultaneous forecasting of grasping actions and future attention. We introduce a novel constrained loss that considers action forecasting and gaze prediction as primary and auxiliary tasks, respectively, and successfully handles challenging cases where providing an accurate attention map from candidate distributions is hard.
3) We present a new dataset, SHARON, for natural grasping action forecasting that is complementary to previous datasets in the field, focusing on efficient system deployment and anticipated action prediction. The SHARON dataset will published with the paper to promote the research in the field.
4) We provide a comprehensive set of experiments in three datasets to assess our technical contributions, and compare our approach with other losses for dynamic multi-task, state-of-the art architectures for general action forecasting, and other models particularly designed for natural grasping intention prediction.

The remainder of this paper is as follows: in section II we analyze the most relevant related work in the fields of gaze-based intention estimation and multi-task learning. Section III describes our problem and potential application scenarios. Section IV is the central section of the paper, where we present our proposed interpretable model for task-based visual attention and the constrained loss for asymmetric multi-task learning. In section V we present an experimental analysis of our method and compare it with other state-of-the-art approaches from the literature. Section VI concludes this work and outlines its perspectives.

II. Related Work

A. Gaze-driven intention estimation

Many works in the literature use gaze to dig into human internal mental states and discover short-term (proximal, motor) intentions. In [19] gaze is used to predict human intention with subjects interacting with virtual environments. The work in [20] combines gaze and model-based AI planning to build probability distributions over a set of possible intentions during a multi-player board game. Some works [15], [16], [21] have shown that using gaze to predict intention during tele-operation tasks provides great advantages in reducing an operator's workload and a task's difficulty as well as enhancing the task performance. Recently, gaze has also been considered a fundamental cue to enable robots with predictive capabilities and support a more natural Human-Robot-Interaction (HRI), as in [22] and [23] for cooperative beverage preparation, and in [1] for more complex breakfast recipes. The interested reader is referred to the work of Bellardinelli [12] for an excellent discussion of cognitive theories exploring the relation between gaze and intention and an extensive survey of computational approaches using gaze to predict human intention.

Some other previous works fit particularly well our scenario of grasping intention using gaze as a fundamental cue. In [24] the authors combine features modeling gaze dynamics, a motion model using the Kalman filter, and different classifiers, to predict human intention in the form of interacting object and action. The same authors extend their proposal in [17], and combine Yolo-based object detectors and sequence modeling (HMMs and RNNs) techniques to predict grasping intention. This work additionally introduces a novel dataset, called Invisible, which deals with the relevant Midas problem by including videos with both grasping and only viewing actions. The Midas problem, firstly introduced and addressed in the work of [25], states that not everything that we gaze upon, is something we want to interact with.

All discussed methods focus on motor intention, and consider as valid segments for intention prediction those that go from the moment just before the object is being grasped until the end of the action execution. In this work, however, we put more focus on grasping action *forecasting*, which implies estimating intention before the action actually takes place (proximal and anticipative motor intention). We consider that this goal adapts more naturally to scenarios in which the action cannot be carried out by humans (e.g. shared control of neurosprostheses for amputees or assistive robots bringing utensils closer to impeded humans). Previous works, as [14] and [2], tackle similar problems by combining active object detectors with Recurrent Neural Networks to provide grasping action forecasting.

Furthermore, previous methods limit the use of gaze as a sensory signal to compute features modeling the dynamics of visual attention, on top of which models for intention prediction are developed. Our proposal, on the contrary, incorporates gaze in the process of intention prediction more ambitiously:

we also aim to predict the future visual attention and exploit the existent synergies between both tasks: action forecasting and visual attention prediction. In other words, we think that anticipating how humans will use their gaze will help to understand their intentions and forecast their actions. To the best of our knowledge, this is the first time that both tasks are addressed simultaneously in a unified model, through the use of a novel asymmetric multi-task loss with constraints.

B. Multi-task learning in action-prediction

Multi-task Learning (MTL) is a technique that aims to concurrently learn multiple objectives from shared representations [26]. In some scenarios, like object detection [27], MTL is inherently required due to end-user task definition, which is decomposed in several necessary subtasks (e.g. object category identification and bounding box regression). In other cases, losses are combined in MTL to provide complementary views of the same task: e.g. in semantic image segmentation, cross-entropy, Dice or Jaccard losses are often linearly combined to account for different factors influencing segmentation performance [28]. Finally, there are cases where MTL is proposed to leverage the synergies between closely-related tasks and improve the performance with respect to single-task learning [29]. In those cases, it is usual to distinguish a primary task and a set of auxiliary tasks. The latter are incorporated into the learning process to regularize learning and obtain better shared representations. Thus, the performance of the primary task is improved.

Action prediction is a task where MTL is prevalent. In [30] action prediction and estimation of the temporal progress are tackled together within a MTL approach. The work in [31] addresses the joint long-term forecasting of future actions and their duration. Future-transformer [32] combines the tasks of temporal action segmentation in the past with the prediction of the actions in the future, and Self-Regulated Learning (SRL) [33] incorporates a contrastive loss to generate future spatio-temporal features on top of which action forecasting is made. The work in [34] is close to our proposal as it aims to simultaneously predict semantic actions and generate trajectories representing future human motions. Our work, however, focuses on the predicting future eye-gaze movement (instead of head and body motion) and relies on a set of interpretable and well-known gaze-motion patterns from visual psychology.

All the discussed multi-task methods follow a shared-bottom structured model [35], and either simply sum individual losses to form the combined multi-task loss, or perform linear combinations where weights are fixed a priori. Some limitations arise in these simple approaches: first, the performance of such systems is strongly dependent on the relative weighting between each task's loss, requiring extensive empirical tuning; second, task weights are often static throughout the course of training, potentially diverting training resources to unnecessary tasks or samples [36]; third, shared-bottom model structures (e.g. architecture with shared bottom layers, and then parallel task-dependent top layers), although reduce the risk of overfitting, can suffer from optimization conflicts caused by task differences.

Regarding the third problem, several works have designed specific architectures that leverage the common elements and identify differences between tasks: the works in [35] and [37] propose to use Mixture-of-Experts (MoE) architectures, which explicitly learn to model task relationships from data that can be also combined with multi-task losses to leverage these relationships. These architectures make use of a set of experts, each one providing an alternative representation of the data, which are then combined using a weighted linear combination that is task-dependent. Hence, two tasks can both assign high weights to those experts modeling their similarities and different weights to track their particular aspects. Of similar inspiration, the work in [38] proposes a encoder-decoder architecture, in which the encoder is shared by all tasks being addressed, and a hybrid decoder is implemented through attention modules, including a shared block that exploits inter-task dependencies and a set of individual blocks that obtain particular task representations. The use of mixture-of-experts is a complementary solution to our proposal and could be easily plugged into our architecture with slight adaptations. However, our scenario of asymmetric multi-task learning, which aims to minimize overfitting and increase regularization in the representation, together with the requisite of real-time execution in our application, suggests adopting of simpler shared-bottom model structures and focusing on the other two limitations stated.

To overcome the first two limitations, several methods have been proposed that aim to dynamically set the mixing coefficients of the linear combination of losses. Self-paced MTL [39] considers both sample and task difficulty to automatically compute the mixing weights for samples and tasks in a multi-task problem. In [40] the authors proposed to weigh multiple loss functions by considering the homoscedastic uncertainty of each task, and apply their method to a combined problem of semantic image segmentation and pixel depth estimation from monocular images. Dynamic task prioritization [36] automatically prioritizes more difficult tasks by adaptively adjusting the mixing weight of each task's loss objective, where difficulty is inversely proportional to performance (measured through differentiable KPIs). The method is successfully applied to combined problems of visual classification, segmentation, detection, and pose estimation.

The work in [41] is of special closeness to our approach, as it considers an asymmetric MTL, in which one task is considered a primary, whereas the others are auxiliary. In this proposal, task weighting is dynamically optimized to reduce negative transfer in multi-task learning: by checking when adding auxiliary losses decreases the performance compared to the single-learning with the primary loss, the method adjusts the mixing weights accordingly. Our problem is also asymmetric, with a well-defined primary task of action forecasting, and the supporting task of future gaze prediction. Our approach, however, diverges to deal with a particular challenge in our scenario: in our model, attention prediction relies on a set of pre-defined spatial distributions, each one corresponding to a well-known gaze motion pattern from psychology. As sometimes none of the candidate maps points to the true future gaze location, achieving accurate predictions is very hard

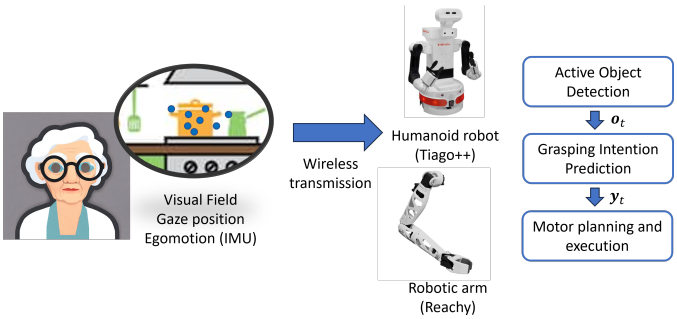

Fig. 1. An overview of our scenario of HRI with predictive robotics.

in these cases and cannot be realistically addressed without leading to overfitting. Consequently, we have developed a constrained loss that simply aims to perform well enough in the auxiliary task.

III. Application Scenario

Our application scenario of HRI enabled by predictive robotics is depicted in Fig. 1. We aim to help people with motor disabilities and potential cognitive impairments, using assistive robotics that naturally bring them out-of-reach objects while they are performing activities of daily living (ADL). We envisage the use of different types of robots, from neuroprostheses (e.g. Reachy[1] [42]) to humanoid robots (e.g. as Tiago++[2]) and, in order to bring a more natural and efficient collaboration with humans, we limit the use of voice commands. Instead, we resort to multimodal signals, as electromyograms, exocentric vision (e.g. advanced pose and joint-bone analysis as in [43]) or egocentric vision, to build *predictive robots* that decode human intention and act proactively without the need of explicit human requests. In this work, we focus on egocentric vision, and equip subjects with Pupil Invisible glasses[3], a wearable multi-sensor that incorporates: a) a camera recording their visual field at 30Hz, b) an eye-tracker that captures their gaze at 120Hz, c) an IMU unit to record egomotion caused by head and body movements, and d) a microphone. The glasses also come with a prescription lens kit that accommodate a variety of wearers and correct a wide diopter range, a relevant requisite to work with senior patients.

The sensed data are transferred via wireless connection to the assistive robot, which performs real-time processing and consequent execution of physical actions to help humans. The Artificial Intelligence (AI) that governs the robot is made up of three processing modules. The first one is the *Active Object Detection (AOD)* module. Video frames showing the visual field of the human and the corresponding gaze locations are its input. The AOD, using gaze as a guiding signal, identifies the active object among all elements in the scene, i.e., the object that the user is interacting with or aims to interact with. The AOD module operates on a frame-by-frame basis, providing a vector $\mathbf{o}_t \in \mathbb{R}^{C_o+1}$, where t is the time instant, and C_o is the number of object categories considered in our problem (we include an additional category 'background' to account for cases where attention is not directed to any object of interest). This vector encodes the probability of each object category being the active object at the current instant. The AOD module is out of the scope of this work, as it has been already described in detail in previous works (the interested reader is referred to [2], [14]). Here we simply highlight that this module is trained using video-label weak annotations indicating the active object for a video sequence, and using gaze as guiding signal, thus avoiding the need of time-consuming annotation though bounding boxes. Henceforth, a user can easily train the detector for a new acquired object by simply looking at it for a brief time, from different angles and viewpoints, and indicating the object category (e.g. using a voice command like 'Train new object with category frying pan').

The outputs of the AOD, along with the remaining sensed data, are passed to the second module, the *Grasping Intention Prediction (GIP)* module. It identifies when the human wants to grasp and manipulate an object that is out of reach. The GIP generates a new vector $\mathbf{y}_t \in \mathbb{R}^{C_o+1}$ containing the grasping action probabilities. We envisage grasping action as a multiclass problem with C_o+1 categories: one category for the non-grasping action (e.g. the user does not need to grasp any object) and the others indicating that the user aims to grasp each of the considered C_o object categories. The GIP system represents the central contribution of this paper and will be thoroughly described in section IV.

Finally, the vector $\mathbf{y}_t$ is passed to our *Motor Planning and Execution (MPE)* module that, when a grasping action is required, plans its physical execution and carries out the action, grasping the object or bringing it close to the patient. The MPE is however out of the scope of this paper, including the physical execution of the grasping actions.

IV. AMT-GAF: Asymmetric Multi-Task learning for Gaze-driven grasping Action Forecasting

A. General description of the module

AMT-GAF: Asymmetric Multi-Task system for Gaze-driven grasping Action Forecasting is a multi-task model that simultaneously predicts future human visual attention and grasping intention. Our hypothesis is that subjects drive their attention based on the task being performed, so both concepts (visual attention and intentionality) are closely linked. In consequence, by addressing them simultaneously, we aim to improve the performance of a baseline isolated action prediction system. Furthermore, our work aims to scientifically contribute towards the design of more human-centric AI [44] in two ways:

1) *Providing more interpretable AI tools*, so that the robot actions can be better understood by non-expert humans, thus improving AI trustworthiness. In this regard, we have designed an interpretable model that defines task-based human attention through a set of eye movement patterns that are well-known in visual psychology.
2) *Providing AI tools that incorporate ingredients from cognitive sciences and emulate the functioning of the human natural intelligence*: in our case, besides the eye movement patterns, we have incorporated a short-term *Visual Working Memory* that enables efficient redirection

[1]https://www.pollen-robotics.com/reachy/
[2]https://pal-robotics.com/es/robots/tiago/
[3]https://pupil-labs.com/products/invisible/

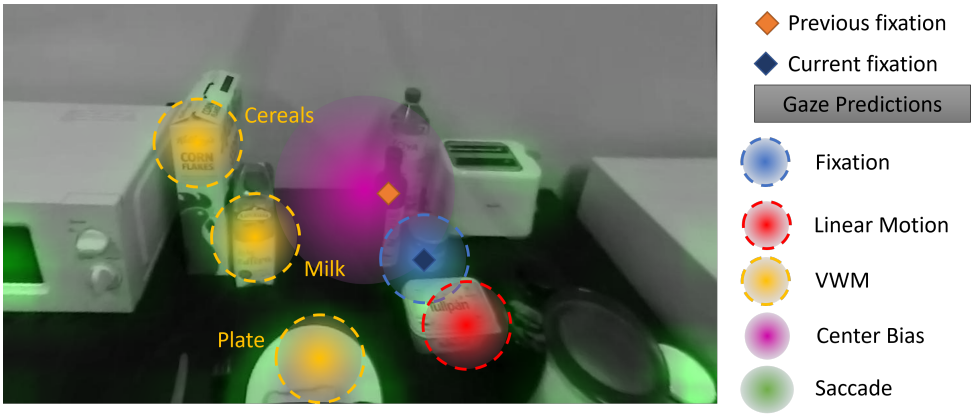

Fig. 2. Interpretable model for task-based visual attention. Each state is associated to a probabilistic spatial map predicting future gaze location.

of visual attention to objects that have been fixated in the recent past.

In the following sections, we will thoroughly describe each scientific contribution: the interpretable model and the VWM will be described in detail in section IV-B and our multi-task architecture will be detailed in section IV-C. Finally, in section IV-D we will introduce a novel loss for multi-task learning, besides considering the asymmetry between both tasks, naturally deals with situations where the estimation of the future attention is poor and the overall performance is degraded.

B. An interpretable model for task-based visual attention prediction

It is known that humans use their gaze to generate the necessary sensory data to guide their perception of the environment [45], and thus be able to decide what to do next. Given this close relationship between perception and intention, we hypothesize that understanding how subjects guide their visual attention to the different elements in the scene will provide clues, not only about the actions they are currently carrying out, but also about their intentions and future actions. In consequence, besides simply decoding human intention, we also aim to forecast the spatial location of gaze.

To do so, and since one of our goals is to achieve a system that is interpretable and understandable for humans, we have modeled the dynamics of task-based attention as a competitive process between a *finite set of eye movement patterns*. Each pattern is a potential state that defines the dynamics of the gaze and sets a particular spatial distribution of future attention along the visual field of the subject. We have considered motion patterns that are well-known in visual psychology and, therefore, become interpretable and understandable for humans. Furthermore, we treat the state encoding future visual attention as a latent variable (we will not have labels indicating the gaze state at each moment). It must be inferred from other observed information as the position of the gaze in the current frame and other sensed data. The goal is to learn a model that automatically identifies the current state of attention and selects the corresponding spatial map to predict the gaze location in the near future.

We next describe the states of visual attention, indicating how they drive gaze and produce eye movements, as well as their corresponding spatial distributions of future visual attention, which are further depicted in Fig. 2.

1) *Fixation (including microsaccades)*: In this case, we model visual attention during a *fixation*, in which a subject fixes his gaze over a stable area covering a particular element in the scene [46]. During a fixation, gaze shifts are only due to micro-saccadic movements (very short shifts) towards different points within the element of interest [47]. When visual attention is in this state, the candidate gaze map is built upon the current fixation, defining a two-dimensional Gaussian distribution centered at its location. In Fig. 2, it can be seen how the map proposed by this state is centered on the current fixation (e.g. blue Gaussian centered on the blue diamond).
2) *Predictable Eye Movements*: this state covers predictable eye movements that may happen in two scenarios: a) smooth pursuits: eyes move to maintain a moving object of interest on the fovea [18]; and b) slow saccadic movements, which allow subjects to perform a scanning of the scene. In order to generate the corresponding attention map, we predict the future gaze position using a constant-velocity model, so that we first compute an eye motion vector (see Eye Movement definition in sec. IV-C), and use it to shift the current fixation and define a two-dimensional Gaussian distribution (see red Gaussian at Fig. 2). It is noteworthy that, when eye movements are negligible, this map is similar to the Fixation one.
3) *Visual Working Memory* (VWM): the VWM, which will be explained below, allows us to store the locations of objects that have previously attracted subject's attention. Each memory position will produce a candidate attention map for future frames, emulating when subjects drive their attention back to an object that they have previously identified (see the yellow Gaussian distributions located at milk, cereals and plate in Fig. 2.
4) *Center Bias*: Human eye-tracking studies have shown that gaze is often biased towards the center of natural scene stimuli ('center bias') [48]. We model this dependency with a large Gaussian located at the center of the visual field (pink Gaussian at Fig. 2).
5) *Fast saccadic movements*: saccades are rapid, ballistic

eye movements between fixations that bring an area of the visual scene onto the fovea [46], and model those instants in which subjects want to drive their attention to a new element in the scene and shift their gaze quickly to another area which as not been yet fixated. This movement is therefore highly unpredictable using previous information and instead driven by competitive mechanisms of visual saliency. Hence, the most striking regions compete to attract the subject's attention based on their low-level characteristics (color, lighting, structure). In this work, a bottom-up attention map is generated using a low-level visual saliency algorithm [49] and, subsequently, the areas covered by Fixation, Predictable Eye Movements and VWM maps are masked to remove the influence of previously attended elements in the scene. In Fig. 2 the probability map is represented in green tones.

It is known that the VWM plays an important role in task-based visual attention. In this work, we have designed a VWM inspired by [50]: 1) It is a short-term memory that stores visual and spatial information of 4-5 relevant objects that have been previously fixated by the human; and 2) each object stored in the memory has an associated attentional weight to be used in competitive processes of attention and during visual searches. From these premises, we have designed a tabular VWM in which each entrance i is defined through the following fields:

- *Object category*: Category of the stored object c_i, with $c_i = 1...C_o$.
- *Object location*: coordinates $\mathbf{x}_i$ of the object center with respect to the current fixation. Coordinates of previously stored objects are continuously updated using a geometric alignment module (see [2] for additional details) so that they are always relative to the current fixation.
- *Attentional weight*: the attentional weight w_i will be used in competitive processes of attention, and is also updated for each new instant t modeling short-term persistence:

$$w_{i,t} = \max(\gamma w_{i,t-1}, o_{c_i,t}) \tag{1}$$

where $o_{c_i,t}$ is the score of the object category c_i in the current instant t, and γ is a memory factor that sets the speed at which objects are forgotten and removed from the memory in case they are not fixed again. We have heuristically set its value to $\gamma = 0.9$, which roughly corresponds to a memory that keeps objects during approximately 1 second (at 30fps) in case they are not longer re-fixated.

In practice, we set a memory of size of 5 entrances, so we just consider the objects with the highest attentional weight among the considered categories. We use this VWM with two purposes: a) to generate candidate attention maps, as explained above; and b) to model subjects' intention and forecast their actions.

Once the states have been conceptually defined together with their associated attention spatial maps, we next introduce the mathematical model of interpretable task-based future attention prediction. We model the state of visual attention through a latent discrete variable that takes S potential values

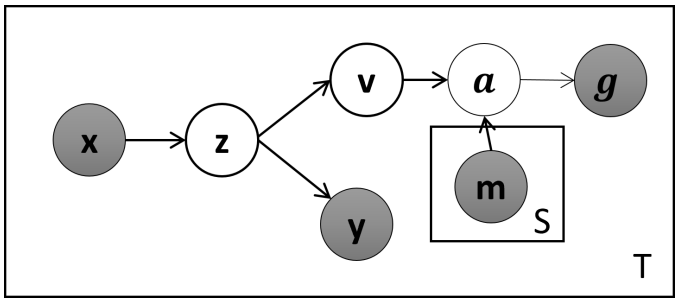

Fig. 3. Graphical representation of our probabilistic AMT-GAF for simultaneous prediction of visual attention and grasping intention. Nodes represent random variables and arrows stand for dependencies. White nodes are latent variables to be inferred, while shaded nodes are observable variables. Boxes mean repetitions of a process/variable.

(the number of considered states or patterns of eye movement) based on the probabilities stored in the vector $\mathbf{v}$. Next, the spatial distribution of future visual attention $\mathbf{a} \in \mathbb{R}^{H \times W}$ is modeled using a probabilistic mixture model over the S latent states (for simplicity, we omit temporal index t):

$$\mathbf{a} = \sum_{s=1}^{S} v_s \mathbf{m}_s \tag{2}$$

where $\mathbf{m}_s \in \mathbb{R}^{H \times W}$ is the candidate map associated to the candidate state s (as shown in Fig. 2). In other words, the final attention map is generated as a linear combination of the candidate maps in which the mixing coefficients are taken from the discrete state variable $\mathbf{v}$. In order to leverage gradient-descent learning over the parameters of the discrete distribution ($\mathbf{v}$), we apply the re-parametrization trick with a Gumbel-Softmax distribution [51]. This distribution represents a soft-approximation of the maximum operator and enables the competition between attention states. Furthermore, let us note that, thus defined, the model is interpretable, as the weights v_s to learn express the importance of each attention component.

Finally, the map $\mathbf{a}$ is used to predict the coordinates of the future gaze location $\tilde{\mathbf{g}}_{t+\Delta} = f_g(\mathbf{a}_t)$ for a future instant $t' = t + \Delta$, computed as a nonlinear regression $f_g(\cdot)$. Here Δ represents the anticipation step, which has been heuristically set to 0.2 secs (6 frames at 30fps), to consider significant but yet predictable shifts in the gaze location.

C. Architecture of AMT-GAF

Once we have presented our interpretable model for visual attention, we now introduce our proposed probabilistic multi-task model for simultaneous prediction of future visual attention and grasping intention, called AMT-GAF. The graphical model of AMT-GAF is shown in Fig. 3, where shaded nodes represent observable and white nodes represent hidden variables, respectively. We rely on two latent variables: a) the discrete variable $\mathbf{v}$ that models the underlying state of the visual attention; and b) a continuous latent variable $\mathbf{z}$, shared by both tasks in our multi-task model, which encodes the dynamics of the visual field and the human intention.

For each instant $t = 1...T$, our model receives an input $\mathbf{x}_t$ that contains a concatenation of the following variables:

1) *Current Gaze location* $\mathbf{g}_t \in \mathbb{R}^2$: Assuming the importance of the central bias [48], we consider that users

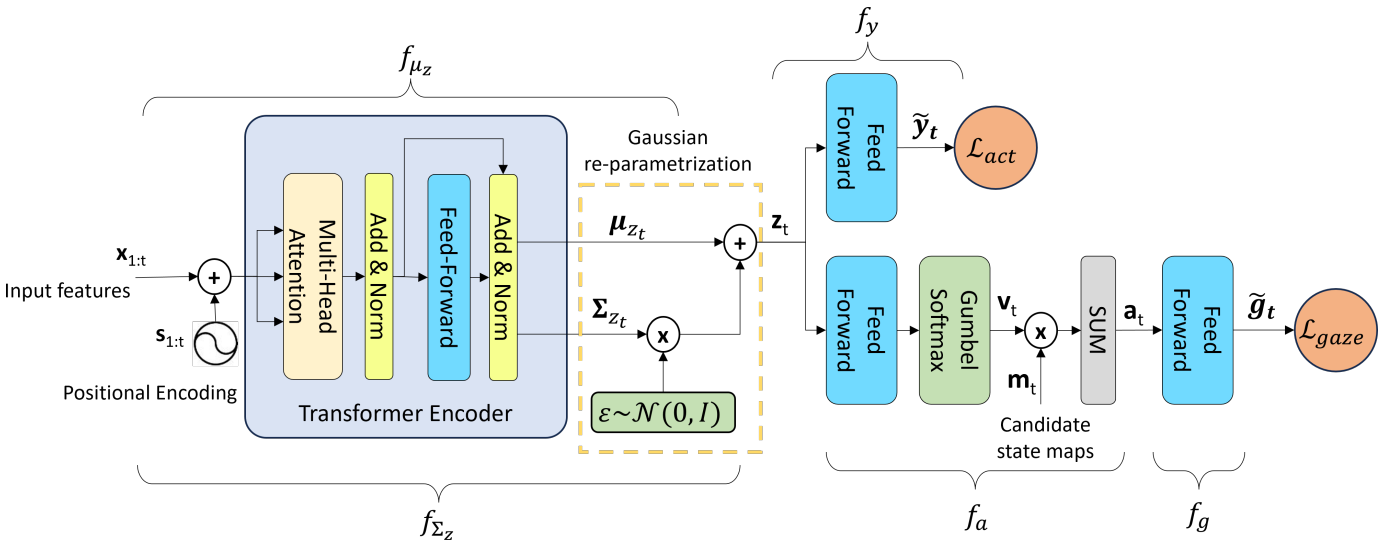

Fig. 4. Computational architecture that implements the graphical model proposed in Fig. 3. Mean $\boldsymbol{\mu_{zt}}$ and covariance $\boldsymbol{\Sigma_{zt}}$ of the latent state $\mathbf{z_t}$ are generated by a Transformer Encoder [52], which is fed by the set of previous observations $\mathbf{x_{1:t}}$ and a positional encoding variable $\mathbf{s_t}$. Then, our multi-task model implements two processing branches, one to predict the actions $\hat{\mathbf{y}}_t$ and one to predict the future gaze coordinates $\hat{\mathbf{g}}_t$. Feed-forward blocks are implemented as stacks of Linear layers and Relu activations. Add & Norm blocks implement residual layer, that add original and processed signals, followed by a layer normalization. SUM means summation over channels for each spatial position of the attention maps. Gaussian and Gumbel-Softmax re-parametrizations, as well as both losses (red circles), are explained in text.

tend to center their field of view on the active object when they aim to interact with it. Note, this does not happen under other states of visual attention, such as scene scanning.

2) *Eye movement* $\mathbf{u}_t \in \mathbb{R}^2$: we measure eye movements by subtracting consecutive gaze locations $(t-1, t)$. However, in order to remove the influence of the egomotion and vestibulo-ocular movements (eye movements that compensate head motion) [53], we first need to geometrically project the previous point $\mathbf{g}_{t-1}$ with respect to the current frame, yielding $\mathbf{g}^t_{t-1}$ [2]. Then, the eye motion vector is $\mathbf{u}_t = \mathbf{g}_t - \mathbf{g}^t_{t-1}$. Eye displacement is a physiological measure of interest to know user intention, and helps to identify the different types of eye movements/attentional states (scanning of the scene, saccadic shifts between objects, fixations over the object of interest, smooth pursuit over hands during manipulation, etc.)
3) *Ego-motion* $\mathbf{v}_t \in \mathbb{R}^2$: we additionally include the motion of head and body of the human wearing the glasses to better understand their physical interaction with the environment (approaching objects, stabilizing pose before grasping an object, manipulating objects, etc.). We compute ego-motion by subtracting the eye movement from the total gaze displacement between consecutive instants: $\mathbf{v}_t = \mathbf{g}_t - \mathbf{g}_{t-1} - \mathbf{u}_t = \mathbf{g}^t_{t-1} - \mathbf{g}_{t-1}$.
4) *Vector of active objects* $\mathbf{o}_t \in \mathbb{R}^{C_o}$: provided by the AOD, as explained in section III.
5) *Information from the VWM*: using the information contained in the VWM, we build two additional input vectors: $\mathbf{w}_t \in \mathbb{R}^{C_o}$: which contain the attention weights for each object category (with zero values if an object is not encoded in the VWM); and $\mathbf{d}_t \in \mathbb{R}^{C_o}$: the Euclidean distance between the object locations and the image center.

For each instant $t = 1...T$ during HRI, our model is fed with the concatenated input $\mathbf{x}_t \in \mathbb{R}^{3C_o+6}$ and the S spatial attention maps $\mathbf{m}_{it}, i = 1...S$, each one associated with one candidate state of attention, and performs the following generative process:

1) It samples the latent shared state variable $\mathbf{z}_t$ from a Gaussian distribution:

$$\mathbf{z}_t \sim \mathcal{N}(\mu_{zt}, \Sigma_{zt}); \text{ with} \quad (3)$$
$$\mu_{zt} = f_{\mu_{zt}}(\mathbf{x}_{1:t}) \quad (4)$$
$$\Sigma_{zt} = f_{\Sigma_{zt}}(\mathbf{x}_{1:t}) \quad (5)$$

where we can observe that the parameters of the distribution at the instant t depend on the inputs in the current and previous instants $\mathbf{x}_{1:t}$. Distribution parameters are computed applying learnable non-linear transforms $f_{\mu_{zt}}$ and $f_{\Sigma_{zt}}$.
2) Next, it samples the latent attention state $\mathbf{v}_t$ from a discrete distribution, approximated by a Gumbel-Softmax distribution [51] $\mathcal{G}$:

$$\mathbf{v}_t \sim \mathcal{G}(f_a(\mathbf{z}_t)) \quad (6)$$

where the vector of probabilities in the distribution is computed as a non-linear transformation f_a of the shared state variable $\mathbf{z}_n$
3) It generates the spatial distribution of future visual attention $\mathbf{a}_t$, using a mixture model over the states (see eq. (2).
4) It predicts the coordinates of the gaze location $\tilde{\mathbf{g}}_{t+\Delta} = f_g(\mathbf{a}_t)$ applying a nonlinear regression f_g over the spatial distribution $\mathbf{a}_t$.
5) It predicts the vector of probabilities of the next grasping action $\tilde{\mathbf{y}}_t$ as $\tilde{\mathbf{y}}_t = f_y(\mathbf{z}_t)$, where the variable $\tilde{\mathbf{y}}_t$ follows a discrete distribution in which the probability of an action is computed from the shared latent state variable $\mathbf{z}_t$ using a nonlinear transform f_y.

Figure 4 brings details about the computational architecture that implements the graphical model proposed in Fig. 3. $f_{\mu_{zt}}$ and $f_{\Sigma_{zt}}$, in charge of modeling the latent shared state $\mathbf{z}_t$, are both implemented with a shared Transformer Encoder [52], which is fed by the set of previous observations $\mathbf{x}_{1:t}$ and a positional encoding variable $\mathbf{s}_t$. This network generates the mean μ_{zt} and the covariance Σ_{zt} of the Gaussian variable $\mathbf{z}_t \in \mathbb{R}^{512} \sim \mathcal{N}(\mu_{zt}, \Sigma_{zt})$ representing the latent state of the human visual dynamics.

To allow optimizing the model parameters using gradient-descent methods, we again leverage the re-parametrization trick, sampling an auxiliary variable ϵ from a normal distribution $\epsilon \sim \mathcal{N}(0, I)$, and applying an affine transform so that $\mathbf{z}_t = \mu_{zt} + \Sigma_{zt} \cdot \epsilon$. We found out that modeling $\mathbf{z}_t$ as a random variable is especially effective in our scenario, in which data augmentation through random transformation of input variables is hard to apply due to the strong inter-dependence between the different sensed signals.

During inference, we simply set $\mathbf{z}_t \sim \mu_{zt}$ to remove noise and maximize performance. The latent state is in turn passed through two different feed-forward networks: a) the network f_y addresses the main task, forecasting the grasping action $\mathbf{y}_t \in \mathbb{R}^{C_o+1}$; and b) a sequential network composed of two processing blocks, f_a and f_g, that addresses the auxiliary task. It computes the fused map of visual attention $\mathbf{a}_t$ and the coordinates of a future $(t + \Delta)$ gaze location in $\tilde{\mathbf{g}}_{t+\Delta}$, respectively.

D. A constrained loss for asymmetric multi-task learning

The multi-task architecture of AMT-GAF is trained in a unified form, using a loss that combines two individual losses, each one associated with a task of interest. Without loss of generality, we will consider that each training batch contains just one video so that we can avoid video indexes that will increase the complexity of the descriptions unnecessarily. The extension to multi-video batches is nevertheless straightforward.

Given a video with T temporal instants, we consider: 1) an *action-prediction loss* $\mathcal{L}_{act}(\tilde{\mathbf{y}}, \mathbf{y})$, which compares the set of T grasping action predictions $\tilde{\mathbf{y}}$ and the corresponding labels $\mathbf{y}$; and 2) a *gaze-prediction loss* $\mathcal{L}_{gaze}(\tilde{\mathbf{g}}_\Delta, \mathbf{g}_\Delta)$, which compares, for an anticipation step Δ, our gaze predictions $\tilde{\mathbf{g}}_\Delta$ with the corresponding ground truth annotations $\mathbf{g}_\Delta$.

We use a $\mathcal{L}_{act}$ implemented through a set of binary cross-entropy losses (one per action category, using sigmoids to transform unbounded outputs into probabilities), that incorporates sample weighting to robustly handle two challenges that are particularly relevant in our scenario: inaccurate active object predictions from the AOD and temporally weak annotations for frame-level grasping action prediction. The interested reader is referred to [2] for a detailed description of the loss.

In addition, $\mathcal{L}_{gaze}$ has the form:

$$\mathcal{L}_{gaze}(\tilde{\mathbf{g}}_\Delta, \mathbf{g}_\Delta) = \frac{1}{T-\Delta} \sum_{t=1}^{T-\Delta} ||\tilde{\mathbf{g}}_{t+\Delta} - \mathbf{g}^t_{t+\Delta}||^2 \quad (7)$$

$\mathcal{L}_{gaze}$ is the mean square error between frame-level predictions $\tilde{\mathbf{g}}_{t+\Delta}$ and annotations $\mathbf{g}^t_{t+\Delta}$ (let us note that ground truth annotations in $t+\Delta$ are aligned with respect the current frame t).

A basic solution enabling multi-task learning consists of a *linear combination* of both losses, with a fixed α parameter that controls the influence of each term:

$$\mathcal{L}_{lin}(\theta, \alpha) = \mathcal{L}_{act}(\tilde{\mathbf{y}}, \mathbf{y}; \theta) + \alpha \mathcal{L}_{gaze}(\tilde{\mathbf{g}}_\Delta, \mathbf{g}_\Delta; \theta) \quad (8)$$

where we have included θ, which represents the learnable parameters of our model (e.g. the weights in the neural networks). Our hypothesis is that the minimization of $\mathcal{L}_{gaze}$ will help to regularize the latent representation of the visual dynamics $\mathbf{z}_t$, as we include additional but closely related information to drive the learning process. During test, this branch might be neglected without any impact on the performance. However, we consider that the gaze prediction branch deserves being computed in test as it actually enhances the model interpretability, as we will show in the experimental section.

In our scenario of application, the prediction of actions represents our primary focus, whereas the prediction of future attention remains an auxiliary task. Besides, our explainable model, in which the next visual attention is predicted from a series of candidate maps, may lead to situations where none of the candidate maps points to the true gaze location, providing poor estimations $\tilde{\mathbf{g}}_\Delta$ of the future attention. During optimization, we observed that these cases produce large loss values and gradients (specially after several epochs of training, when models have already been partially adjusted) that dominate the learning process in detriment of the primary loss. This phenomenon leads to degradation in the system performance. To overcome this issue, we have replaced the linear combination of losses with a *novel constrained loss function for asymmetric multi-task learning*. Hence, for each training video, we aim to solve the following optimization problem:

$$\begin{aligned} &\underset{\theta}{\text{minimize}} \;\; \mathcal{L}_{act}(\tilde{\mathbf{y}}, \mathbf{y}; \theta) \\ &s.t. \;\; \sum_{t=1}^{T-\Delta} \mathcal{L}_{gaze}(\tilde{\mathbf{g}}_{t+\Delta}, \mathbf{g}^t_{t+\Delta}; \theta) \leq TH_{gaze} \end{aligned} \quad (9)$$

With this new approach, we aim to minimize the primary loss (the action loss) subject to obtaining a gaze prediction that is good enough (below a threshold TH_{gaze}). If the constraint is hold for a video, the gaze loss no longer affects the optimization, which will primarily focus on action forecasting. This approach prevents from indiscriminately learning weights that aim to optimize an ill-posed problem, where the available candidate maps do not allow to find a solution for the future gaze.

To provide a solution for the problem in eq. (9), we re-define the constrained problem in its Lagrangian form, leading to a multi-task loss $\mathcal{L}_{mt}$:

$$\begin{aligned} \mathcal{L}_{mt}(\theta, \lambda) &= \mathcal{L}_{act}(\tilde{\mathbf{y}}, \mathbf{y}; \theta) \\ &+ \lambda \left(\sum_{t=1}^{T-\Delta} \mathcal{L}_{gaze}(\tilde{\mathbf{g}}_{t+\Delta}, \mathbf{g}^t_{t+\Delta}; \theta) - TH_{gaze} \right) \end{aligned} \quad (10)$$

where $\lambda \geq 0$ is the Lagrange multiplier associated to the problem restriction. The dual function corresponding to these problem is:

$$g(\lambda) = \inf_{\theta} \ \mathcal{L}_{mt}(\theta, \lambda) \tag{11}$$

And the dual optimization problem is:

$$\underset{\lambda}{\text{maximize}} \ \ g(\lambda) \tag{12}$$

which is concave with respect to the Lagrange multiplier λ, and allows us to use a Projected Gradient Ascent (PGA) for optimization, only requiring the computation of the gradient as:

$$\frac{\partial g(\lambda)}{\partial \lambda} = \sum_{t=1}^{T-\Delta} \mathcal{L}_{gaze}(\tilde{\mathbf{g}}_{t+\Delta}, \mathbf{g}^{t}_{t+\Delta}; \theta) - TH_{gaze} \tag{13}$$

The optimal solution of the dual problem is a lower bound for the solution of the primal formulation (concept known as weak duality). In our problem, $\mathcal{L}_{act}$ (see [2]) and $\mathcal{L}_{gaze}$, although both are convex with respect to the network outputs $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{g}}_{t+\Delta}$, they are generally not with respect to the network parameters θ, something characteristic in deep learning architectures. In consequence, strong duality cannot be ensured, and the solution $\theta*$ of the dual problem is just a lower bound that may be more or less tight [54] (*dual gap*). However, we have found that, for reasonable values of TH_{gaze}, and due to the partial independency between the paths that compute $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{g}}_{t+\Delta}$ (e.g. layers starting from the latent representation $\mathbf{z}_t$ on), our optimization process usually finds feasible points obeying the constraints, thus ensuring that the obtained bounds are tight.

Finally, to optimize our model with respect to this novel loss, we perform the following process. We initialize $\lambda = 0$ and iterate as follows:

1) Given the current model parameters θ, we use PGA and eq. (13) to update λ towards the direction that maximizes the dual.
2) For the new obtained λ value, we perform Stochastic Gradient Descend (SGD) to find the optimal model parameters θ that minimize the dual function for the given value of λ (see eq. (11)).

It is worth giving some notes on the implementation process. For each new value of λ computed in step 1, we should perform several iterations of SGD in order to find the optimal parameters of the model for this λ (the dual value in eq. (11)). Then, we can update λ and proceed again, until convergence (e.g. $\lambda = 0$ if there exists a feasible solution). In practice, without observing a noticeable lack of performance and due to the time required to optimize over model parameters, we limit this loop to 3 repetitions, when we finish even if the constraints are not yet satisfied, and limit the number of θ updates in step 2 to 1. Overall, this leads to a 1-to-1 gradient ascent-descent between dual (λ) and primal (θ) parameters.

V. Experimental Section

A. Datasets and Experimental Setup

We have considered three egocentric datasets in our experiments: GITW and SHARON, which focus on proximal and motor intention forecasting (milliseconds to seconds), and the complementary Invisible dataset, which considers the prediction of motor intention not only before but also during the action execution.

Grasping In The Wild (GITW)[4] has been recorded with Tobii Glasses 2 worn by subjects performing activities of everyday life in an ecological environment (7 kitchens). It contains 404 egocentric videos of lengths varying between 3.5 and 26 seconds, with a total length of 62 minutes, and 16 categories of objects being grasped: bowl, can of coca-cola, frying pan, glass, jam container, pan lid, milk container, mug, oil bottle, plate, rice container, sauce pan, sponge/scourer, sugar container, vinegar bottle, and washing up liquid. Videos have been recorded with a resolution of 1920×1080px @25Hz, whereas gaze points were acquired at 50 Hz. The same dataset is used to train the AOD and GIP modules.

Symmetric HumAn RobOt iNteraction (SHARON)[5] contains two datasets: SHARON-OBJECTS and SHARON-GRASP. The database has been recorded with Pupil Invisible Glasses at 1088×1080px @30Hz, providing gaze points at a rate of 200Hz. The dataset includes 21 utensils and ingredients of interest in breakfast recipes: bowl, butter, cereals, coffee, cup, cutting board, fork, fridge, jam, knife, microwave, milk, nesquick, nutella, olive oil, plate, sliced bread, spoon, sugar, toaster, tomato sauce, water. SHARON-OBJECTS contains 128 videos recorded by 6 subjects in which the object of interest is placed alone over a smooth surface and the user looks and manipulates the object in isolation. In all cases, volunteers signed the corresponding informed consent. We use this dataset to learn the AOD module following the efficient gaze-driven approach in [2], which minimizes the effort of human annotation. SHARON-GRASP contains 236 videos with lengths varying from 1.5 and 11 seconds, with a total length of 32.6 minutes, recorded by 6 subjects, and is used to learn the GIP module.

The recording protocol is similar in both datasets: each subject first listened to the instruction with the name of the object-to-grasp. Next, he/she explored the visual scene to find the location of target object and finally grasped it. The main difference between them, besides the use of different eye-tracker glasses to acquire the data (different data rates, images dimensions) and the sets of objects of interest, lies in the annotation process. In GITW the videos have been annotated by labeling the temporal segment (t_{start}, t_{end}) starting when the user fixates the active object for the first time (after searching the scene) and ending at the instant just before the object is grasped. In SHARON-GRASP, labeling is efficiently implemented through voice commands and Automatic Speech Recognition (ASR) systems: first, the subject receives a voice command to indicate the object to be grasped, second, the subject indicates that the object is grasped at the exact moment of touching the object. We set the segment of interest for grasping action prediction as ranging from the last sample of the first voice command and the starting sample of the second one. It is noteworthy that this annotation yields to a

[4]www.labri.fr/projet/AIV/dossierSiteRoBioVis/GraspingInTheWildV2.htm
[5]To be published upon paper acceptance

TABLE I
ASSESSMENT OF OUR CONSTRAINED LOSS FOR ASYMMETRIC MT ON GITW DATASET.

Model	AP@Δt(secs)											Gaze Error(%) (mean $\pm$ std)
	$\Delta t = 0.2$	$\Delta t = 0.4$	$\Delta t = 0.6$	$\Delta t = 0.8$	$\Delta t = 1.0$	$\Delta t = 1.2$	$\Delta t = 1.4$	$\Delta t = 1.6$	$\Delta t = 1.8$	$\Delta t = 2.0$	AVG	
Single-task (no VWM)	0.176	0.322	0.428	0.494	0.550	0.580	0.600	0.608	0.622	0.624	0.536	–
Single-task (with VWM)	**0.186**	**0.330**	0.440	0.516	0.578	0.608	0.630	0.646	0.656	0.660	0.564	–
MT Fixed ($\alpha = 0.01$)	0.124	0.230	0.352	0.472	0.568	0.618	0.646	0.660	0.670	0.676	0.544	7.60 ± 0.71
MT Fixed ($\alpha = 0.1$)	0.124	0.232	0.354	0.472	0.574	**0.620**	**0.648**	**0.664**	**0.672**	**0.676**	0.546	4.96 ± 0.61
MT Fixed ($\alpha = 1.0$)	0.174	0.294	0.404	0.496	0.580	0.608	0.634	0.646	0.656	0.662	0.553	4.54 ± 0.67
LB-MT [41]	0.160	0.288	0.400	0.496	0.574	0.610	0.636	0.650	0.658	0.66	0.552	4.52 ± 0.71
DTP-MT [36]	0.174	0.294	0.406	0.496	0.576	0.608	0:632	0.648	0.656	0.660	0.553	5.06 ± 0.68
MT Constrained	**0.186**	**0.330**	**0.444**	**0.520**	**0.584**	0.612	0.638	0.658	0.668	0.670	**0.569**	6.16 ± 0.43

TABLE II
ASSESSMENT OF OUR CONSTRAINED LOSS FOR ASYMMETRIC MT ON SHARON DATASET.

Model	AP@Δt(secs)										Gaze Error (%) (mean $\pm$ std)
	$\Delta t = 1.0$	$\Delta t = 1.5$	$\Delta t = 2.0$	$\Delta t = 2.5$	$\Delta t = 3.0$	$\Delta t = 3.5$	$\Delta t = 4.0$	$\Delta t = 4.5$	$\Delta t = 5.0$	AVG	
Single-task (no VWM)	0.132	**0.296**	0.442	0.536	0.586	0.606	0.620	0.628	0.628	0.498	–
Single-task (with VWM)	0.140	**0.296**	0.442	0.538	0.612	0.630	0.642	0.652	0.654	0.512	–
MT Fixed ($\alpha = 0.01$)	0.134	0.292	0.416	0.524	0.592	0.614	0.622	0.634	0.634	0.496	3.00 ± 0.23
MT Fixed ($\alpha = 0.1$)	0.134	0.292	0.442	0.544	0.622	0.640	0.650	0.658	0.662	0.516	1.28 ± 0.08
MT Fixed ($\alpha = 1.0$)	**0.142**	**0.296**	0.448	0.548	0.620	0.640	0.650	0.662	0.664	0.519	0.86 ± 0.11
LB-MT [41]	0.128	0.286	0.340	0.528	0.600	0.626	0.632	0.644	0.648	0.503	0.92 ± 0.11
DTP-MT [36]	0.136	**0.296**	0.446	0.542	0.620	0.642	0.650	0.660	0.664	0.517	1.16 ± 0.11
MT Constrained	0.140	**0.296**	**0.450**	**0.556**	**0.626**	**0.648**	**0.658**	**0.668**	**0.672**	**0.524**	1.44 ± 0.11

more challenging problem, as we are not considering the time required by the subject to, once the command is listened and processed, scan the scene and search the object to be grasped.

We have measured the performance to anticipate grasping actions using the Average Precision (AP), which is computed by accumulating detections (true and false) along the videos in the test set for different values of the detection threshold. Furthermore, to assess the anticipation time, AP has been computed at different times $\Delta t = t - t_{start}$. The considered range of Δt varies between datasets (see Tables I and II), due to the different annotation protocols discussed in the previous paragraph, which allow lower values of Δt in GITW. We compute AP in our multiclass problem with $C_o + 1$ classes considering a true positive only when the system detects that the subject aims to grasp the right object during the period $t \in (t_{start}, t_{end})$, and a false positive if the detection is either associated to a wrong object or is done before the valid period. To establish a fair comparison, every compared method has been trained using the same feature set described in sec. IV-C and using the $\mathcal{L}_{act}$ proposed in [2]. Furthermore, in order to obtain statistically stable results, we have followed a 5-fold cross validation in both datasets, leading to final mAP values. We will use these two datasets to assess our technical contributions and to establish a comparison with state-of-the-art methods of action forecasting.

Invisible Dataset [17] is a very recent dataset for gaze-driven natural object grasping detection. This dataset complements GITW and SHARON with one appealing property: apart from regular grasping sequences, it also contains videos in which users look at specific objects without the intention of grasping them. Hence, the original set of 10 object categories leads to a 21-multiclass action prediction problem: 1 class for no object of interest (the user simply scans the scene), 10 categories for *viewing* objects, and 10 categories for *grasping objects*. Discriminating between *viewing* and *grasping* and object is rather challenging in advance, when a subject has not yet used their hand. On the other hand, and due a different scenario of application [17], intention is not only predicted in advance in this dataset but also detected during the action execution. This means that the whole videos are labeled with the same category, including frames in which the subject is actually grasping or even holding the object. Although this set-up is not realistic in scenarios where action forecasting is required, as control of neuroprostheses or predictive robotics in HRI, we have followed the original set-up of the user-based experiments in the paper (see [17]) to obtain comparable results with other methods, and used the same performance metrics: frame-based average action prediction Accuracy, and Success Rate. We used this dataset to establish a comparison with particular methods for gaze-based grasping intention prediction.

B. Assessment of model contributions: VWM and constrained loss for asymmetric multi-task learning

In this section, we provide a comprehensive analysis of the effect of the main technical contributions of our proposal. Results are provided for GITW and SHARON datasets in Tables I and II, respectively. We have conducted studies comparing our full approach (AMT-GAF) with three ablated versions by incrementally adding contributions: (1) a single-task approach (considering only $\mathcal{L}_{act}$ and removing the processing blocks associated to the gaze prediction) that does not include the VWM in the input feature set; and (2) a single-task approach but including VWM as features, and (3) the classical linear multi-task (MT fixed), defined in eq. (8) and trained using fixed values of the mixing hyperparameter α. Furthermore, we have extended the comparison with two other

TABLE III
A COMPARISON WITH STATE-OF-THE-ART METHODS FOR ACTION FORECASTING ON GITW DATASET.

Model	AP@Δt(secs)											Gaze Error(%) (mean $\pm$ std)
	$\Delta t = 0.2$	$\Delta t = 0.4$	$\Delta t = 0.6$	$\Delta t = 0.8$	$\Delta t = 1.0$	$\Delta t = 1.2$	$\Delta t = 1.4$	$\Delta t = 1.6$	$\Delta t = 1.8$	$\Delta t = 2.0$	AVG	
Gonzalez et al. [2]	0.170	0.308	0.426	0.498	0.556	0.592	0.604	0.614	0.626	0.628	0.539	–
RU-LSTM [55]	0.162	0.296	0.422	0.508	0.582	**0.612**	0.632	0.644	0.652	0.654	0.556	–
SRL [33]	0.158	0.274	0.388	0.464	0.528	0.568	0.588	0.606	0.610	0.614	0.516	–
AMT-GAF (Ours)	**0.186**	**0.330**	**0.444**	**0.520**	**0.584**	**0.612**	**0.638**	**0.658**	**0.668**	**0.670**	**0.569**	6.16 ± 0.43

TABLE IV
A COMPARISON WITH STATE-OF-THE-ART METHODS FOR ACTION FORECASTING ON SHARON DATASET.

Model	AP@Δt(secs)										Gaze Error (%) (mean $\pm$ std)
	$\Delta t = 1.0$	$\Delta t = 1.5$	$\Delta t = 2.0$	$\Delta t = 2.5$	$\Delta t = 3.0$	$\Delta t = 3.5$	$\Delta t = 4.0$	$\Delta t = 4.5$	$\Delta t = 5.0$	AVG	
Gonzalez et al. [2]	0.152	0.276	0.404	0.482	0.536	0.562	0.572	0.580	0.582	0.461	–
RU-LSTM [55]	**0.156**	**0.318**	0.444	0.536	0.592	0.616	0.628	0.634	0.640	0.507	–
SRL [33]	0.132	0.286	0.422	0.508	0.562	0.588	0.600	0.606	0.612	0.480	–
AMT-GAF (Ours)	0.140	0.296	**0.450**	**0.556**	**0.626**	**0.648**	**0.658**	**0.668**	**0.672**	**0.524**	1.44 ± 0.11

advanced approaches for dynamic multi-task learning: Loss-balanced task-weighting (LBTW-MT) [41], and Dynamic-Task Prioritization (DTP-MT) [36].

Regarding the use of the VWM, we can observe that VWM is useful even as a feature, as it keeps along time scores of objects that have been previously fixated in the short past.

Concerning multi-task learning, in SHARON dataset, the auxiliary task of gaze prediction always helps to regularize the solution for the action prediction task. In GITW, the linear combination of losses with fixed weights, although yields competitive results for some values of Δt, fails to provide stable performance and even achieves worse average AP than the single-task approach. This makes us think that the existence of ill-posed restrictions over the visual attention forecasting degrades the optimization process. This intuition is also supported by the fact that, for both datasets, the best performance is always achieved by the proposed constrained loss, which demonstrates that the use of restrictions successfully prevents over-fitting when none of the candidate maps adjusts well to the real future gaze. Furthermore, dynamic MT methods LBTW-MT an DTP-MT do not yield good results in our scenario. These methods aim to balance the importance of each task during optimization based on its difficulty, so that learning focuses more on those tasks for which the system is not performing well yet. However, although this might be beneficial to some extent, it will lead to severe over-fitting in the cases discussed previously. Hence, we can conclude that our multi-task architecture is beneficial, as it regularizes the solution for action prediction, but requires a dynamic loss that correctly balances the influence of both tasks and handles the existence of ill-posed problems during optimization. In our case, we have provided a dynamic balancing: λ in eq. (10) is computed for each video and learning iteration.

TABLE V
A COMPARISON WITH STATE-OF-THE-ART METHODS FOR NATURAL GRASPING INTENTION ON INVISIBLE DATASET. RESULTS FOR ALL COMPARED METHODS HAVE BEEN TAKEN FROM [17]

Method	Accuracy	Success Rate
Video-Net [56]	60.40 ± 7.28	43.72 ± 7.98
Midas-Net [25]	65.40 ± 6.49	47.94 ± 5.88
TAGMM [24]	78.43 ± 4.19	60.74 ± 4.72
GIRSDF [17]	88.12 ± 3.13	79.48 ± 5.49
AMT-GAF (Ours)	$\mathbf{89.20 \pm 3.33}$	$\mathbf{81.7 \pm 7.59}$

C. A comparison with state-of-the-art architectures of action forecasting

In this section we compare our proposal with other computational architectures in the literature that tackle prediction of proximal intention in the form of human action forecasting in egocentric video. The results of this experiment are given for GITW and SHARON datasets in Tables III and IV, respectively. In particular, we compare our method with:

1) Gonzalez et. al [2]: the method used to tackle grasping action prediction in the paper proposing GITW dataset. The method used an LSTM for sequence modeling, and originally introduced the $\mathcal{L}_{act}$ that we use in the present work. We have simply substituted its simpler feature set by the one presented in this paper.
2) RU-LSTM [55]: Rolling-Unrolling LSTMs have demonstrated a great performance to encode previous sequential information and decode future information for action prediction, and become a reference method for general action forecasting as EGTEA [57] and EPIC-Kitchens [58], [59].
3) SRL [33]: a recent paper that has achieved state-of-the-art results in both EGTEA and EPIC-Kitchens, which leverages the generation of future features on top of which future actions are predicted. Similarly, the method has been adapted and trained in our scenario of application.

RU-LSTM and SRL have been designed for action forecasting in EGTA and EPIC-Kitchens datasets and consider a slightly different action prediction problem which, from our point of view, is not very realistic and does not fit well our scenario of application: they pose action prediction as a multiclass problem over fixed segments of analysis (e.g. 2 seconds before the next action begins), which requires setting the beginning of actions not only in training but also during test; in consequence, they do not consider "no action" as a possible class either. We have therefore adapted both RU-LSTM and SRL to our online scenario by setting a prediction horizon of one frame, adding the additional class 'no-grasp', and taking our proposed loss $\mathcal{L}_{act}$ for training.

The tables show that AMT-GAF clearly outperforms the rest of approaches. RU-LSTM and SRL, although have shown impressive results in EGTA and EPIC-Kitchens, adapt worse

to our particular scenario of online grasping action prediction. The rationale behind is that both approaches have been designed to simultaneously provide a vector of predictions at several anticipation horizons (e.g. at 0.25, 0.5, 0.75 ... until 2 secs of anticipation) while our predictions are intended in the very short-term and in an online manner. Consequently the impact of both the unrolling LSTM in RU-LSTM and the future feature prediction module in SRL is strongly limited. In our approach, the attention-based encoder model encodes the internal state of subject visual dynamics and predictions are inherently instantaneous, with a horizon of one frame. Furthermore, it seems that attention establishes better long-term relationships between the current state and past observations, and our multi-task approach leverages the analysis and prediction of the gaze to regularize the learning process of the grasping action prediction.

Let us note that long-term relationships captured by our transformer encoder, although generally beneficial and leading to the optimal resultsin almost cases, are however outperformed by RNN-based solutions in SHARON dataset when $\Delta t < 2secs$ (see LSTM in [2] and RU-LSTM [55]). We have further analyzed this issue and found that RNNs yield similar recall but higher detection precision than our solution in these two particular cases, which means that our approach leads to more false grasping action detections. Our hypothesis is that during those early instants, given the high variability in the training videos (e.g. in some videos subjects may be still searching for the object, in others they have already located it), our transformer encoder is learning long-term relationships that do not generalize well on the test videos.

D. A comparison with state-of-the-art methods predicting natural grasping intention

In this section we compare our method with several solutions specifically-tailored to the prediction of natural grasping intention using the Invisible dataset: Video-Net [56], Midas-Net [25], TAGMM [24] and GIRSDF [17].

In this case, and following the approach of Gaze-Yolo in [17], we trained our object detector to concurrently identify the active object and also its associated action (viewing or grasping). Results for the challenging subject-based experiments are provided in Table V, in which a cross-validation scheme was applied leaving for testing all videos from one subject at a time. Results show that our method outperforms the rest of the considered approaches, even when all of them have been carefully designed for the particular task of grasping intention prediction. Furthermore, AMT-GAF outperforms the remaining methods by a large margin, in except for the GIRDSF. GIRDSF was proposed together with the Invisible dataset and therefore is especially fine-tuned to perform well on it. In any case, our method achieves better performance and shows one relevant advantage compared to GIRDSF: it does not require to annotate objects using bounding boxes, as our AOD module is trained using weak labels indicating the object of interest in each video and the action being carried. This strongly improves the usability of our approach and its deployment in scenarios where novel unseen objects

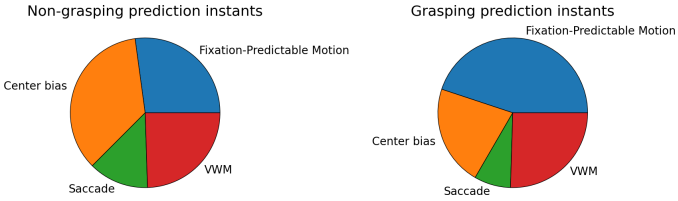

Fig. 5. Distribution of motion patterns conditioned on the action predicted by our model (no grasping, grasping an object) in GITW dataset. Note that fixation and predictable motion have been fused to improve visualization.

are dynamically introduced: our weak-learning approach and voice command-based annotations (see SHARON description) would only require users to look at the new object for some time and from different angles and viewpoints to automatically learn visual models of the object and incorporate it to the AOD module.

E. Interpretability

Including the auxiliary task of future gaze prediction not only impacts the performance on the primary task, but also enhances the interpretability of our approach, as the gaze prediction module has been designed in accordance with cognitive theories of eye movements [18], [46], [47]. To demonstrate our claims, we have conducted a post-hoc study that combines temporal and statistical analyses of eye motion patterns conditioned on model predictions, with the application of an inherently interpretable global surrogate model [60] for action prediction. We did all the experiments in GITW as the action annotations are more accurate in this dataset (e.g. annotations identify the first moment when the subject fixates the object to be grasped).

In order to assess the relevance of each state s of visual attention for each time instant t, we have computed the posterior probability of its associated component in our mixture model, evaluated at the predicted future gaze location $\tilde{\mathbf{g}}_{t+\Delta}$ (see sec. IV-C) as:

$$p(S_t = s|\tilde{\mathbf{g}}_{t+\Delta}) = \frac{v_{ts}\mathbf{m}_{ts}(\tilde{\mathbf{g}}_{t+\Delta})}{\sum_{s'=1}^{S} v_{ts'}\mathbf{m}_{ts'}(\tilde{\mathbf{g}}_{t+\Delta})} \quad (14)$$

With this posterior probability we can effectively measure the influence of each attention state on the final prediction of the future gaze location.

Fig. 5 shows the values of the posteriors conditioned on the model grasping actions predictions $\tilde{y}_t$: (left) $p(S_t = s|\tilde{\mathbf{g}}_{t+\Delta}, \tilde{y}_t = 0)$ if the decision is non-grasping; and (right) $p(S_t = s|\tilde{\mathbf{g}}_{t+\Delta}, \tilde{y}_t > 0)$ if the decision is grasping an object. Let us note that fixation and predictable motion have been aggregated to improve visualization as they often drive visual attention to the same area in the scene (e.g. in the absence of large displacements). Similarly, all maps encoding the location of objects in the VWM have also been aggregated into a unified map.

We can see that various patterns change between non-grasping and grasping instants: central bias and, to a lesser extent, saccadic movements, are more dominant during non-grasping instants, whereas fixation-predictable motion behave oppositely. We have observed that gaze is much more unstable

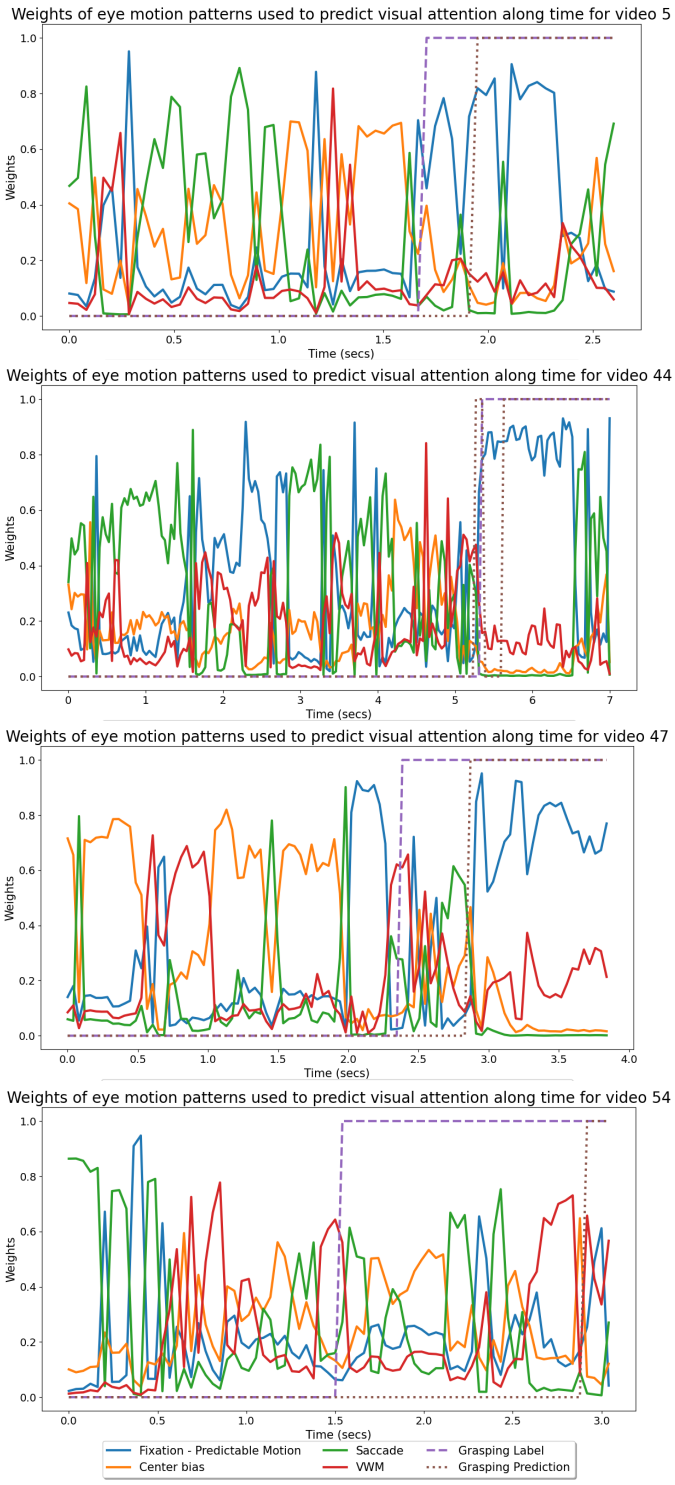

Fig. 6. Plots of intepretable weights for different eye motion patterns along time in various video sequences for GITW dataset, together with two binary signals: the ground-truth action grasping label (indicating the segment between the moment when the subject has fixated the object of interest and the moment just before he/she grasps the object), and model's action prediction (0 no intention to grasp, 1 intention to grasp an object).

during the phase previous to the grasping preparation, as the subject is scanning the scene or even approaching the object. In those cases, different motion patterns follow one after another to drive visual attention and, quite often, attention

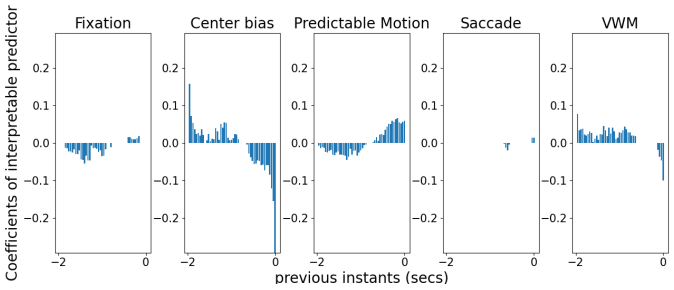

Fig. 7. Coefficients of an interpretable logistic regressor fed by a buffer of the previous 2 secs of relevances for our considered attention states.

is directed towards the center of the scene, which explains the high relevance of center-bias. During the prediction of grasping actions, in contrast, fixation-predictable motion are the dominant states, as subjects tend to keep their attention over the object of interest. Finally, we observe that VWM is a relevant state and attracts attention during both instant types.

Previous intuitions can be confirmed looking at Figure 6, which shows the temporal variation of the posterior $p(S_t = s|\tilde{\mathbf{g}}_{t+\Delta})$. The first three examples clearly show that, during the instants previous to the first fixation of the object of interest (e.g. before the label changes from 0 to 1), attention is modeled as a sequence of different eye motion patterns, including some stable fixations, some unpredictable shifts and a notable influence of the central bias phenomenon. In addition, VWM is recurrently used to redirect gaze to previously fixated elements. We can also see, as we presumed from cognitive theories [46], that states in gaze dynamics are generally short (hundreds of milliseconds) and quickly evolve into new ones (in except of central bias, which tends to be dominant during longer periods when the other states fail to accurately represent visual attention).

Once the active object is fixated, gaze looks more stable and fixation-predictable motions become the clear dominant patterns. Although this is the usual behavior, we can also find cases where gaze exhibits a different behavior. As we show in the fourth example, during the instants just after the label changes to 1, attention is still dominated by central bias and saccadic movements. The consequence is that our system is not able to infer the right action until then, when gaze becomes more stable over the object to be grasped (fixation and VWM patterns). After analyzing videos, we found that this is sometimes caused by subjects that, after identifying the location of the object of interest, approach the object using their peripheral vision. In other cases, it may be due to an inaccurate labeling process, e.g. annotators identify the first active object fixation when gaze passes over the object during a saccadic movement.

To get deeper insights about the importance of visual attention and gaze motion patterns in the intention-action processes, we have additionally developed a global surrogate model [60] that tackles action prediction using a linear classifier fed by the attention state relevances. In particular, we have chosen a logistic regressor with L1-regularization to predict a binary signal (0 no grasp, 1 grasp an object), working over the sequence of posteriors $p(S_t = s|\tilde{\mathbf{g}}_{t+\Delta})$ corresponding to the last second (e.g. 25 frames at 25 fps in GITW dataset).

Considering 5 states (fixation, central bias, predictable motion, saccade and VWM), this leads to a total dimension of D=125. We have further normalized the inputs by standardization. Our choice of a linear classifier and L1-regularization lies in its simplicity and its inherent interpretability through the study of the coefficients, which are different than zero only for relevant features. Furthermore, to emulate our full model in online action forecasting (e.g. generate the variable $\tilde{\mathbf{y}}_t$), we combine the binary decision provided by the surrogate model (grasp, no grasp) with the output of the AOD, to allow identifying the object to be grasped.

Our surrogate model addresses the binary prediction problem (grasping, non-grasping) with an $AP = 0.689$. Combined with the outputs of the AOD in the multi-class prediction problem yields a reasonable average $AP = 0.335$. Although this value is notably below the performance of the full model (average $AP = 0.569$, as stated in Table III), it demonstrates that our modeling of visual attention through simple and interpretable motion patterns encodes very valuable information about human intention. Of course, the difference in the performance between full and surrogate models lies in the fact that the first relies on the richer and more expressive latent variable $\mathbf{z}_t$, which is not restricted to the attention states but encodes instead the full dynamics of the visual field.

Finally, we have further analyzed the values of the coefficients in the classifier, as they provide meaningful cues about the relationships between eye motion patterns and human grasping intention. Results are shown in Fig. 7. Let us note that, due to L1-regularization, less relevant features have associated null coefficients. From the figure, we can see that finding the central bias pattern in the short-term past (up to 0.5 secs) is very indicative of non-grasping instants, whereas recent occurrences of predictable motion and fixations suggest grasping actions. If instead we observe the more distant past (from 1 sec. on), the behavior of these patterns is the opposite, indicating the complementary actions. In addition, we can see that VWM has some medium-term anticipation capability (grasping action is going to happen in 1-2 seconds). Finally, saccadic movements were found to be certainly irrelevant in our surrogate model.

VI. Conclusions and Further Work

In this paper, we have presented AMT-GAF, a multi-task model that simultaneously predicts future visual attention and forecast grasping actions by decoding human intention. Based in the hypothesis that perception, in the form of visual exploration, and intention are strongly coupled, we have designed an asymmetric multi-task learning approach in which action forecasting is the main and future visual attention is the auxiliary task, respectively. In addition, we have proposed a novel constrained optimization problem that minimizes the action forecasting loss subject to achieving good enough results in future visual attention prediction.

Our experiments have demonstrated that our multi-task model successfully regularizes learning and improves the performance of action forecasting. Therefore, the use of our asymmetric multi-task loss is fundamental to exploit the synergies between both tasks and avoid learning degradation. Furthermore, AMT-GAF also outperforms two architectures that represent the state-of-the-art in general action forecasting, and a large set of methods that have been specifically designed for grasping intention detection.

Our module for future visual attention is inherently interpretable, as it relies on a set of states associated to well-known eye motion patterns from visual psychology. These states model how humans direct their gaze towards the different elements in a scene during a task. Our experiments have demonstrated the strong links that exist between these attentional states and human intention, and leveraged this relationship to provide meaningful explanations to system decisions.

Further work will follow three lines of research. First, we are currently making experiments with real-time HRI between humans and assistive robotics to demonstrate that predictive robotics (enabled by AMT-GAF) can successfully anticipate human needs and act proactively, leading to more natural interactions and saving time with respect to reactive robots responding to explicit human requests (e.g. voice commands). We aim to additionally couple AMT-GAF with existing system for shared control of neuroprosthesis [4], integrating data from natural arm movement and gaze information [61]. Second, we will extend the anticipation horizon from the current short-term (milliseconds) to medium and long-terms (seconds, minutes) and incorporate action prediction to more complex activities as cooking or working. With these horizons, gaze losses importance as guiding signal, and should be replaced by a compositional and sequential analysis of the duple tasks-atomic actions (e.g. a cooking recipe can be defined as a sequence of atomic actions involving ingredients and utensils). Third, we aim to dig into the model usability and adaptation to dynamic and changing environments. Currently, our weakly-supervised AOD can be easily adapted to new acquired utensils by simply recording videos of the objects. However, we consider action detection as a multi-class problem, which requires that the action categories are defined a priori. This limits the incorporation of new utensils in case they lead to new unseen object categories and, in consequence, to new actions (e.g. grasping a new unseen object or putting the object into another). Hence, our goal is to develop zero or few-shot learning approaches that can adapt better to dynamic environments and learn from little demonstration.

Acknowledgments

This work has been partially supported by the National Grants PID2020-118504GB-I00 from the Spanish Ministry of Science and Innovation, by the Madrid Government under the grants Y2020/NMT-6660 COMPANION-CM and the Multiannual Agreement with UC3M in the line of "Fostering Young Doctors Research" (SHARON-CM-UC3M), in the context of the V PRICIT (Regional Programme of Research and Technological Innovation).

References

[1] S. Abal-Fernández, C. Caramazana-Zarzosa, M. B. Loureiro-Casalderrey, S. Martínez, C. Balaguer, F. Díaz-de María, and

I. González-Díaz, "Learning rl policies for anticipative assistive robots by simulating human-robot interactions in real scenarios using egocentric videos," in *2023 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2023, pp. 1–8.

[2] I. González-Díaz, J. Benois-Pineau, J.-P. Domenger, D. Cattaert, and A. de Rugy, "Perceptually-guided deep neural networks for ego-action prediction: Object grasping," *Pattern Recognition*, vol. 88, pp. 223 – 235, 2019.

[3] S. Mick, S. Effie, L. Dure, H. Christophe, J. Benois-Pineau, G. Loeb, D. Cattaert, and A. Rugy, "Shoulder kinematics plus contextual target information enable control of multiple distal joints of a simulated prosthetic arm and hand," *Journal of NeuroEngineering and Rehabilitation*, vol. 18, 01 2021.

[4] E. Segas, S. Mick, V. Leconte, O. Dubois, R. Klotz, D. Cattaert, and A. de Rugy, "Intuitive movement-based prosthesis control enables arm amputees to reach naturally in virtual reality," *eLife*, vol. 12, p. RP87317, oct 2023.

[5] A. R. Mele, *Springs Of Action: Understanding Intentional Behavior*. Oxford University Press, 02 1992.

[6] E. Pacherie, "The phenomenology of action: A conceptual framework," *Cognition*, vol. 107, no. 1, pp. 179–217, 2008.

[7] Y. Liu, J. Yuan, and Z. Tu, "Motion-driven visual tempo learning for video-based action recognition," *IEEE Transactions on Image Processing*, vol. 31, pp. 4104–4116, 2022.

[8] J. J. Gibson, *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979, ISBN: 978-1-848-72578-2.

[9] M. Bratman, *Intention, plans, and practical reason*. Cambridge, MA: Harvard University Press, 1987, ISBN: 978-0-674-45818-5.

[10] B. F. Malle and J. Knobe, "The folk concept of intentionality," *Journal of Experimental Social Psychology*, vol. 33, no. 2, pp. 101 – 121, 1997.

[11] D. BALLARD, M. HAYHOE, and J. PELZ, "Memory representations in natural tasks," *JOURNAL OF COGNITIVE NEUROSCIENCE*, vol. 7, no. 1, pp. 66–80, WIN 1995.

[12] A. Belardinelli, "Gaze-based intention estimation: principles, methodologies, and applications in hri," 2023.

[13] R. Johansson, G. Westling, A. Bäckström, and J. Flanagan, "Eye–hand coordination in object manipulation," *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 21, pp. 6917–32, 10 2001.

[14] I. González-Díaz, J. Benois-Pineau, J. Domenger, and A. de Rugy, "Perceptually-guided understanding of egocentric video content: Recognition of objects to grasp," in *ACM International Conference on Multimedia Retrieval, ICMR*, 2018, pp. 434–441.

[15] S. Li, M. Bowman, H. Nobarani, and X. Zhang, "Inference of manipulation intent in teleoperation for robotic assistance," *Journal of Intelligent & Robotic Systems*, vol. 99, 07 2020.

[16] A. Belardinelli, A. R. Kondapally, D. Ruiken, D. Tanneberg, and T. Watabe, "Intention estimation from gaze and motion features for human-robot shared-control object manipulation," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 9806–9813.

[17] B. Yang, X. Chen, X. Xiao, P. Yan, Y. Hasegawa, and J. Huang, "Gaze and environmental context-guided deep neural network and sequential decision fusion for grasp intention recognition," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 3687–3698, 2023.

[18] M. Land and B. Tatler, *Looking and Acting: Vision and eye movements in natural behaviour*. Oxford University Press, 07 2009.

[19] B. David-John, C. Peacock, T. Zhang, T. S. Murdison, H. Benko, and T. R. Jonker, "Towards gaze-based prediction of the intent to interact in virtual reality," in *ACM Symposium on Eye Tracking Research and Applications*, ser. ETRA '21 Short Papers. New York, NY, USA: Association for Computing Machinery, 2021.

[20] R. Singh, T. Miller, J. Newn, L. Sonenberg, E. Velloso, and F. Vetere, "Combining planning with gaze for online human intention recognition," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, ser. AAMAS '18. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2018, p. 488–496.

[21] S. Fuchs and A. Belardinelli, "Gaze-based intention estimation for shared autonomy in pick-and-place tasks," *Frontiers in Neurorobotics*, vol. 15, 2021.

[22] C.-M. Huang and B. Mutlu, "Anticipatory robot control for efficient human-robot collaboration," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2016, pp. 83–90.

[23] A. Haji Fathaliyan, X. Wang, and V. J. Santos, "Exploiting three-dimensional gaze tracking for action recognition during bimanual manipulation to enhance human–robot collaboration," *Frontiers in Robotics and AI*, vol. 5, 2018.

[24] B. Yang, J. Huang, X. Chen, X. Li, and Y. Hasegawa, "Natural grasp intention recognition based on gaze in human–robot interaction," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 4, pp. 2059–2070, 2023.

[25] P. Festor, A. Shafti, A. Harston, M. Li, P. Orlov, and A. A. Faisal, "Midas: Deep learning human action intention prediction from natural eye movement patterns," *Journal of Vision*, vol. 21, no. 9, 2021.

[26] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, pp. 41–75, 1997.

[27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *International Conference on Neural Information Processing Systems*, ser. NIPS'15. Cambridge, MA, USA: MIT Press, 2015, pp. 91–99.

[28] S. Jadon, "A survey of loss functions for semantic segmentation," in *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, 2020, pp. 1–7.

[29] L. Liebel and M. Körner, "Auxiliary tasks in multi-task learning," *ArXiv*, vol. abs/1805.06334, 2018.

[30] T. Yu, C. Liu, Z. Yan, and X. Shi, "A multi-task framework for action prediction," *Information*, vol. 11, no. 3, 2020.

[31] S. B. Loh, D. Roy, and B. Fernando, "Long-term action forecasting using multi-headed attention-based variational recurrent neural networks," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022, pp. 2418–2426.

[32] D. Gong, J. Lee, M. Kim, S. Jong Ha, and M. Cho, "Future transformer for long-term action anticipation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[33] Z. Qi, S. Wang, C. Su, L. Su, Q. Huang, and Q. Tian, "Self-regulated learning for egocentric video activity anticipation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 6715–6730, 2023.

[34] X. Liu and J. Yin, "Multi-head trajectorycnn: A new multi-task framework for action prediction," *Applied Sciences*, vol. 12, no. 11, 2022.

[35] J. Ma, Z. Zhao, X. Yi, J. Chen, L. Hong, and E. H. Chi, "Modeling task relationships in multi-task learning with multi-gate mixture-of-experts," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 1930–1939. [Online]. Available: https://doi.org/10.1145/3219819.3220007

[36] M. Guo, A. Haque, D.-A. Huang, S. Yeung, and L. Fei-Fei, "Dynamic task prioritization for multitask learning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[37] Z. Chen, Y. Shen, M. Ding, Z. Chen, H. Zhao, E. Learned-Miller, and C. Gan, "Mod-squad: Designing mixtures of experts as modular multi-task learners," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[38] Y. Xu, Y. Yang, and L. Zhang, "Demt: Deformable mixer transformer for multi-task learning of dense prediction," in *Proceedings of the The Thirty-Seventh Conference on Artificial Intelligence (AAAI)*, 2023.

[39] C. Li, F. Wei, J. Yan, W. Dong, Q. Liu, and H. Zha, "Self-paced multi-task learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, 04 2016.

[40] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[41] S. Liu, Y. Liang, and A. Gitter, "Loss-balanced task weighting to reduce negative transfer in multi-task learning," in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019.

[42] S. Mick, M. Lapeyre, P. Rouanet, C. Halgand, J. Benois-Pineau, F. Paclet, D. Cattaert, P.-Y. Oudeyer, and A. de Rugy, "Reachy, a 3d-printed human-like robotic arm as a testbed for human-robot control strategies," *Frontiers in Neurorobotics*, vol. 13, 2019.

[43] Z. Tu, J. Zhang, H. Li, Y. Chen, and J. Yuan, "Joint-bone fusion graph convolutional network for semi-supervised skeleton action recognition," *IEEE Transactions on Multimedia*, vol. 25, pp. 1819–1831, 2023.

[44] E. Comission, "White paper on artificial intelligence -a european approach to excellence and trust," Brussels, 19.2.2020.

[45] K. Friston, R. Adams, L. Perrinet, and M. Breakspear, "Perceptions as hypotheses: Saccades as experiments," *Frontiers in Psychology*, vol. 3, p. 151, 2012.
[46] R. Hessels, D. Niehorster, M. Nyström, R. Andersson, and I. Hooge, "Is the eye-movement field confused about fixations and saccades? a survey among 124 researchers," *Royal Society Open Science*, vol. 5, p. 180502, 08 2018.
[47] S. Martinez-Conde, S. Macknik, and D. Hubel, "The role of fixational eye movements in visual perception," *Nature reviews. Neuroscience*, vol. 5, pp. 229–40, 04 2004.
[48] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *Trans. Img. Proc.*, vol. 13, no. 10, p. 1304–1318, oct 2004.
[49] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR07). IEEE Computer Society*, 2007, pp. 1–8.
[50] W. Schneider, "Selective visual processing across competition episodes: A theory of task-driven visual attention and working memory," *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 368, p. 20130060, 10 2013.
[51] E. Jang, S. Gu, and B. Poole, "Categorical reparametrization with gumbel-softmax," in *Proceedings International Conference on Learning Representations 2017*. OpenReviews.net, Apr. 2017.
[52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.
[53] H. Straka and N. Dieringer, "Basic organization principles of the vor: lessons from frogs," *Progress in Neurobiology*, vol. 73, no. 4, pp. 259–309, 2004.
[54] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.
[55] A. Furnari and G. M. Farinella, "Rolling-unrolling lstms for action anticipation from first-person video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, pp. 4021–4036, 2020.
[56] D. Kim, B. Kang, K. B. Kim, H. Choi, J. Ha, K.-J. Cho, and S. Jo, "Eyes are faster than hands: A soft wearable robot learns user intention from the egocentric view," *Science Robotics*, p. 2949, 01 2019.
[57] Y. Li, M. Liu, and J. M. Rehg, "In the eye of beholder: Joint learning of gaze and actions in first person video," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.
[58] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Scaling egocentric vision: The epic-kitchens dataset," in *European Conference on Computer Vision (ECCV)*, 2018.
[59] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, J. Ma, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100," *International Journal of Computer Vision (IJCV)*, vol. 130, p. 33–55, 2022.
[60] C. Molnar, *Interpretable Machine Learning*, 2nd ed., 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book
[61] B. Lento, E. Segas, V. Leconte, E. Doat, F. Danion, R. Péteri, J. Benois-Pineau, and A. de Rugy, "3d-arm-gaze: a public dataset of 3d arm reaching movements with gaze information in virtual reality," *bioRxiv*, 2024.

Iván González-Díaz Iván González-Díaz received the Telecommunications Engineering degree from Universidad de Valladolid, Valladolid, Spain, in 1999, the M.Sc. and Ph.D. degree from Universidad Carlos III de Madrid, Madrid, Spain, in 2007 and 2011, respectively. After holding a postdoc position in the Laboratoire Bordelais de Recherche en Informatique at the University Bordeaux, he currently works as Associate Professor at the Signal Theory and Communications Department in Universidad Carlos III de Madrid. His primary research interests include object recognition, category-based image segmentation, scene understanding and content-based image and video retrieval systems. In these fields, he is co-author of several papers in prestigious international journals, two chapters in international books and a few papers in revised international conferences.

Miguel Molina-Moreno Miguel Molina Moreno received his Telecommunications Engineering Degree from Universidad de Granada in 2015 and his M. Sc. in Telecommunications Engineering and Multimedia and Communications from Universidad Carlos III de Madrid in 2017, and his PhD in Multimedia and Communications in Universidad Carlos III de Madrid in 2023. He currently works as a postdoctoral researcher in Yale School of Medicine. His research interests comprehend computer vision applied to medical image, object detection, signal processing, Machine Learning and Deep Learning.

Jenny Benois-Pineau Jenny Benois-Pineau is a professor of Computer Science at the University Bordeaux of exceptional class. Her topics of interest include artificial intelligence in image/multimedia analysis and pattern recognition. She is the author and co-author of more than 230 papers in international journals, conference proceedings, books and book chapters. She has tutored and co-tutored 30 PhD students. She is associated editor of ACM MTAP, senior associated editor JEI SPIE journals. She has served in numerous program committees in international conferences: ACM MM, IEEE ICIP, ICPR, ICMR, CIVR, CBMI, IPTA, MMM, and organized WS at the major conference as ACM MM, IEEE ICIP, ICPR ... She has been coordinator or leading researcher in EU – funded and French national research projects and projects with French companies. She was invited for lecturing as a distinguished researcher at the Universities of Madrid (Spain), Klagenfurt (Austria), Ben Gurion (Israel), NJIT (USA), UAM, CITEDI (Mexico) and gave invited lectures at the UNC at Chapel Hill, Carnegie Melon, Brooklynn Polytechnic (USA), University of Sussex (GB) and UCL (Belgium). She is a member of IEEE SPS TC MMSP 2023-2025. She has Knight of Academic Palms grade.

Aymar de Rugy Aymar de Rugy received a Ph.D. in human movement science (2001) from the University of the Mediterranean, Marseille, France. After 2 years postdoctoral training in the department of kinesiology of the Pennsylvania State University (US), he spent 9 years as a research fellow at the center for sensorimotor neuroscience in the University of Queensland (Brisbane, Australia), which is dedicated to the understanding of human sensorimotor control using a range of non-invasive electrophysiology and brain stimulation techniques. He is currently a research director at the CNRS, in a large neuroscience institute (INCIA, Bordeaux, France). His research focuses on hybrid systems, mixing biological controls with artificial devices, increasing our understanding of the fundamental mechanisms of human sensorimotor control and exploiting this knowledge to restore and optimize movement. Applications of his work include prosthesis control, robotic assistance and teleoperation, and human-machine interfaces.