VLMAH: Visual-Linguistic Modeling of Action History for Effective Action Anticipation

Victoria Manousaki*[1,2], Konstantinos Bacharidis[1,2], Konstantinos Papoutsakis[3], and Antonis Argyros[1,2]

[1]Computer Science Department, University of Crete
[2]Institute of Computer Science, FORTH
[3]Department of Management, Science & Technology, Hellenic Mediterranean University

Abstract

Although existing methods for action anticipation have shown considerably improved performance on the predictability of future events in videos, the way they exploit information related to past actions is constrained by time duration and encoding complexity. This paper addresses the task of action anticipation by taking into consideration the history of all executed actions throughout long, procedural activities. A novel approach noted as Visual-Linguistic Modeling of Action History (VLMAH) is proposed that fuses the immediate past in the form of visual features as well as the distant past based on a cost-effective form of linguistic constructs (semantic labels of the nouns, verbs, or actions). Our approach generates accurate near-future action predictions during procedural activities by leveraging information on the long- and short-term past. Extensive experimental evaluation was conducted on three challenging video datasets containing procedural activities, namely the Meccano, the Assembly-101, and the 50Salads. The results confirm that using long-term action history improves action anticipation and enhances the SOTA Top-1 accuracy.

1. Introduction

Anticipating future actions during an observed complex activity is a critical ability that enables humans to recognize intended goals and outcomes to proactively plan and engage in interactions with other humans and the environment in a timely, efficient, and safe manner. We accomplish this task naturally by perceiving visual information and learning from a few activities as well as based on self-experimentation; thus, it encompasses harnessing relevant kinematic and contextual knowledge rooted in perception, personal experience, and skills. These competencies are regarded as fundamental constituents of human intelligence.

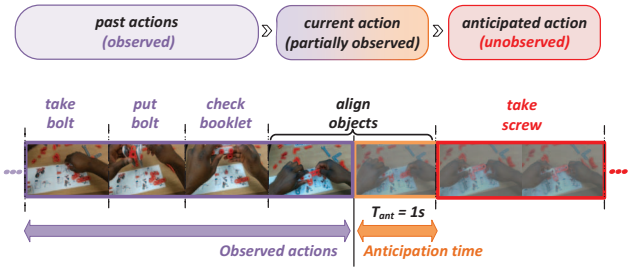

Figure 1. We consider the problem of action anticipation in untrimmed videos of procedural activities. At a certain moment in time (decision point), the proposed framework (VLMAH) anticipates the action (i.e., the unobserved action "take screw") that is most likely to be performed after some anticipation time T_{ant} (depicted with orange color). This is performed on the basis of the history of all past actions up to the decision point (depicted with purple) which is modeled by integrating visual input regarding the immediate past and a linguistic description of the distant past.

Deriving effective solutions for similar competencies is also beneficial to AI-enabled agents and robots that operate in industrial and domestic environments in a multitude of real-world applications [22]. In particular, the anticipation of near or long-term future actions can efficiently be used to advance autonomous navigation or driver-assistance systems, leverage the ability of industrial or home/socially assistive robots towards fluent human-robot collaboration and interaction, drive optimization of industrial workflows and enhance human safety through real-time hazard/anomaly identification to preemptively signal alerts and aids [38].

To enhance AI agents' capabilities, researchers have concentrated on video-based human understanding, yielding impressive outcomes in tasks like recognition, detection, and short- or long-term action prediction during ex-

*vmanous@ics.fort.gr (Corresponding author), kback@ics.forth.gr, kpapoutsakis@hmu.gr or papoutsa@ics.forth.gr, argyros@ics.forth.gr
https://projects.ics.forth.gr/cvrl/vlmah/

tended activities [22]. Among these, action anticipation stands out, involving forecasting upcoming action labels based on partial ongoing action observation and recent action history [42, 7], as depicted in Figure 1. The ability to use recent action history is crucial for proposing potential actions at the decision point T_{ant} before the expected start time of the next action. This anticipation time captures valuable insights and the sequence of actions leading to the anticipated one. We identify the following questions towards this challenging task, which effective solutions have to deal with by assessing the best trade-off between the complexity of spatiotemporal visual feature modeling and the accuracy performance of action anticipation:

- How much of the action history should be considered to accurately predict future actions during complex activities?
- What is the most efficient way to model the temporal ordering of action history (past actions)?
- What information modalities could enhance action anticipation accuracy?

We tackle the challenge of anticipating actions within instructional activities by merging visual and linguistic data from ongoing actions. This encompasses recent and distant history, vital for predicting the future. While visual features offer rich information, they are resource-intensive for storage and computation. In contrast, language-based action descriptions are less detailed but more storage and processing efficient. Our approach balances these aspects by integrating high-cost visual features for recent events and low-cost language features for remote ones.

We explore action anticipation in the context of procedural activities, where variations of the temporal ordering of actions are usually more constrained. Based on that, it is not surprising that the majority of existing works [14, 15, 58, 37, 31, 46, 61, 43, 18] aspire to tackle this problem using video datasets [7, 8, 27, 53, 24, 47, 41, 5] containing procedural activities. For instance, EpicKitchens [8] is one of the largest and most popular video datasets, among others [27, 53, 24, 50, 62], deals with the task of action anticipation featuring videos of cooking activities. Another popular domain of instructional activities that regard complex assembly activities [47, 41, 5, 19, 25, 39] in the context of industrial and non-industrial scenarios.

In particular, we focus on videos of assembly activities using the Meccano [41] and the Assembly-101 [47] video datasets. Those two can are considered complementary with respect to the types of the target activities, as participants in the former are provided with specific instructions to accomplish the assembly process of a toy vehicle, whereas in the latter participants were free to disassemble a fixed toy vehicle and then to assemble it from its parts, following a less constrained process.

Our contributions can be summarized as follows:

- We propose the Visual-Linguistic Modeling of Action History (VLMAH) framework that combines short-term visual and longer-term lexical information of observed past actions to estimate the label of the near-future anticipated action.
- We show that the combination of cost-effective processing and integration of linguistic information along with visual information can greatly benefit prediction accuracy in various types of procedural activities.
- An extensive experimental evaluation was conducted with state-of-art results on three challenging datasets, Meccano [41], Assembly-101 [47] and 50-salads [53], for a large set of different experimental setups, and anticipation times. VLMAH improves the noun/verb/action predictions for the Meccano and Assembly-101 dataset while for the 50Salads dataset, our method is amongst the top performing.

2. Related Work

Action/Activity Recognition sets the thematic base upon which more fine-grained video understanding tasks, such as action detection, early action recognition, and action anticipation/prediction have been defined. In its most challenging form, it comprises the recognition of actions that involve human-object interactions, and action sets with high intra- and inter-class variability. With the advent of deep learning, video action recognition methods have become extremely efficient and effective in modeling short-range dependencies of actions with CNN-centered models [52, 6]. Moreover, the ability to model long-range dependencies of complex actions or long, composite activities has also been considerably improved using memorization layers, such as RNNs and their variants [60, 28], attention mechanisms [57, 2], and temporal frame dependency modeling at multiple time scales [11, 59].

The significant performance gains that have been witnessed in this field have also been fueled by the emergence of large-scale datasets [7, 36], that contain diverse action sets, viewing conditions (egocentric [7, 51, 16] or third-person [1, 26]) and videos in various contexts providing rich, multi-level annotation data and different information modalities. Such datasets enabled the design of multi-modal models that apart from appearance and motion, also exploit audio, gaze-related data, and most importantly language [20, 17]. In the concept of multi-modal action/activity modeling, the visual-linguistic fusion scheme is shown to be extremely effective at representing the variability of complex actions and activities. This mainly relies

on the action-related knowledge that is extracted using the lexical description of the action sequence and transitions, which is presented in the form of a simple text label or rich transcription/captions per action [20]. This information can be further processed using text statistics [45]. Recently, deep learning language models [54, 56], have also been proposed acting as a complimentary information source to the visual representation, expressed with handcrafted [44, 45] or deep learned [32, 23, 3, 4] descriptors.

Action Anticipation/Forecasting is defined as the task of predicting the class(es) of one or more future actions for which no observations are available at the decision time [22, 26]. The tasks of prediction and anticipation have been well-explored for actions of various complexity that range from simple motion primitives of a single human action [34] or a human-object interaction [22, 18, 35] to long, composite, procedural or unconstrained activities [48, 33]. Anticipating the near-future actions is performed towards a limited set or even thousands of action categories [7, 47]. Forecasting of the next actions is performed at "anticipation time" in the video that can be set at variable time horizons ranging from short- to long-term predictions. Many existing approaches fix this important task parameter to 1 second prior to the start of the action of interest [30, 14], while others explore the predictability of actions for several seconds [40, 12, 31, 1, 21]. The problem was initially introduced in third-person videos [18, 1], but it has recently gained significant popularity in first-person (egocentric) videos [7, 16], too.

The prominent method of Furnari *et al*. [13] explored the problem of action anticipation using "rolling-unrolling" LSTMs in order to summarize past actions and make predictions for the verb, noun and action of the next segment for multiple anticipation times. In [49] a multi-scale temporal model is proposed so that the past actions are aggregated for the future actions to be iteratively predicted. This framework performs predictions for the next action with an anticipation time of 1 second and is also capable of performing dense anticipation considering a large number of anticipated action classes. Our work complies with both methodologies so that the verb, noun, and action predictions are made in the range of $[0.25, 2]$ seconds with a step of 0.25 seconds.

Natural language processing (NLP) initially gained popularity in the cooking domain since recipes naturally contain a large variety of texts with instructions on food preparation. These large texts of instructions have attracted the interest for predictions of the next unobserved steps of the recipe in natural language in the form of sentences. Sener *et al*. [50] created a hierarchical model for learning multi-step procedures of recipe datasets with text and visual context. Their zero-shot anticipation framework is able to transfer knowledge from large-scale text corpora to the visual domain for the prediction of coherent and plausible recipe instructions. The same authors improved their framework by integrating a temporal segment proposal method into the video encoder and additional losses at the recipe encoder to improve convergence [48]. By comparing to recipe generation networks they showed that this method can perform better even for unseen recipes and dishes. Contrary to methods [50, 48] that exploit text to provide information to the visual domain, Mahmud *et al*. [33] proposes a two-step approach where information on the visual spatiotemporal context of the observed actions and the linguistic labels of the anticipated actions along with scene context are incorporated for caption prediction. Text and/or captions of the observed actions are not utilized.

Our framework deviates from the aforementioned approaches that use NLP, as we do not focus on the prediction of captions/sentences of the near-future, still unobserved actions. Instead, we focus on using linguistic information complementary to the vision module [3, 4] for the encoding of the short- and long-term history of the observed past.

3. Proposed Approach

The proposed Visual-Linguistic Modeling of Action History framework noted as VLMAH, is shown in Figure 2. It features a two-stream three branch deep neural network design that comprises (a) a vision-based action anticipation sub-network, (b) an activity-level sub-network for temporal modeling based on natural language processing (NLP), and, (c) a vision-based action recognition sub-net. The action anticipation visual sub-net is able to estimate the next action given the visual representation of the current/ongoing action segment exploring the short- and long-term action dynamics. The action recognition sub-net exploits the same input to provide estimates for the current action class. Additionally, the NLP-driven activity-centric sub-net is responsible for the long-range temporal modeling of the relation of the current action to the previously observed actions to learn a stochastic model of the forthcoming action.

The last architecture stage combines the two representations (visual action anticipation sub-net & language modeling sub-net) to anticipate one of the following events, (a) the next action (fine-grained label), (b) the active object of the next segment (noun), or (c) the next motion motif (verb).

3.1. Visual Action Anticipation Module

Given an input sequence x_t of the action y_t of an activity video sample $X_i = \{x_1, ..., x_N\} \rightarrow Y$, the visual action-anticipation sub-net aims at learning the representation of the on-going action at a segment-wise level, that will enable the prediction of the forthcoming action y_{t+1}. To achieve this, the proposed module follows a multi-branch design that operates on an ensemble of different vision-driven representations of the entire scene or of the key to the action

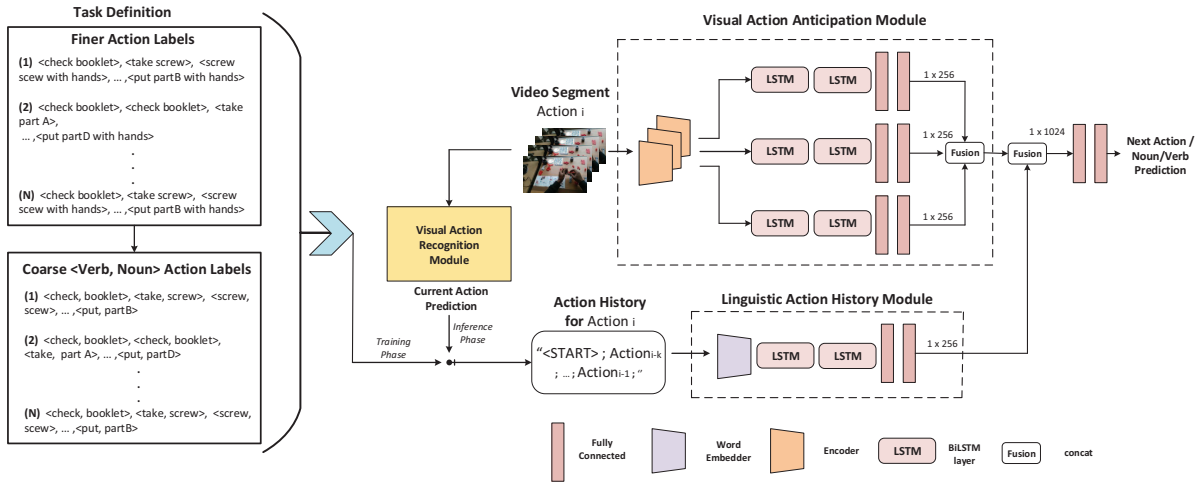

Figure 2. The proposed VLMAH architecture. The Visual Action Anticipation and the Linguistic Action History modules are presented. For the *Meccano* dataset, the encoders of the action module, generate Object, Hands, Gaze representations, whereas, for the *Assembly*-101 dataset, there is a single encoder network, TSM [29] while representations are split into 3 sub-sequences, as mentioned in Section 4.2. The detail level regarding the textual label descriptions is adaptable to the anticipation task at hand (action, motion motif (verb), or object (noun)). The final format also includes two special labels (START, END) that indicate the start and end of the action history sequence.

scene elements, such as the actor's body part regions or the appearance states of the active object.

On a technical basis, each branch of the proposed multi-branch design comprises a two-layer Bidirectional LSTM (BiLSTM) temporal encoder, followed by a Fully-Connected (FC) layer, that further encodes the representation into a $[1 \times 256]$ feature vector. Finally, the representations of all branches are fused via concatenation and forwarded to an FC layer that generates the final representation, which encodes the action segment into a $[1 \times 1024]$ feature vector. To form the inputs of this sub-net, we follow a sparse uniform sampling policy on the input sequence. Regarding the case of visual scene representation in the two datasets of interest, every single action of the action sequence that represents the activity has been encoded using a segment-wise temporal encoder network[1]. Therefore, it corresponds to the feature-based representation of a segment formed based on the adjacent frames. This formulation of the subnet's input enables the modeling of both short- and long-term appearance variations of the scene elements.

3.2. Linguistic Action History Module

We argue that the knowledge of the preceding action occurrences, noted as action history, is important for learning to estimate at a certain time in the video, the label of the next-anticipated action (*action forecasting/anticipation*) or of the active object in that action, as it provides efficient, discriminative features to opt among potential candidate targets. We address this issue using a compact textual description of the preceding actions, in the compact form of action labels, compared to captions that feature extensive textual descriptions of actions. The sentence-based textual description of the preceding actions is processed using the NLP sub-network that comprises a Word Embedding layer followed by the same layer set as the branches of the action-centric visual module. This representation forms a $[1 \times 256]$ feature vector, which is concatenated with the representation of the action-centric module. The combined representation is then forwarded to a set of two FC layers to provide estimations on the next action/object class.

Delving into this representation of the action history, we restructure each label (length, semantic complexity, part-of-speech element position (verb, noun, adverb)), in a specific lexical format depending on the task at hand (action, motion verb, or noun anticipation), to facilitate the learning process. Specifically, in the case of the verb (coarse motion motif) or noun (next-segment active-object) anticipation, we may have to deal with actions of a similar motion and object basis but of a different type of object upon which the action is performed. For example, consider the actions, *screw a screw with hands* and *screw a screw with screwdriver*. When asked to predict the key object(s)[2] of the next anticipated action, the action history module should maintain the key objects of the preceding action segments, and therefore the knowledge that the tool-medium is of no importance in this coarser anticipation problem. A similar convention is also considered for the task of predicting only the coarse motion motif label for the next action.

[1]For example, in Assembly-101 each action instance has been encoded using the effective Temporal Shift Module (TSM) [29]

[2]As key objects we refer to objects that affect the outcome of the activity, e.g. in a toy assembly activity on the parts that can alter the result.

Under this premise, for the tasks of verb/noun next-segment prediction we restructure the available lexical information/labels of actions by discarding parts of the labels that refer to the usage of extra tools (annotated as nouns) to implement the corresponding action, i.e. the action labels are restructured to follow the format *action verb + noun*. In fact, this meta-processing of action labels that allow for a decoupled prediction of the next action verb or next action object(noun), is a common practice followed by the recent datasets targeting isolated motion motif or next-segment object prediction (e.g. Assembly-101 [47]). If such an action label format is available for the dataset in question, this label restructuring is skipped. The gain from such lexical decomposition is that the prediction task becomes simpler since the number of classes decreases, due to the fact that labels sharing the same action verb or action object (noun) are being merged, which allows for more samples to be associated with the specific motion motif or object state. Finally, in the case of the next action prediction (entire action context), we do not restructure the initial labels since the entire context of the preceding action labels is required to disambiguate between actions that share the same motion and object characteristics but differ on the execution medium.

3.3. Visual Action Recognition Module

The two aforementioned modules can be regarded as independent action anticipation models. In addition, a visual action recognition model is incorporated independently which during the inference stage operates on the same input sequence, denoted as x_t, as the action y_t. The purpose of this model is to provide estimates specifically for the current action y_t and fill the language-based action history.

Since the purpose of this model is to fill the action history, it remains independent from the action anticipation modules without any influence or connection, it can be trained separately and applied during the inference stage of the framework. In this work, instead of developing and training an action recognition module from scratch, we leverage the capabilities of state-of-the-art (SOTA) action recognition models that have been documented in the existing literature for each dataset. This approach is motivated by our objective to construct a visual-linguistic action anticipation framework, which can benefit from the advancements achieved by action recognition models specific to each dataset, thereby enhancing its overall performance.

4. Experimental Setup

We evaluate the proposed framework on three popular datasets of procedural activities. The main characteristics of the datasets are described in this section, such as the target activities, camera viewpoints, annotation data as well as multi-modal data and features provided (Section 4.1), followed by the evaluation protocols.

The experimental evaluation for the proposed framework follows a two-way narrative. Firstly, the population of the action history module involves simulating the prediction scores of a realistic action recognition model on a given dataset. This step aims to showcase the model's performance in relation to the latest advancements for each dataset. Subsequently, the complete potential of the model is presented by populating the action history module with past predictions obtained from an ideal visual action recognition model for each respective dataset. We should note that the realistic visual action recognizer performance follows the current SOTA action recognition scores reported for each examined dataset. Finally, we conduct experiments regarding different portions of the linguistic action history to assess the effect of the different action history sizes on the anticipation capabilities of the proposed framework.

4.1. Datasets

Meccano [41] is a multi-modal egocentric dataset created to study the interactions of humans and objects in industrial settings during instructional activities. Twenty different participants were requested to build a toy model of a motorbike. There exist 20 object classes, which include 16 classes that categorize 49 different toy components, 2 tool classes namely the screwdriver and the wrench, the instructions booklet, and a special class, noted as a partial model, for the under-construction toy object. Also, the dataset contains 12 verb classes and 61 action classes. In total, 20 videos are provided, 11 of which are used for training while the rest 9 videos are used for validation and testing.

The Meccano dataset provides gaze, object-centric features, and hands-centric features. The former type of features are computed based on the occurrences of the objects in each frame following the work of [12, 13]. Gaze features have been obtained by weighting the object-centric features with the distance between the center of objects bounding boxes and the gaze position in the image. The hand annotations of the dataset that contain the bounding boxes of both hands were used as hands-related features.

Assembly-101 [47] is a large-scale video dataset for the analysis and understanding of procedural activities regarding assembling and disassembling 101 "take-apart" toy vehicles captured from multiple viewpoints. In total 362 unique data sequences were captured synchronously by 4 egocentric and 8 static cameras and annotated with more than 100K coarse and 1M fine-grained action segments, targeting the challenging tasks of action recognition, action anticipation, temporal action segmentation, and mistake detection. Participants were instructed to disassemble and then assemble a toy vehicle without any instructions, which enhances the variability of the temporal ordering of actions performed by the participants during the procedural activities. A set of 90 object classes is considered that includes 5

tools together with the "hand". Also, 24 verbs are included along with the object classes form 1380 fine-grained action classes. A 60% of the available videos is used for training, while the rest 15% and 25% are utilized for validation and testing, respectively. Of the 101 toys, 25 of them are shared between all splits which sets the dataset even more challenging. For the RGB input, 2048-D frame-wise features are calculated using TSM [29] with an 8-frame input.

50Salads [53] is a multi-modal third-person instructional dataset of cooking-related activities. Twenty-five different participants prepared a set two mixed salads. The dataset provides RGB videos, depth maps, accelerometer data, and high- to low-level activity annotations. The dataset consists of 17 action classes. We report top-1 accuracy averaged over the 5 pre-defined splits following the work of [42].

4.2. Training, Testing & Input Configurations

As noted in Section 3, the structure of the action-centered temporal modeling sub-net follows a three-branch design, that acquires three vision-centered input sequences.

For the Mecanno dataset [41], input refers to the available feature representations for a) *Gaze*, b) *Objects*, and c) *Hands*. For the Assembly-101 [47], the available TSM [29] features for the RGB videos are utilized, which refer to frame-wise $[1 \times 2048]$ feature vectors. We restructure this representation to fit in the action-centric visual anticipation sub-net, as follows: a) split feature vectors into a set of two $[1 \times 1024]$ feature vectors to drive input to the first two branches and b) uniform sub-sampling is applied on the feature vector of the current frame of size $[1 \times 2048]$ into a $[1 \times 1024]$ and then calculate discrepancies between the sub-sampled feature representation of the previous frame to form the input feature vector for the third branch. For 50Salads [53] we utilized pre-extracted I3D features from [10, 42], which correspond to frame-wise $[1 \times 2048]$ feature vectors, which were restructured in the form described for the ones of the Assembly dataset.

Regarding the training configurations, the batch size was set to 4 for all datasets. The loss minimization is performed using the Adam optimizer, with a learning rate of 0.001. The input sequence length was set to 8 frames, while a random clip cropping sampling scheme was utilized. During the inference phase, we simulated the performance of the realistic visual action recognizer, by exploiting the SOTA performance of SlowFast [11] for Meccano, with 49.66% top1 accuracy, of TSM [29] for Assembly101 with 39.2% top1 accuracy, and, of Therbligs [9] for 50Salads with 76.5%.

5. Experimental Results

5.1. Action Anticipation

Predicting future actions is challenging, while modeling and performance greatly depend on the designated time horizon of the predictions. More specifically, predictions can be made at different decision points in time (timesteps) prior to the start of the next segment. In order to establish an extensive performance assessment of the proposed framework, we adopt the evaluation protocol reported in Furnari *et al*. [13] where predictions are made at 8 different anticipation timesteps before the start of the near-future anticipated action. Noted as τ_{ant}, the set of anticipation time refers to discrete values in the range of $[2s, 0.25s]$ for a timestep of $0.25s$. At the same time, the upper limit of this interval, that is $0.25s$ is closest to the start of the anticipated action.

Meccano: For the prediction of each action, the input to our framework regards information originating from the selected anticipation time point and runs backward, toward the start of the video (see Figure 1). As described in the previous sections, we exploit visual information related to the recent past (visual-action module) for modeling the short-term action history and the long-term past with the linguistic action history module. We report Top-1/Top-5 accuracy of the predicted action of the next segment, according to the [41]. In this work, the authors proposed the RULSTM framework [13] to anticipate the next action. We employ the publicly available code[3] of RULSTM for Meccano to replicate the experiments and also provide accurate results for the prediction of the noun and the verb of the next action-segment. We utilized a combination of information based on haze, object-centric and hand-centric features that are provided by [41], as those are the most discriminative features according to their experimental evaluation.

We evaluate the proposed VLMAH framework for action forecasting using different anticipation timesteps (see Table 1), and under the use of a realistic and an ideal (oracle) action predictor (denoted as VLMAH and VLMAH$_{GT}$ respectively) for past actions that populate the action history subnet. Under the use of a realistic visual action recognizer for past actions, our framework is compared to [41] which is the baseline and currently the SOTA method for the Meccano dataset. Our method outperforms the SOTA in Top-1 accuracy for the noun, verb, and action scenarios for almost every anticipation time, by a considerable margin. We present to have a slight decrease in performance in the Top-5 accuracy for the verb and action scenarios. This happens due to the impact of the action recognizer in the linguistic action history from which we draw information for making predictions. Our accuracy margin increases significantly from 4.1% up to 9.05% if we consider an ideal (oracle-like) visual action recognizer that feeds the linguistic action history module with the true past action classes. Any enhancement in action recognition accuracy is expected to similarly boost action anticipation, too.

[3] https://github.com/fpv-iplab/MECCANO

Top-1 / Top-5 Accuracy% @ different τ_{ant}									
		Timesteps							
Method		**2s**	**1.75s**	**1.5s**	**1.25s**	**1s**	**0.75s**	**0.5s**	**0.25s**
Meccano [41]		30.89/65.14	30.50/65.11	30.99/66.17	30.85/65.92	30.53/66.49	31.10/67.06	31.10/67.84	31.24/70.00
VLMAH		**33.12/77.85**	**32.12/77.78**	**31.48/78.49**	**32.33/80.41**	**31.25/76.30**	**32.17/82.39**	**34.07/78.58**	**38.34/79.19**
VMAH$_{GT}$	**Noun**	*15.91/72.58*	*27.63/69.46*	*25.37/65.83*	*28.93/73.29*	*26.21/70.31*	*25.08/71.73*	*28.83/69.81*	*29.50/70.88*
VLMAH$_{GT}$		*37.57/79.40*	*41.33/82.88*	*35.09/80.75*	*35.65/79.33*	*39.35/82.31*	*40.55/84.94*	*39.55/81.24*	*40.63/80.54*
Meccano [41]		36.06/ **93.19**	35.11/**93.01**	34.96/**92.98**	35.92/**93.19**	35.32/**93.38**	35.39/ **93.62**	34.75/ **93.76**	35.00/ **93.83**
VLMAH		**36.35**/93.00	**35.42**/92.33	**35.61**/91.31	**35.96**/92.88	**36.73**/91.08	**36.30**/90.62	**37.19**/91.14	**39.06**/90.93
VMAH$_{GT}$	**Verb**	*25.71/87.85*	*29.75/87.64*	*25.71/88.06*	*29.11/89.48*	*27.48/87.99*	*25.92/85.51*	*25.78/86.57*	*31.25/84.30*
VLMAH$_{GT}$		*40.76/91.40*	*41.26/93.39*	*40.83/92.61*	*43.39/92.96*	*39.91/91.69*	*40.98/93.18*	*43.67/92.68*	*43.55/91.65*
Meccano [41]		23.37/**54.65**	23.48/**55.99**	23.30/**56.56**	**23.97/57.73**	24.08/**58.23**	24.50/**59.96**	25.60/**61.31**	28.87/**63.40**
VLMAH		**24.75**/54.23	**24.35**/55.16	**24.22**/53.09	22.79/53.98	**28.90**/58.13	**25.29**/53.16	**26.47**/56.71	**29.12**/ 58.01
VMAH$_{GT}$	**Action**	*27.20/49.08*	*28.91/51.63*	*26.99/48.57*	*28.98/52.20*	*28.62/50.49*	*26.99/49.94*	*27.77/49.86*	*28.03/51.70*
VLMAH$_{GT}$		*34.73/67.75*	*36.86/69.53*	*35.01/67.18*	*34.30/69.24*	*35.15/68.25*	*33.59/67.89*	*34.65/66.90*	*33.09/65.98*

Table 1. Action anticipation accuracy for different timesteps (prior to the beginning of the next segment) for the **Meccano dataset**. *VLMAH*$_{GT}$ and *VMAH*$_{GT}$ represent the two variants of the proposed method when *ground truth annotations* are used as the linguistic action history. VLMAH makes use of the Linguistic Action History module while the action history is generated from the visual action recognition module. The comparison is between the [41] and the VLMAH methods.

Top-1/Top-5 Accuracy% @ $\tau_{ant} = 1s$			
Method	**Noun**	**Verb**	**Action**
TempAgg [47]	17.19 / 55.65	24.20 / 75.38	08.62 / 27.73
TempAgg [47]*	18.99 / 57.29	28.52 / 77.16	09.00 / 29.79
VLMAH	**27.70 / 54.37**	**42.17 / 82.52**	**14.18 / 30.95**
VMAH$_{GT}$	*22.68 / 55.32*	*40.59 / 85.11*	*13.14 33.98*
VLMAH$_{GT}$	*55.27 / 83.89*	*61.12 / 93.03*	*34.26 58.89*

Table 2. Top-1/Top-5 accuracy results of [47] and the VLMAH variants on the **Assembly-101 dataset** for anticipation time $\tau_{ant} = 1s$, with or without the use of the linguistic action history module. TempAgg$*$ denotes the single-task learning variant.

Top-1 Acc% @ $\tau_{ant} = 1s$	
Method	**Action**
DMR [55]	06.20
RNN [1]	30.10
CNN [1]	29.80
TempAgg [49]	40.70
AVT [14]	**48.00**
VLMAH	43.58
VLMAH$_{GT}$	*55.49*

Table 3. Top-1 accuracy results on the 50Salads dataset for the anticipation time $\tau_{ant} = 1s$.

Assembly101: In [47] that have also introduced the Assembly-101 dataset, action anticipation is performed at the fixed timestep $\tau_{ant} = 1s$. To assess action anticipation performance in [47], the TempAgg [49] method is used[4]. Both the VLMAH and the TempAgg methods are trained to generate predictions at anticipation time $\tau_{ant} = 1s$ that are evaluated using the Top-1 and Top-5 accuracy measures. Since the test split of the dataset is not yet available, we train and test both methods on the training and validation splits, respectively, using the egocentric viewpoint and data captured by the *e4* camera which yields the best results according to the experiments reported in [47]. Both the proposed VLMAH and the TempAgg methods have been trained/tested on data captured by this specific viewpoint.

Table 2 presents the accuracy results at $\tau_{ant} = 1s$. We provide two results for our framework. We compare our work with the state-of-art on Assembly-101 dataset, the TempAgg [49] framework. Our work is a single-task learning framework so for a fair comparison we test TempAgg [49] under two learning settings, a multi-task and a single-task. The single-task setting is denoted with $*$ in Table 2. The proposed approach outperforms state-of-the-art performance for the verb, noun, and action predictions by a large margin for this large and challenging dataset, even in the case that the linguistic action history module is not used. In particular, by using a realistic visual action recognizer to populate the action history module, an increase in accuracy of 13.65% for the verb prediction, 8.71% for the noun prediction, and 5.18% for the action prediction for $\tau_{ant} = 1s$ was reported. Similarly to *Meccano*, the use of an oracle-like visual action recognizer to verify/correct past estimates in the history module further increases the action anticipation performance of the proposed method. Even if we use only the visual information (VMAH), we outperform the TempAgg$*$ framework in general for a minimum of 4% up to 12%.

50-Salads: In Table 3 we present the accuracy scores at $\tau_{ant} = 1s$, and compare our proposed framework with recent works that tackle action anticipation in this dataset. We can observe that under the use of realistic action, recognizer to validate/correct the past action estimates stored in the action history module, our method is only surpassed by AVT [14] ($\approx 4\%$), with our proposed action anticipation method however, having a vastly lower number of trainable parameters (AVT: 378M, Ours: 10M), and ease in adapt-

[4]Code online at`https://github.com/assembly-101`

ing/incorporating the current action recognition advancements in each dataset.

5.2. How much history is enough?

In this study, we conducted ablation analyses to evaluate the performance of our proposed framework under various scenarios that pertain to the linguistic action history module's role and the required amount of linguistic action history to enhance the predictability power of the framework. Despite the fact that action history can obtain long-term information faster and with less cost compared to the visual features one question to be answered is "*how much history is enough?*". To answer this we evaluate our framework on the Assembly-101 dataset with different lengths of linguistic action history. From the previous sections, we have acquired the results of the evaluation of our framework with the full linguistic history of the observed actions[5]. In this experiment, we assess our framework by reducing the linguistic history to different percentages. The history percentages are in the range from 0% to 100%. Zero percent indicates the use of the VMAH$_{GT}$ framework while all the other percentages imply the use of the VLMAH$_{GT}$ framework with different percentages of action history. In this experiment, we use the VLMAH$_{GT}$ instead of VLMAH in order to assess the effect of the available size of action history in case no errors from the Visual Action Recognition module are present in the action history. As seen in Table 4, the results differ between the action and the verb/noun predictions considering different amounts of observed history.

Initially, all experiments were performed using 100% of the textual action history, which referred to a memorization capacity of 854 actions (slowest assembler). Our experiments show that, for the task of fine-grained action anticipation (full label), considering the entire linguistic history was the best strategy since it allowed us to disambiguate between cases of candidate actions that exhibited high similarity in their preceding action history.

In contrast, for the prediction of the coarse-grained verb and noun classes our experiments indicate that considering a more recent history is the best strategy. We observe that considering a larger percentage of the action history on these cases introduces noise that results in a considerable decrease in prediction accuracy, potentially due to similarities in the sequence of verb/noun transitions between different assembly scenarios. This is a valid assumption since, as stated in Section 3.2, in these tasks the initial action labels were restructured into a two-part-of-speech label (*verb+noun)*. This way, we discarded the fine-grained context of the label that refers to the mediums (tools) utilized to perform the action. For example, in the case of the action label pair "*screw cabin with screwdriver*" and "*screw cabin with hands*", which are two different action classes, the restructuring operation merged the two classes into the action "*screw cabin*". We note that in Assembly-101 similar format is provided as annotation data.

[5]A full history refers to the number of actions the slowest assembler from the training set performed to complete the assembling task.

Top-1/Top-5 Accuracy% @ $\tau_{ant} = 1s$			
History	**Noun**	**Verb**	**Action**
0%	22.68 / 55.32	40.59 / 85.11	13.14 / 33.98
1%	56.98 / 83.35	62.33 / 92.78	28.49 / 53.69
12.5%	56.86 / 84.08	62.83 / 93.40	28.20 / 51.38
25%	53.86 / 83.03	62.92 / 93.06	28.96 / 53.15
50%	56.92 / 84.54	63.99 / 93.16	27.13 / 51.02
75%	56.19 / 84.53	63.20 / 93.21	29.83 / 53.75
100%	52.16 / 83.81	61.12 / 93.03	34.51 / 58.44

Table 4. The Top1 and Top5 accuracy scores achieved by the proposed framework using variable lengths of the linguistic action history on the **Assembly-101 dataset**. Zero percent (0%) is equivalent to the use of VMAH$_{GT}$ variant, while other action history percentage values refer to the use of the VLMAH$_{GT}$.

6. Conclusions and Future Work

This paper assessed the impact of a language-driven history-logging method on action anticipation. This mechanism complements visual action representation by memorizing prior actions. We explored its performance and resilience across diverse past action misclassification rates and the length of encoded action history in anticipation tasks (action, motion motif, object). Our experiments reveal the strategy's benefits, notably enhancing scores on tough video datasets showing procedural activities. Moreover, the proposed method proves robust even with limited memory and high misclassification rates. Future research will investigate the effects of incorporating the history of preceding actions on long-range action anticipation and examine the impact of the temporal positions of miss-classifications (e.g., short-term and long-term past) on action anticipation accuracy.

Acknowledgements

This research was (a) co-financed by Greece and the European Union (European Social Fund-ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning" in the context of the Act "Enhancing Human Resources Research Potential by undertaking a Doctoral Research" Sub-action 2: IKY Scholarship Programme for PhD candidates in the Greek Universities, and (b) supported by the Hellenic Foundation for Research and Innovation (HFRI) under the "1st Call for HFRI Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment" (Project I.C.Humans, Number: 91) and under the "3rd Call for H.F.R.I. Research Projects to support Post-Doctoral Researchers" (Project InterLinK, Number: 7678).

References

[1] Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what?-anticipating temporal occurrences of activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5343–5352, 2018.

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

[3] Konstantinos Bacharidis and Antonis Argyros. Improving deep learning approaches for human activity recognition based on natural language processing of action labels. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[4] Konstantinos Bacharidis and Antonis Argyros. Cross-domain learning in deep har models via natural language processing on action labels. In *International Symposium on Visual Computing*, pages 347–361. Springer, 2022.

[5] Yizhak Ben-Shabat, Xin Yu, Fatemeh Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 847–859, 2021.

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[7] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018.

[8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23.

[9] Eadom Dessalene, Michael Maynord, Cornelia Fermüller, and Yiannis Aloimonos. Therbligs in action: Video understanding through motion primitives. In *CVPR*, 2023.

[10] Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3575–3584, 2019.

[11] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[12] Antonino Furnari and Giovanni Maria Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6252–6261, 2019.

[13] Antonino Furnari and Giovanni Maria Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4021–4036, 2020.

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

[15] Dayoung Gong, Joonseok Lee, Manjin Kim, Seong Jong Ha, and Minsu Cho. Future transformer for long-term action anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3052–3061, 2022.

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

[17] Jenhao Hsiao, Yikang Li, and Chiuman Ho. Language-guided multi-modal fusion for video action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3158–3162, 2021.

[18] De-An Huang and Kris M. Kitani. Action-reaction: Forecasting the dynamics of human interaction. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 489–504, Cham, 2014. Springer International Publishing.

[19] Youngkyoon Jang, Brian Sullivan, Casimir Ludwig, Iain Gilchrist, Dima Damen, and Walterio Mayol-Cuevas. Epic-tent: An egocentric video dataset for camping tent assembly. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[20] Evangelos Kazakos, Jaesung Huh, Arsha Nagrani, Andrew Zisserman, and Dima Damen. With a little help from my temporal context: Multimodal egocentric action recognition. *arXiv preprint arXiv:2111.01024*, 2021.

[21] Qiuhong Ke, Mario Fritz, and Bernt Schiele. Time-conditioned action anticipation in one shot. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9917–9926, 2019.

[22] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022.

[23] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[24] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 780–787, 2014.

[25] Sateesh Kumar, Sanjay Haresh, Awais Ahmed, Andrey Konin, M Zeeshan Zia, and Quoc-Huy Tran. Unsupervised

action segmentation by joint representation learning and online clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20174–20185, 2022.

[26] Tian Lan, Tsung-Chuan Chen, and Silvio Savarese. A hierarchical representation for future action prediction. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 689–704, Cham, 2014. Springer International Publishing.

[27] Yin Li, Miao Liu, and James 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)*, pages 619–635, 2018.

[28] Zhenyang Li, Kirill Gavrilyuk, Efstratios Gavves, Mihir Jain, and Cees GM Snoek. Videolstm convolves, attends and flows for action recognition. *Computer Vision and Image Understanding*, 166:41–50, 2018.

[29] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019.

[30] Miao Liu, Siyu Tang, Yin Li, and James M Rehg. Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 704–721. Springer, 2020.

[31] Tianshan Liu and Kin-Man Lam. A hybrid egocentric activity anticipation framework via memory-augmented recurrent and one-shot representation forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13904–13913, 2022.

[32] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, pages 852–869. Springer, 2016.

[33] Tahmida Mahmud, Mohammad Billah, Mahmudul Hasan, and Amit K Roy-Chowdhury. Prediction and description of near-future activities in video. *Computer Vision and Image Understanding*, 210:103230, 2021.

[34] Victoria Manousaki and Antonis Argyros. Segregational soft dynamic time warping and its application to action prediction. In *International Conference on Computer Vision Theory and Applications (VISAPP 2022)*, pages 226–235, 2022.

[35] Victoria Manousaki, Konstantinos Papoutsakis, and Antonis Argyros. Graphing the future: Activity and next active object prediction using graph-based activity representations. In *International Symposium on Visual Computing*, pages 299–312. Springer, 2022.

[36] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019.

[37] Megha Nawhal, Akash Abdu Jyothi, and Greg Mori. Rethinking learning approaches for long-term action anticipation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*, pages 558–576. Springer, 2022.

[38] Rashmiranjan Nayak, Umesh Chandra Pati, and Santos Kumar Das. A comprehensive review on deep learning-based methods for video anomaly detection. *Image and Vision Computing*, 106:104078, 2021.

[39] Konstantinos Papoutsakis, George Papadopoulos, Michail Maniadakis, Thodoris Papadopoulos, Manolis Lourakis, Maria Pateraki, and Iraklis Varlamis. Detection of physical strain and fatigue in industrial environments using visual and non-visual low-cost sensors. *Technologies*, 10(2), 2022.

[40] Zhaobo Qi, Shuhui Wang, Chi Su, Li Su, Qingming Huang, and Qi Tian. Self-regulated learning for egocentric video activity anticipation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[41] Francesco Ragusa, Antonino Furnari, Salvatore Livatino, and Giovanni Maria Farinella. The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1569–1578, 2021.

[42] Ivan Rodin, Antonino Furnari, Dimitrios Mavroeidis, and Giovanni Maria Farinella. Predicting the future from first person (egocentric) vision: A survey. *Computer Vision and Image Understanding*, 211:103252, 2021.

[43] Ivan Rodin, Antonino Furnari, Dimitrios Mavroeidis, and Giovanni Maria Farinella. Untrimmed action anticipation. In Stan Sclaroff, Cosimo Distante, Marco Leo, Giovanni M. Farinella, and Federico Tombari, editors, *Image Analysis and Processing – ICIAP 2022*, pages 337–348, Cham, 2022. Springer International Publishing.

[44] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. A database for fine grained activity detection of cooking activities. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1194–1201. IEEE, 2012.

[45] Marcus Rohrbach, Michaela Regneri, Mykhaylo Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele. Script data for attribute-based recognition of composite activities. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*, pages 144–157. Springer, 2012.

[46] Debaditya Roy and Basura Fernando. Action anticipation using latent goal learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2745–2753, 2022.

[47] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022.

[48] Fadime Sener, Rishabh Saraf, and Angela Yao. Transferring knowledge from text to video: Zero-shot anticipation for procedural actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[49] Fadime Sener, Dipika Singhania, and Angela Yao. Temporal aggregate representations for long-range video understanding. In *European Conference on Computer Vision*, pages 154–171. Springer, 2020.

[50] Fadime Sener and Angela Yao. Zero-shot anticipation for instructional activities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 862–871, 2019.

[51] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020.

[52] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[53] Sebastian Stein and Stephen J McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 729–738, 2013.

[54] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019.

[55] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 98–106, 2016.

[56] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021.

[57] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[58] Xinyu Xu, Yong-Lu Li, and Cewu Lu. Learning to anticipate future with dynamic context removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12734–12744, 2022.

[59] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 591–600, 2020.

[60] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015.

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

[62] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Anticipating Object State Changes in Long Procedural Videos

Victoria Manousaki[1,3,*], Konstantinos Bacharidis[1,2,*], Filippos Gouidis[1,2,*], Konstantinos Papoutsakis[3], Dimitris Plexousakis[1,2], and Antonis Argyros[1,2]

{vmanous, gouidis, kpapoutsakis}@hmu.gr, {kbach, dp, argyros}@ics.forth.gr

[1]Institute of Computer Science, FORTH
[2]Computer Science Department, University of Crete
[3]Management Science & Technology Department, Hellenic Mediterranean University
[*]Equal Contribution

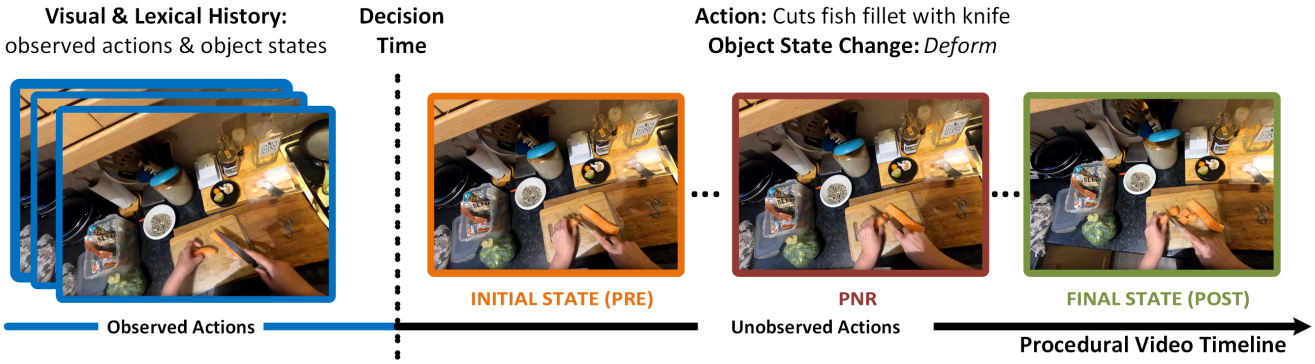

Figure 1. We introduce the new problem of anticipating object state changes in videos of procedural activities, noted OSCA. The decision point in the timeline is placed at the onset of the next anticipated, yet unobserved action. The objective is to predict accurately, at this point, the object state change class that will occur, if any, during the subsequent, yet unobserved action. This involves understanding the dynamics of past and current interactions and how they will affect the object's state. An object state change (e.g. deform) refers to a physical and possibly functional change in an object's attributes/properties. It is realized based on the transition from a pre-state (initial) to a post-state (final) occurring at the Point of No Return (PNR) time during an object state-modifying action (e.g. cut fish fillet using a knife).

Abstract

In this work, we introduce (a) the new problem of anticipating object state changes in images and videos during procedural activities, (b) new curated annotation data for object state change classification based on the Ego4D dataset, and (c) the first method for addressing this challenging problem. Solutions to this new task have important implications in vision-based scene understanding, automated monitoring systems, and action planning. The proposed novel framework predicts object state changes that will occur in the near future due to yet unseen human actions by integrating learned visual features that represent recent visual information with natural language (NLP) features that represent past object state changes and actions. Leveraging the extensive and challenging Ego4D dataset which provides a large-scale collection of first-person perspective videos across numerous interaction scenarios, we introduce an extension noted Ego4D-OSCA that provides new curated annotation data for the object state change anticipation task (OSCA). An extensive experimental evaluation is presented demonstrating the proposed method's efficacy in predicting object state changes in dynamic scenarios. The performance of the proposed approach also underscores the potential of integrating video and language cues to enhance the predictive performance of video understanding systems and lays the groundwork for future research on the new task of object state change anticipation. The source code and the

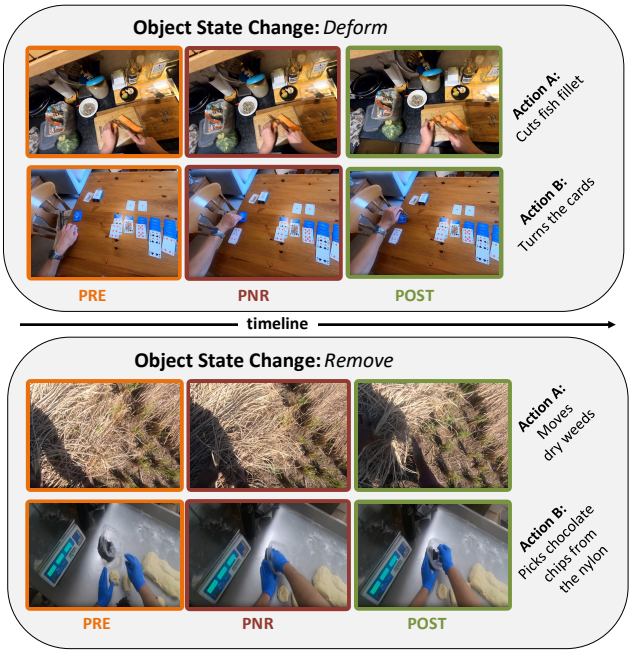

Figure 2. Examples of modifying actions from the "deform" and "remove" object state change classes represented by a triplet of frames (pre-state, PNR, post-state). Each state change is associated with various actions occurring in diverse environments/scenarios, emphasizing the complexity and challenges introduced in the OSCA problem.

new annotated data will be made publicly available[1].

1. Introduction

When observing human-object interactions, we can effortlessly reason about and anticipate changes in object states [3, 6, 35]. Imagine, for example, that while preparing the table for a dinner, somebody brings a bottle of wine. Even before opening it, we can infer that in the near future, the bottle will be "opened", and glasses will be "filled". Recognizing and anticipating object states and their changes is crucial for any entity that interacts with objects because the state of an object significantly affects its physical and functional properties and plays a decisive role in activity understanding, reasoning, and task planning.

While it is almost effortless for humans, the capability of predicting object state changes still lies beyond the competencies of current AI-powered systems [62, 66]. Understanding object states and their changes in the context of interactions relates to several challenging tasks, such as visual object perception, next-active object prediction, action recognition and anticipation, and object state estimation, that have been well-explored by the research community. Surprisingly, the problem of anticipating object state changes remains undefined and unexplored. However, recognizing and anticipating object state changes would be an important ability of AI-powered agents toward understanding human activities and task planning [57, 65].

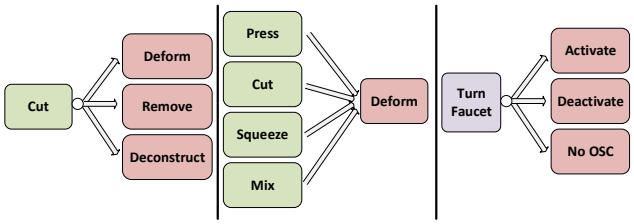

Figure 3. The intricate relation between verb/object/action and object state change. From left to right: one verb may signify different state changes; different verbs might signify the same state change; an action might lead to a variety of object state changes.

The problem of Object State Classification (OSC) is defined as the multi-class recognition of an object's state in a still image [16, 23], or the initial and the final object states in a video that demonstrates one or more state-modifying actions [19]. The binary object state change classification variant is also related to detecting state change occurrences in a short video clip [7, 19]. Researchers have only recently started to focus on methods for the representation and understanding of object state changes in videos in the context of state-modifying actions [41, 44, 47], which can also be seen as transformations [54]. Existing benchmarks largely ignore object state changes and focus on traditional types of annotations related to object type, location, or shape, attributes, affordances, and human actions.

In this work, we take one step beyond the Object State Classification and the Action Anticipation tasks by *introducing the new task of Object State Change Anticipation* in videos. OSCA focuses on the multi-class prediction of the state change occurring on an object during the next, yet unseen at inference time, action during a long procedural activity. Specifically, as shown in Fig. 1, at a certain decision point in time that is at the start of the next, yet unobserved action, we aim to predict the object state change class that will occur. The object state change will occur at the "Point of No Return" timestamp during the next action [19].

The OSCA task differs significantly from action anticipation, as it focuses on predicting imminent changes in the object's state, if any, without requiring the prediction of the verb and possibly the noun categories associated with the anticipated state-modifying action. Although the OSCA task involves fewer target state change categories than action categories, it remains challenging. A single action may apply to various objects and contexts, resulting in a range of possible state changes, as shown in Fig. 2 and 3. In addition, while it might seem that, given a prediction for the next

[1] https://projects.ics.forth.gr/cvrl/osca/

action, determining the next object state change would be straightforward, estimating such predictions in instructional videos is a challenging task, as current benchmarks suggest. Based on the Ego4D leaderboard for short-term action anticipation benchmark[2] (Top5 mAP: Overall approaches score 7%, verb and noun combined scores close to 17%, and noun-only score around 37% as of 11/2024), it is clear that a lot of research effort is still needed to devise efficient solutions to action anticipation in procedural activities and consequently their contribution to the estimation of the next state of the action-participating objects.

To tackle this new task, we introduce a new formulation leveraging on the cumulative history of the textual description of recognized preceding actions and object states, up to the decision point where this prediction is performed. The integration of the former information with visual information concerning the recent past is a key idea to model this historical context effectively. Our approach may also predict that no object state change will occur in the forthcoming action which implies the anticipation of a no-state-modifying action. We evaluate the proposed approach toward the newly proposed OSCA problem to establish baseline results for automated change state anticipation in long instructional videos and to also investigate the potential impact and effectiveness of the extracted information in other related vision-based anticipation tasks. We assess the performance of our proposed method for the OSCA problem and show initial results based on a proposed extension of the popular Ego4D video dataset. We hence build on the large-scale and challenging Ego4D dataset [19] which provides egocentric videos by augmenting the available annotation data with labels for the object state changes based on the initial and the final object states for any state-modifying actions in a subset of videos related to the Hand & Object Interactions benchmark[3]. This results in the Ego4D-OSCA, a variant of the Ego4D dataset that will become available to the community. Thus, the main contributions are:

- The introduction of the new problem of anticipating an object state change (OSCA) that will occur in the next, yet unseen, action in instructional videos.
- The introduction of the Ego4D-OSCA dataset, a new benchmark for evaluating solutions to the OSCA problem based on a subset of the Ego4D video dataset.
- The proposal of the first approach to tackle OSCA which integrates visual and language features to model the history of performed actions, object states, and their changes. We also present initial baseline results.

[2] https://eval.ai/web/challenges/challenge-page/1623/leaderboard/3910/Overall

[3] https://ego4d-data.org/docs/benchmarks/hands-and-objects/

2. Related Work

Object states capture dynamic aspects of object appearance and/or functionality and are subject to visually perceivable changes, as a result of state-modifying actions. They are also known as object fluents related to changeable object attributes [2, 26, 27]. Since there is no prior work on the problem of Object State Change Anticipation that we introduce, we review the literature on the closely related topics of action and next-active object anticipation, object state classification in images as well as the interplay between object state estimation and action recognition in videos.

2.1. Object State Classification/Recognition

Object State Classification in Images: Object states are typically considered as a special subset of "visual attributes", i.e. visual concepts that are related to the physical and functional properties of objects [23]. Object states and their changes are perceivable by humans and should be perceivable by AI-enabled agents [10]. The majority of the attribute classification approaches follow a similar approach to that of object classification by training a convolutional neural network with discriminative classifiers on annotated image datasets [43]. Few works focus explicitly on state classification [16], while most of them rely on the same assumptions used for the attribute classification task. A prominent direction to tackle this task refers to zero-shot learning. It gained considerable attention in recent years due to its practical significance in real-world applications, mitigating the problem of collecting and learning training data for a very large number of object classes [58]. One such prevalent approach involves the utilization of semantic embeddings to represent objects and their attributes in a low-dimensional space [55]. The work in [18] leverages Knowledge Graphs (KGs) and semantic knowledge in the context of zero-shot object classification. In a similar vein, the work in [17] combines KGs and Large Language Models (LLMs) to address object-agnostic state classification. A recent work by [41] focuses on object state recognition based on the compositional generation of novel object-state images, while the method in [47] introduced a novel conditioned diffusion model that focuses on generating temporally consistent and physically plausible images of actions and object state transformations based on an input image and a text prompt describing the targeted transformation.

Object State Change Estimation & Action Recognition in Videos: Object state changes have been considered a meaningful information source in video-based human action understanding and recognition (HAR). In HAR, object state changes are often considered complementary attributes to the visual representation of actions. These changes are typically derived within the visual domain through the utilization of explicit models for object detec-

tion and state estimation [11, 28, 44], or indirect modeling of object states based on general scene changes resulting from action execution [1, 4, 5]. Several methods exploit object states implicitly to estimate the type of action performed. The work in [1] was among the first to propose a method to automatically discover object states and the associated manipulation actions from videos by leveraging a discriminative clustering framework that jointly models the temporal order of object states and manipulation actions. The work in [27] explored the recognition of object fluents (changeable object attributes) and tasks (goal-oriented human activities) in egocentric videos using a hierarchical model that represents tasks as concurrent and sequential object fluents. Moreover, [29] focuses on understanding human actions within videos by analyzing complex interactions across multiple interrelated objects by recognizing their different state changes. In [44, 46] a multi-task self-supervised framework is proposed that allows the temporal localization of object state changes and state-modifying actions in uncurated web videos.

Furthermore, [60] introduced the novel VidOSC approach for understanding object state changes by segmenting object parts related to those changes in videos from an open-world perspective. A recently proposed framework [39] can recognize object-centric actions by relying only on the initial and final object states. The model can also generalize across unseen objects and different video datasets. The method proposed in [40] aims at disentangling visual embeddings that distinctly represent object states alongside identities, enabling effective recognition and generation of novel object-state compositions through a compositional learning framework. Finally, the InternVideo [7] video foundation model was adapted to tackle the tasks of object state change classification and action anticipation in the context of the Ego4D Challenges.

2.2. Action & Next-Active Object Anticipation

Action anticipation involves predicting the label of an action that is expected to occur in the future but has not yet been started/observed [21, 56, 64]. This challenge has been studied in both egocentric [8, 19] and exocentric [22, 42] videos, with the latter becoming increasingly popular in recent years. Short-term action anticipation [20, 34, 38] focuses on predicting actions or events in the immediate future, whereas long-term action anticipation [33, 63] extends to predicting actions or events over a longer period, ranging from several seconds to minutes.

In the context of human-object interactions in videos, *active objects* [36] and *next-active objects* [14] refer to specific items that are involved in ongoing or anticipated actions. The active object is the item that a person is currently interacting with within the video. In contrast, the next-active object is the item predicted to be used in the near future, based on the current interaction [9, 25, 30]. Although not yet in use, it is likely to become involved in subsequent actions. These concepts are crucial in video analysis for understanding and predicting human behavior, as they help anticipate the sequence of actions and interactions within a scene. The concept of next-active object anticipation has also been the subject of the short-term anticipation challenge in [19] and is described as the next object that will be touched by the user (either with their hands or with a tool) to initiate an interaction. Several methods have been proposed in this challenge for the solution of this problem [34, 37, 49, 50].

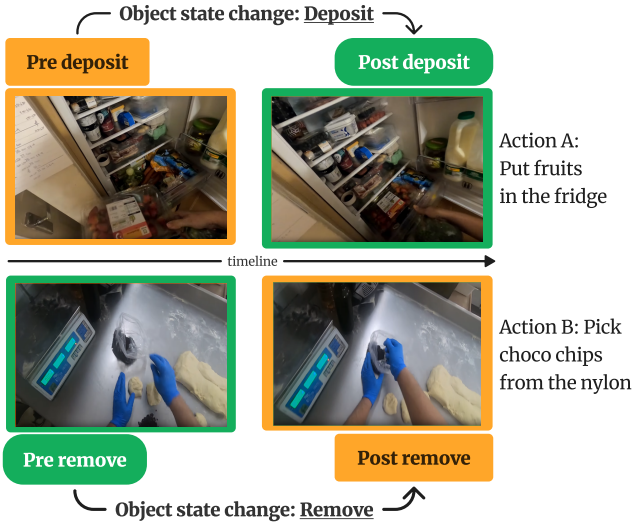

Figure 4. Two of the nine object state change super-annotated classes in the Ego4D-OSCA dataset, 'deposit' and 'remove'. Pre-/post-state labels for these actions are shown as distinct video segments. 'Deposit' and 'remove' are inverse changes, where pre-deposit matches post-remove, and pre-remove matches post-deposit, indicated by frames and shapes of the same color.

Anticipating the state change of an object involves predicting how the object's condition or form will alter as it becomes involved in an activity, whether it is currently active or will become active in the near future. This process goes beyond merely identifying which object will be used next(*next-active*); it focuses on understanding how the object's state will evolve during or after its involvement in the interaction. This anticipatory process requires analyzing the current interaction and understanding the transformations that occur as objects are used. By predicting these state changes, we gain insights into how objects will behave as they become active or next-active, enhancing the ability to interpret the sequence of events and interactions in a video.

3. Ego4D-OSCA Dataset

We introduce Ego4D-OSCA as a new partition of the large-scale Ego4D dataset that aims to serve as a benchmark for the assessment of methods for object state change antici-

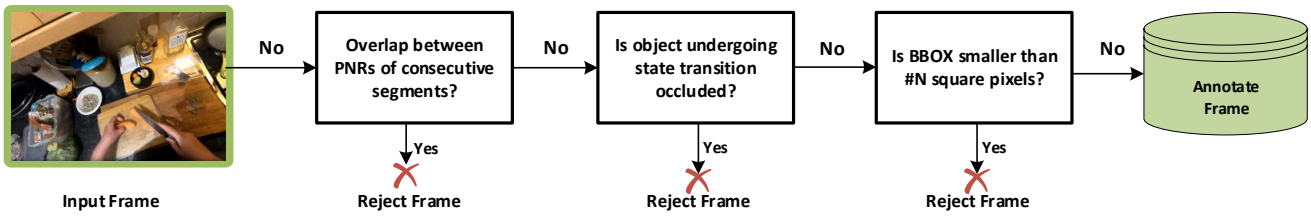

Figure 5. **Annotation pipeline:** Occlusions are checked in the pre- & post-frames. A threshold value for the BBOX area of N square pixels (N=100) for each object annotation. Ego4D-SCOD benchmark data are used to automatically annotate the change states per clip.

pation. The volume and diversity of the Ego4D dataset make Ego4D-OSCA a very challenging dataset for OSCA, as shown in Fig. 2. Ego4D-OSCA is tailored from the long-term activity (LTA) prediction benchmark, which aims to forecast the sequence of activities that will unfold in future video frames. Due to an ongoing challenge, the official test set for this benchmark has not yet been released, prompting us to re-purpose the validation split as a stand-in test set.

Dataset annotation: To enrich the LTA benchmark, we integrate object state annotations extracted from the dataset as follows. The original Ego4D dataset does not include annotations for the specific state labels of individual video frames. Instead, annotations about state changes are provided, which relate to entire video segments. Additionally, the dataset includes annotations for bounding boxes and object classes across seven critical frames within each video segment. These frames are temporally centered around the occurrence of the state change that occurs within each video segment. Based on this information, we super-annotate certain critical frames of each video segment with state-related labels as follows. For each video segment, we annotate the initial and final frames as pre_X and $post\_X$, respectively, where X denotes the label of the state change. Furthermore, in line with the semantic implications of these changes, we establish three pairs of state changes. Each pair is constructed under the premise that the first action is the inverse of the second concerning the resulting state change. For instance, if X and Y represent inverse state changes, then the labels pre_X and $post\_Y$ are considered samples of identical states. A similar correspondence applies between pre_Y and $post\_X$. For example, the states pre_remove and $post\_deposit$ are considered identical, since $remove$ and $deposit$ constitute a pair of inverse state changes. Figure 4 delineates the specifics for these two super-annotated state change classes. The same condition is true for the state change classes of activate-deactivate and construct-deconstruct. The full set of pre- and post- object state pairs that constitute the target set of object state changes appear in the supplementary material. It is important to note that there exists a subset of video segments containing actions that do not induce state changes and, therefore, these segments are not considered for annotation.

By incorporating these detailed annotations, Ego4D-OSCA offers researchers a comprehensive platform to explore and refine methods for anticipating object behavior and activity sequences in egocentric video contexts. The Ego4D dataset offers eight distinct state change labels: *activate, deactivate, deposit, remove, construct, deconstruct, deform, and other*. However, we contend that there are actions that do not alter the state of an object. To address this, we propose adding a state change category called "No Object-State-Change (No OSC)". This new class will help capture instances where actions occur without affecting the state of an object, thereby providing a more comprehensive framework for understanding and categorizing interactions.

Details on the annotation process: The annotation process for the state transitions is applied to the pre_Y and $post\_X$ frames in each video segment. Overall, the annotation process consists of the next 4 steps (a schematic representation of the annotation pipeline is shown in Fig. 5). First, the PNR moment of the video segment being examined for annotation is compared to the PNR moment of the segment that has been previously annotated. If the PNR of the previously annotated segment is located after the PNR of the segment under examination, then the segment under examination is rejected. The reasoning behind this decision has to do with the learning of the segment features related to state transitions. This alignment of the two PNRs signifies that there is an overlap between the state transition actions of the two segments and therefore the feature learning becomes more challenging. Subsequently, it is examined if the object undergoing state change is occluded. If this is the case, the frame is rejected. Then, the bounding box area of the object is evaluated, and if it is below 100 square pixels—a threshold empirically chosen and commonly used in annotation tools like Voxel-51—the frame is discarded. Finally, the frame that has passed all the previous checks is annotated with the appropriate state label that pertains to the state transition action.

Dataset Statistics: The proposed *Ego4D-OSCA* dataset is compiled using a subset of the popular, large-scale Ego4D v2 dataset [19] that contains egocentric videos for a large variety of human daily living or work activities. In Table 1 we compare the proposed *Ego4D-OSCA* dataset with existing image and video datasets that also provide annotations related to object states. Ego4D-OSCA contains 61.858

Datasets	Modalities	OSC related task	Actions per video	Samples	Obj. State Classes	Actions	Objects
Fire et al. [13]	Videos	Detection	Single	490	17	-	13
ChangeIt [45]	Videos	Temporal Localization	Single	34.428	-	44	-
HowToChange [59]	Video & Text	Temporal Localization	Single	498.475	20	-	134
VSCOS [61]	Video	Segmentation	Single	1.905	4	271	124
VOST [51]	Video	Segmentation	Single	713	-	-	155
Ego4D [19]	Video & Text	Detection & Classification	Single	92.864	8	-	478
Ego4D-OSCA (Ours)	Video & Text	**Anticipation**	**Multiple**	1610	9	1500	477

Table 1. Comparison with other image and video datasets that contain annotations related to object state changes. Modalities refer to the available data source for the object state change-related tasks.

training and 31.846 testing clips. The target tasks performed using each of the datasets are also noted. More dataset statistics can be found in the supplementary material.

4. Object State Change Anticipation - Baseline

The proposed framework, depicted in Fig. 6, draws inspiration from the efficacy of combining visual and lexical information for semantic action/activity encoding. To achieve this, it adopts a three-stream architecture. Within this design, a visual encoding module is tasked with capturing the visual attributes of ongoing actions, while two lexical-based encoders are employed to extract the semantic nuances from a procedural-oriented representation of past actions and object states. The framework fuses these distinct representations towards the unified objective of anticipating the next object state. This task entails the estimation of the forthcoming state in which the object of interest will reside during the subsequent action. The framework tries to holistically capture the underlying dynamics and contextual intricacies governing object-state transformations across sequential actions by integrating visual and lexical cues.

The design of our framework draws from the recent VLMAH model [31] that was specifically tailored for the task of *action* anticipation. We augment this architecture by introducing specialized object state history encoding modules. Additionally, we redesign the action history module to facilitate disjoint encoding, capturing both the motion motifs in actions (verbs) and the transitions of objects-in-use (nouns) between actions. This refined architecture enables a more nuanced representation of the sequential dynamics between actions and object states, empowering the framework to achieve enhanced performance in the task of next-state anticipation within dynamic environments.

As illustrated in Fig. 6, the proposed framework consists of two primary components: (a) the current action and object state estimation module, and (b) the object state anticipation module, depicted within the thin-dotted rectangle. Our contribution resides in (a) the conceptualization of this framework and (b) the development of the object state anticipation module. Concerning the latter, it encompasses some constituent components.

Visual Encoder: For this module, we employ a lightweight visual encoder consisting of a single-branch bidirectional long short-term memory (BiLSTM) component followed by a multi-layer perceptron (MLP). We selected this simplified design for the visual encoder based on the objective of temporally encoding the enduring relationships among encoded short-term segments extracted from the input video. Our model relies on an external pretrained human action recognition model, such as SlowFast [12] or TSN [53] to provide encodings of short-term spatio-temporal dependencies between the frames inside a single segment.

Action & State History Encoders: As illustrated in Fig. 6, both encoders exploit the model design of the lexical encoder of the VLMAH model [31], which follows a NLP neural network design consisting of BiLSTM and MLP components. The decision to employ a simple NN for encoding the history, instead of utilizing LLMs was motivated by several factors. Firstly, the computational efficiency of LLMs such as GPT- or LLaMA, often entails significant resource requirements for training and inference [52], whereas a simpler neural network architecture mitigates computational overhead. Secondly, LLMs are pre-trained on general text corpora and may not capture the domain-specific nuances inherent in the textual data related to action histories and object states. Additionally, the simplicity of the chosen architecture facilitates interpretability, data efficiency, and customization, affording greater control over the model's behavior and adaptation to the task's requirements.

Learning Objective: The objective for training the model was exclusively focused on evaluating the anticipated state estimate. This deliberate choice stemmed from the aim to prioritize the accurate prediction of object states, which was the study's primary objective. This objective was formulated using the cross-categorical entropy loss, which is well-suited for multi-class classification tasks, such as predicting object states across different categories:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c}), \tag{1}$$

where N is the number of samples in the dataset, C is number of object state categories, $y_{i,c}$ is the ground truth next state label for the object-in-use in the current action sample i and $\hat{y}_{i,c}$ is the predicted next state probability.

During training, the proposed framework leverages oracle action and state detectors to provide the action and state

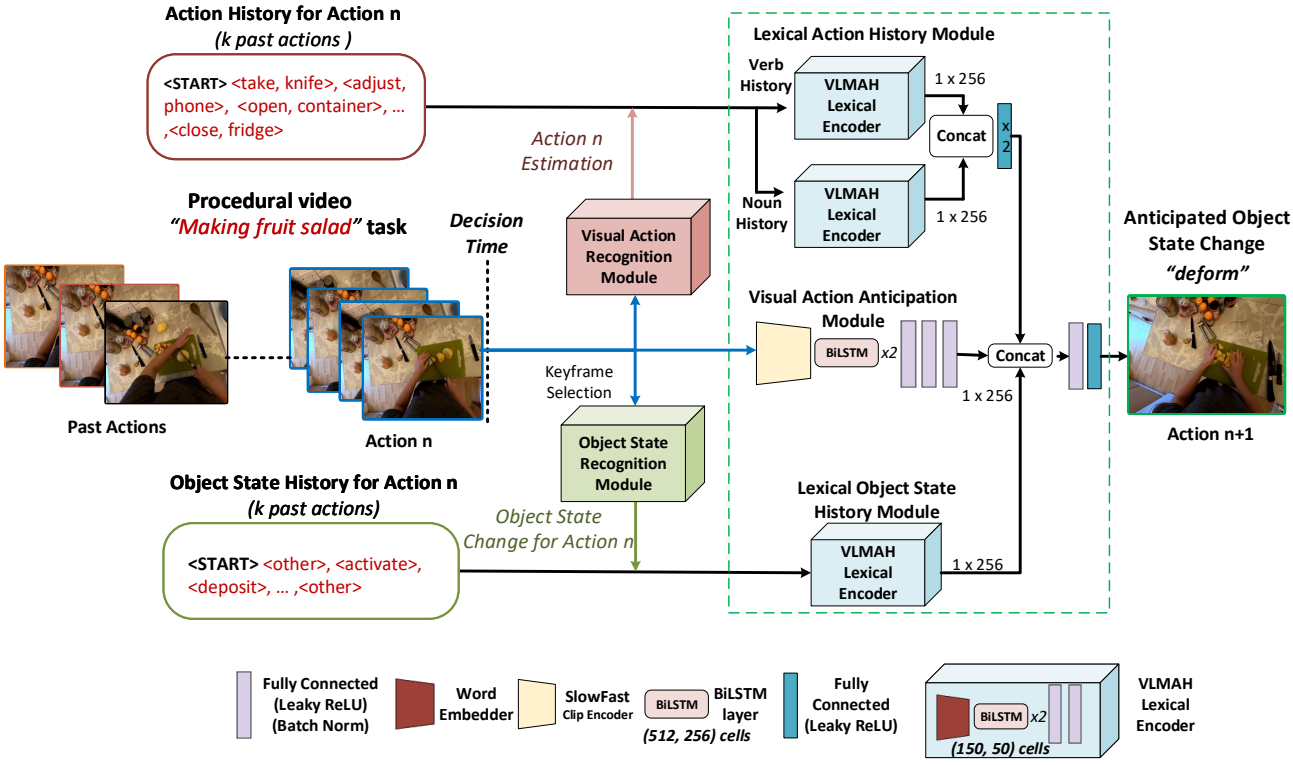

Figure 6. Overview of the proposed baseline framework for the object state change anticipation task. The proposed framework anticipates object state changes by integrating real-time visual data and a historical record of past actions and object state changes.

history, respectively, for each clip in the dataset (see Fig. 6). These detectors estimate the current action and object state observed in the clip, serving as ground truth annotations for training purposes. However, it is important to note that for inference toward real-world applications, there is a requirement for current action and object state recognition models to provide input to the framework. Consequently, our model is solely tasked with the learning objective of next-object anticipation, focusing exclusively on predicting the future state of the object. By decoupling the training and inference phases in this manner, the model can effectively learn the dynamics of object-state transitions without the added complexity of simultaneously predicting the current action.

5. Implementation, Experiments and Results

Implementation Details: The proposed state anticipation model (dotted rectangle in Fig. 6) is trained on a single NVIDIA TITAN GPU using the Adam optimizer, a batch size of 32, a learning rate of $1e-4$, without any temporal augmentations (clip or frame cropping). Short-term associations between neighboring segments of an input video are represented using the pre-extracted SlowFast frame level features from Ego4D. Regarding the selection of the pre- and the post-state keyframes that are introduced in the object recognition model, we exploit the PNR annotations of Ego4D, which correspond to the first frame in each clip when the state change/transition is visible. We should note that in real-world inference, the action and state lexical histories in the proposed anticipation model will be populated by existing action recognition and object state estimation models trained on the respective data of the task.

Evaluation metrics: The evaluation of all examined models was conducted using top-1/5 mean accuracy, and F-score, following standard practices in the relevant literature.

5.1. OSCA Results

In Table 2, we compare variants of an object state anticipation model to highlight the impact of incorporating lexical histories of past actions and object states on the anticipation performance. The vision-only model (VID-A) only relies on the visual representation of the current action. We observe modest performance levels.

OSCA under ideal action and state recognition: When ground-truth lexical histories of past actions are introduced through an oracle recognition model (VNLP (O-Action)), a slight performance improvement indicates the potential benefit of contextual action information. Notably, incorporating lexical histories of past object states from an oracle recognition model (VNLP (O-State)) leads to significant performance gains, that highlight the importance of considering object state dynamics in anticipation tasks. Fur-

Model	Top@1/5 mAcc	F1-score
VID-A	23.93 / 89.10%	11.74%
VNLP (O-Action)	25.59 / 83.06%	24.62%
VNLP (O-State)	**40.07** / 90.83%	33.57%
VNLP (O-Action, O-State)	39.20 / 89.76%	**37.12%**
VNLP (Action [12])	23.04 / 81.31%	22.09%
VNLP (State [16])	**32.72** / 92.16%	21.78%
VNLP (Action [12], State [16])	29.42 / 94.65%	**26.29%**

Table 2. OSCA performance for various model configurations (O-: Oracle recognizer, VID-A: vision-only state anticipation model).

ther improvements are observed when both lexical histories are integrated into the model (VNLP (O-Action, O-State)), demonstrating the synergistic effect of leveraging contextual information from actions and object states. Overall, the low anticipation scores highlight the inherently challenging nature of the task and the intricacy of the dataset scenarios that pose significant challenges for anticipation models.

OSCA under actual action and state recognition: We conducted experiments where the oracles (action recognizer, object's current state estimator), were substituted with existing baseline models. In this setting, the output of these models is utilized to populate the action and state histories that OSCA utilizes. We employed the following models: (a) the well-established SlowFast [12] model for action recognition; (b) the object-agnostic state classification method proposed by [16] as the current state classifier, with minor modifications, as described below.

Action recognizer: Regarding the action classification module, we fine-tuned the SlowFast model [12] on a subset of the original Ego4D dataset. Specifically, we obtained the train/validation/test splits using the training set provided for the (LTA) long-term anticipation task of Ego4D based on the $60/20/20$ split scheme. The adaptation of the LTA data to the action recognition task resulted in 5754 action classes and a total of $\approx 65K$ video clips (with a mean of ≈ 10.7 sample clips per action class). SlowFast achieved 12.86% Top-1 and 33.69% Top-5 accuracy for the task of current action recognition. This low performance can be attributed to the extensive number of action classes and limited samples per class, as well as in Ego4D's inherent motion and appearance similarities across different actions, e.g. *take cup - take bottle, tie string - tie rope*.

Object state recognizer: Additionally, for the object-agnostic state history, we adapt the model of [16]. This model relies on the outputs of two distinct state classifiers. Each classifier receives the first (pre) or the last (post) frame of each video segment as input to predict the object state label for the respective frame. The prediction of the state-change label for the video segment considers both outputs and is derived based on the following rules. If the object state predictions are pre_X and $post\_X$, respectively, the inferred state change for the video segment is denoted as X. Conversely, if the classifiers predict pre_X and $post\_Y$, where X and Y are distinct and represent inverse state changes, it is concluded that no state change has occurred. Finally, if neither of the above conditions is met, the prediction of the state change defaults to the output of the second classifier; that is, if the prediction is $post\_Y$, the state change for the video segment is identified as Y. For example, if the predictions of the two classifiers are $pre\_activate$ and $post\_activate$ the prediction of state change would be $activate$. Likewise, if the predictions are $pre\_activate$ and $post\_deactivate$ the prediction of the state change would be that of no change. This object-agnostic state recognizer showcased 25.4% mean state recognition accuracy.

In our experiments, replacing oracles with realistic recognizers to populate the action and state history buffers that are considered by the baseline OSCA model, we observed a significant accuracy drop (last block of lines in Table 2). This accuracy difference underscores the critical role of precise recognition of the current action and object state for effective anticipation of near-future object states within dynamic environments. Given the class imbalance in the proposed Ego4D-OSCA data set, the F1 score is a more appropriate performance measure. This metric considers both precision and recall while remaining insensitive to the true negatives of majority classes, unlike accuracy, which can be biased toward the majority class. Based on this rationale, the reported F1-scores in Table 2 indicate that a model combining both action and state history (past context) may be more effective for the object state anticipation task.

Noise (Action, State)	Top@1/5 mAcc
(0%, 0%) (Oracle)	**35.60 / 88.14%**
(25%, 25%)	30.46 / 84.42%
(50%, 50%)	26.00 / 81.75%
(75%, 75%)	22.48 / 78.09%

Table 3. The robustness of the object state change anticipation model is tested to the recognizer performance variability.

5.2. Object State & Action Recognition Impact

To further demonstrate the impact of the current action and object state recognizer accuracies on the object state change anticipation task, we conducted experiments that hypothesized recognizers of different accuracy. In this experimental setup, we uniformly introduce noise, representing erroneous estimations, to both the action and state histories, since in the inference stage of the proposed framework, these histories would need to be populated by the outputs of the respective recognizers.

Table 3 presents the results obtained under three varying levels of label noise (rows 2-4), contrasted to the outcomes achieved when employing ground truth labels (where the

noise level is 0%). The noise levels correspond to the rate of erroneous estimates generated by the recognizers, i.e., 25% corresponds to a recognizer with 75% mAcc. As it can be verified based on the obtained results, the performance of the state anticipation task is influenced by the recognizer's accuracy, demonstrating an approximate $4-5\%$ reduction in OSCA accuracy for every 25% decrease in object state and action recognition accuracy. Notably, despite substantial declines in state and action recognition performance, the anticipation model exhibits only a marginal decrease in performance. This finding can be attributed to the compensatory capability of the visual component of the model, which effectively accommodates dynamic and previously unseen sequences of action and state histories.

6. Conclusions

This paper introduced the new problem of object state change anticipation during procedural activities. We proposed a novel framework that integrates lexical histories of past actions and object states with recent visual information to enhance anticipation accuracy in vision-based models. By fusing long-term semantic and recent visual information, our framework demonstrates notable improvements in anticipation accuracy, underscoring the importance of contextual understanding in dynamic environments. To validate our approach, we augmented the Ego4D dataset forming a specialized subset noted Ego4D-OSCA. Future work will explore the applicability of LLMs as a replacement for the NLP processing component of the proposed framework, for leveraging their enhanced semantic understanding and in-context learning abilities. We also plan to explore zero-shot settings to enable anticipation of state changes involving novel, previously unseen, objects or actions.

References

[1] Jean-Baptiste Alayrac, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. Joint discovery of object states and manipulation actions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 407–416, 2017. 4, 2

[2] Jean-Baptiste Baptiste Alayrac, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. Joint Discovery of Object States and Manipulation Actions. In *International Conference on Computer Vision (ICCV)*, pages 2146–2155, 2017. 3

[3] Patric Bach, Toby Nicholson, and Matthew Hudson. The affordance-matching hypothesis: how objects guide action understanding and prediction. *Frontiers in human neuroscience*, 8:254, 2014. 2

[4] Konstantinos Bacharidis and Antonis Argyros. Exploiting the nature of repetitive actions for their effective and efficient recognition. *Frontiers in Computer Science*, 4, 2022. 4

[5] Konstantinos Bacharidis and Antonis Argyros. Repetition-aware image sequence sampling for recognizing repetitive human actions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1878–1887, 2023. 4

[6] Andreja Bubic, D Yves Von Cramon, and Ricarda I Schubotz. Prediction, cognition and the brain. *Frontiers in human neuroscience*, 4:1094, 2010. 2

[7] Guo Chen, Sen Xing, Zhe Chen, Yi Wang, Kunchang Li, Yizhuo Li, Yi Liu, Jiahao Wang, Yin-Dong Zheng, Bingkun Huang, et al. Internvideo-ego4d: A pack of champion solutions to ego4d challenges. *arXiv preprint arXiv:2211.09529*, 2022. 2, 4

[8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2022. 4

[9] Eadom Dessalene, Chinmaya Devaraj, Michael Maynord, Cornelia Fermüller, and Yiannis Aloimonos. Forecasting action through contact representations from first person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):6703–6714, 2021. 4

[10] Kun Duan, Devi Parikh, David Crandall, and Kristen Grauman. Discovering localized attributes for fine-grained recognition. In *2012 IEEE CVPR*, pages 3474–3481. IEEE, 2012. 3

[11] Alireza Fathi and James M Rehg. Modeling actions through state changes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2579–2586, 2013. 4

[12] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 6, 8

[13] Amy Fire and Song-Chun Zhu. Inferring hidden statuses and actions in video by causal reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 48–56, 2017. 6, 2

[14] Antonino Furnari, Sebastiano Battiato, Kristen Grauman, and Giovanni Maria Farinella. Next-active-object prediction from egocentric videos. *Journal of Visual Communication and Image Representation*, 49:401–411, 2017. 4

[15] Filippos Gouidis, Theodore Patkos, Antonis Argyros, and Dimitris Plexousakis. Detecting object states vs detecting objects: A new dataset and a quantitative experimental study. *arXiv preprint arXiv:2112.08281*, 2021. 2

[16] F. Gouidis, T. Patkos, A. Argyros, and D. Plexousakis. Detecting object states vs detecting objects: A new dataset and a quantitative experimental study. In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, pages 590–600, 2022. 2, 3, 8

[17] F. Gouidis, K. Papantoniou, K. Papoutsakis, T. Patkos, A. Argyros, and D. Plexousakis. Fusing domain-specific content from large language models into knowledge graphs for enhanced zero shot object state classification. In *Proceedings of the AAAI 2024 Spring Symposium on Empowering Ma-*

chine Learning and Large Language Models with Domain and Commonsense Knowledge (AAAI-MAKE), 2024. 3

[18] F. Gouidis, K. Papoutsakis, T. Patkos, A. Argyros, and D. Plexousakis. Exploring the impact of knowledge graphs on zero-shot visual object state classification. In *Proceedings of the the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications.*, 2024. 3

[19] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF CVPR*, pages 18995–19012, 2022. 2, 3, 4, 5, 6, 1

[20] Hongji Guo, Nakul Agarwal, Shao-Yuan Lo, Kwonjoon Lee, and Qiang Ji. Uncertainty-aware action decoupling transformer for action anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18644–18654, 2024. 4

[21] Xuejiao Hu, Jingzhao Dai, Ming Li, Chenglei Peng, Yang Li, and Sidan Du. Online human action detection and anticipation in videos: A survey. *Neurocomputing*, 491:395–413, 2022. 4

[22] Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, et al. Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22072–22086, 2024. 4

[23] Phillip Isola, Joseph J. Lim, and Edward H. Adelson. Discovering states and transformations in image collections. *Proceedings of the IEEE Computer Society CVPR*, 07-12-June:1383–1391, 2015. 2, 3

[24] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1383–1391, 2015. 2

[25] Jingjing Jiang, Zhixiong Nan, Hui Chen, Shitao Chen, and Nanning Zheng. Predicting short-term next-active-object through visual attention and hand position. *Neurocomputing*, 433:212–222, 2021. 4

[26] Yang Liu, Ping Wei, and Song-Chun Zhu. Jointly recognizing object fluents and tasks in egocentric videos. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2943–2951, 2017. 3

[27] Yang Liu, Ping Wei, and Song Chun Zhu. Jointly Recognizing Object Fluents and Tasks in Egocentric Videos. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-Octob:2943–2951, 2017. 3, 4

[28] Yang Liu, Ping Wei, and Song-Chun Zhu. Jointly recognizing object fluents and tasks in egocentric videos. In *Proceedings of the IEEE ICCV*, pages 2924–2932, 2017. 4, 2

[29] Chih Yao Ma, Asim Kadav, Iain Melvin, Zsolt Kira, Ghassan Alregib, and Hans Peter Graf. Attend and Interact: Higher-Order Object Interactions for Video Understanding. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 6790–6800, 2018. 4

[30] Victoria Manousaki, Konstantinos Papoutsakis, and Antonis Argyros. Graphing the future: Activity and next active object prediction using graph-based activity representations. In *International Symposium on Visual Computing*, pages 299–312. Springer, 2022. 4

[31] Victoria Manousaki, Konstantinos Bacharidis, Konstantinos Papoutsakis, and Antonis Argyros. Vlmah: Visual-linguistic modeling of action history for effective action anticipation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1917–1927, 2023. 6

[32] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019. 3

[33] Himangi Mittal, Nakul Agarwal, Shao-Yuan Lo, and Kwonjoon Lee. Can't make an omelette without breaking some eggs: Plausible action anticipation using large video-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18580–18590, 2024. 4

[34] Razvan-George Pasca, Alexey Gavryushin, Muhammad Hamza, Yen-Ling Kuo, Kaichun Mo, Luc Van Gool, Otmar Hilliges, and Xi Wang. Summarize the past to predict the future: Natural language descriptions of context boost multimodal object interaction anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18286–18296, 2024. 4

[35] Giovanni Pezzulo. Coordinating with the future: the anticipatory nature of representation. *Minds and Machines*, 18: 179–225, 2008. 2

[36] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2847–2854. IEEE, 2012. 4

[37] Francesco Ragusa, Giovanni Maria Farinella, and Antonino Furnari. Stillfast: An end-to-end approach for short-term object interaction anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3636–3645, 2023. 4

[38] Debaditya Roy, Ramanathan Rajendiran, and Basura Fernando. Interaction region visual transformer for egocentric action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6740–6750, 2024. 4

[39] Nirat Saini, Bo He, Gaurav Shrivastava, Sai Saketh Rambhatla, and Abhinav Shrivastava. Recognizing actions using object states. In *ICLR Workshop on Objects, Structure and Causality*, 2022. 4

[40] Nirat Saini, Khoi Pham, and Abhinav Shrivastava. Disentangling visual embeddings for attributes and objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 4

[41] Nirat Saini, Hanyu Wang, Archana Swaminathan, Vinoj Jayasundara, Bo He, Kamal Gupta, and Abhinav Shrivas-

tava. Chop & learn: Recognizing and generating object-state compositions. In *Proceedings of the International Conference on Computer Vision (ICCV)*. IEEE, 2023. 2, 3

[42] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022. 4

[43] Krishna Kumar Singh and Yong Jae Lee. End-to-end localization and ranking for relative attributes. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 753–769. Springer, 2016. 3

[44] Tomáš Souček, Jean-Baptiste Alayrac, Antoine Miech, Ivan Laptev, and Josef Sivic. Look for the change: Learning object states and state-modifying actions from untrimmed web videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 4

[45] Tomáš Souček, Jean-Baptiste Alayrac, Antoine Miech, Ivan Laptev, and Josef Sivic. Look for the change: Learning object states and state-modifying actions from untrimmed web videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13956–13966, 2022. 6, 2

[46] Tomáš Souček, Jean-Baptiste Alayrac, Antoine Miech, Ivan Laptev, and Josef Sivic. Multi-task learning of object states and state-modifying actions from web videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 4

[47] Tomáš Souček, Dima Damen, Michael Wray, Ivan Laptev, and Josef Sivic. Genhowto: Learning to generate actions and state transformations from instructional videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3

[48] Masatoshi Tateno, Takuma Yagi, Ryosuke Furuta, and Yoichi Sato. Learning object states from actions via large language models. *arXiv preprint arXiv:2405.01090*, 2024. 2

[49] Sanket Thakur, Cigdem Beyan, Pietro Morerio, Vittorio Murino, and Alessio Del Bue. Anticipating next active objects for egocentric videos. *IEEE Access*, 2024. 4

[50] Sanket Thakur, Cigdem Beyan, Pietro Morerio, Vittorio Murino, and Alessio Del Bue. Leveraging next-active objects for context-aware anticipation in egocentric videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8657–8666, 2024. 4

[51] Pavel Tokmakov, Jie Li, and Adrien Gaidon. Breaking the "object" in video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22836–22845, 2023. 6, 2, 3

[52] Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, Mosharaf Chowdhury, et al. Efficient large language models: A survey. *arXiv preprint arXiv:2312.03863*, 1, 2023. 6

[53] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(11): 2740–2755, 2019. 6

[54] Xiaolong Wang, Ali Farhadi, and Abhinav Gupta. Actions $\sim$ Transformations. In *Proceedings of the IEEE Computer Society CVPR*, pages 2658–2667. IEEE, 2016. 2

[55] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-Shot Recognition via Semantic Embeddings and Knowledge Graphs. *Proceedings of the IEEE CVPR*, pages 6857–6866, 2018. 3

[56] Richard Wardle and Sareh Rowlands. Deep-learning based egocentric action anticipation: A survey. 2023. 4

[57] Florentin Wörgötter, Alejandro Agostini, Norbert Krüger, Natalya Shylo, and Bernd Porr. Cognitive agents—a procedural perspective relying on the predictability of object-action-complexes (oacs). *Robotics and Autonomous Systems*, 57(4):420–432, 2009. 2

[58] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Trans. on PAMI*, 41(9):2251–2265, 2018. 3

[59] Zihui Xue, Kumar Ashutosh, and Kristen Grauman. Learning object state changes in videos: An open-world perspective. *arXiv preprint arXiv:2312.11782*, 2023. 6, 2, 3

[60] Zihui Xue et al. Learning object state changes in videos: An open-world perspective. *arXiv preprint arXiv:2312.11782*, 2024. 4

[61] Jiangwei Yu, Xiang Li, Xinran Zhao, Hongming Zhang, and Yu-Xiong Wang. Video state-changing object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20439–20448, 2023. 6, 2, 3

[62] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019. 2

[63] Ce Zhang, Changcheng Fu, Shijie Wang, Nakul Agarwal, Kwonjoon Lee, Chiho Choi, and Chen Sun. Object-centric video representation for long-term action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6751–6761, 2024. 4

[64] Zeyun Zhong, Manuel Martin, Michael Voit, Juergen Gall, and Jürgen Beyerer. A survey on deep learning techniques for action anticipation. *arXiv preprint arXiv:2309.17257*, 2023. 4

[65] X. Zhou, A. Arnab, C. Sun, and C. Schmid. How can objects help action recognition? In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2353–2362, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 2

[66] Yixin Zhu, Tao Gao, Lifeng Fan, Siyuan Huang, Mark Edmonds, Hangxin Liu, Feng Gao, Chi Zhang, Siyuan Qi, Ying Nian Wu, et al. Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense. *Engineering*, 6(3):310–345, 2020. 2

Anticipating Object State Changes in Long Procedural Videos

Supplementary Material

This supplementary material aims to provide a detailed analysis of the proposed *Ego4D-OSCA* dataset. In Section 7, we compare the proposed video dataset with existing image and video datasets proposed to support vision-based tasks related to object state change understanding. Section 8 provides an analysis of the complexity of the proposed dataset in the context of the newly introduced object state anticipation task. In that direction, we highlight its challenging nature that arises from the characteristics and underlying associations among the actions, the activities, the objects, and the annotated object states. Finally, in Section 9 additional sample images retrieved from the proposed dataset are shown to emphasize the complexity and context variability of object state change and action classes.

7. Comparison of datasets related to object state changes

The proposed *Ego4D-OSCA* dataset is compiled using a subset of the popular, large-scale Ego4D v2 dataset [19] comprising egocentric videos for a large variety of human daily living or work activities.

In Table 4 we compare the proposed *Ego4D-OSCA* dataset with existing image and video datasets that also provide annotations related to object states. The target tasks performed using each of the datasets are also noted. In the following, we analyze the most similar ones and argue on the necessity to compile a new dataset as a benchmark for the introduced task of Object State Anticipation in videos.

Ego4D [19]: The Ego4D dataset is among the largest to date, encompassing an extensive collection of videos captured in a wide variety of environments. A subset of this dataset has been utilized for object state detection and classification, as shown in the second-to-last row of Table 4. This subset consists of 92,864 short videos, each featuring a single state-modifying action.

The Ego4D dataset is related to three key challenges concerning visual object state change understanding. Firstly, the Ego4D SCOD (State Change Object Detection) challenge[1] focus on the bounding box-based detection of the object that undergoes a state change in an action segment (short video clip). The State Change Classification task is defined as the multi-class classification of object state changes in a video clip where a state-modifying action occurs, for example, identifying that the state of a cup has changed from filled to empty. Second, the binary state change classification variant is realized as an Ego4D challenge[2], with the aim to detect whether a state change was performed or not in an action segment (video clip). Moreover, the Ego4D State Change Localization challenge[3] involves pinpointing the exact frames in the video where the state change occurs. Accurate localization is crucial for understanding the precise timing and context of the state transitions within the egocentric video perspective. These challenges are designed to advance the understanding and development of AI models in recognizing and interpreting state changes in dynamic and realistic scenarios captured from a first-person viewpoint. Finally, a series of workshops in major conferences have been organized based on the Ego4D dataset and related tasks/benchmarks, such as 2nd International Ego4D Workshop @ ECCV 2022 , 1st Ego4D Workshop @ CVPR 2022 and the Joint 3rd Ego4D and 11th EPIC Workshop on Egocentric Vision @ CVPR2023.

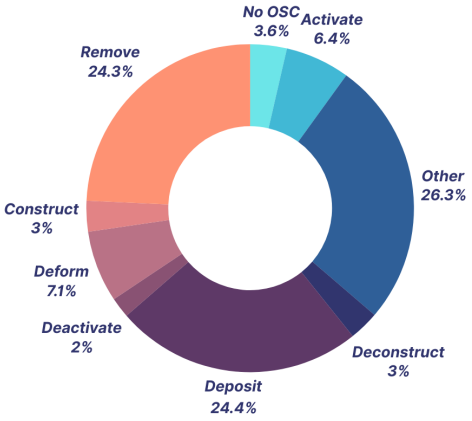

Figure 7. Statistics for the Ego4D-OSCA dataset per object state change class.

Comparison: The Ego4D-OSCA dataset comprises long videos of sequential state-modifying actions that correspond to any of the nine classes of object state changes: *deposit, remove, construct, deconstruct, activate, deactivate, deform, other, and no-state-change*. A distribution of the samples across the 9 object state labels is presented in Table 5 and Fig. 4. In contrast, the current subsets of Ego4D used for detecting and classifying object states (Ego4D SCOD & OSCC benchmark) comprise short videos, each depicting a single action. This renders the subsets unsuitable for addressing the problem of anticipating object state changes in procedural videos that comprise consecutive actions under the same scenario (activity).

[1] Ego4D State Change Object Detection Challenge

[2] Ego4D Object State Change Classification Challenge

[3] Ego4D Object State Change Temporal Localization Challenge

Datasets	Modalities	Task	Year	Actions per video	Samples	Obj. State Classes	Actions	Objects
Isola et al. [24]	Images	OS Classification	2015	N/A	63.440	9	-	18
OSDD [15]	Images	OS Classification & Detection	2021	N/A	19.000	9	-	18
Alayrac et al. [1]	Videos	OS Classification & Act. Localization	2017	Single	630	7	7	5
Fire et al. [13]	Videos	SC Object Detection	2017	Single	490	17	-	13
Task-Fluent [28]	Videos	OS Classification	2017	Single	809	21	14	25
ChangeIt [45]	Videos	OSC Temporal Localization	2022	Single	34.428	-	44	-
HowToChange [59]	Video & Text	OSC Temporal Localization	2023	Single	498.475	20	-	134
VSCOS [61]	Video	SC Object Segmentation	2023	Single	1.905	4	271	124
VOST [51]	Video	SC Object Segmentation	2023	Single	713	-	-	155
MOST [48]	Video	OS Classification	2024	Multiple	61	60	-	6
Ego4D [19]	Video	SC Object Detection & Classification	2022	Single	92.864	8	-	478
Ego4D-OSCA	Video	**OSC Anticipation**	2024	**Multiple**	1498	9	5754	475

Table 4. Comparison with other image and video datasets that contain annotations related to object state changes. Note that in *Ego4D-OSCA*, a sample refers to a video of an entire activity, which might consist of multiple actions.

	No OSC	activate	deactivate	construct	deconstruct	deposit	remove	deform	other
Train	2066	4017	1492	4186	1773	14984	15338	4400	15667
Test	1284	1888	617	2289	966	7613	7608	2149	8715

Table 5. Statistics for the Ego4D-OSCA dataset per object state change class. In total, the dataset has 61858 train and 31846 test clips.

ChangeIt [45]: The ChangeIt dataset comprises unedited videos sourced from YouTube and automatically generated labelling of actions. The designated tasks for analysis on this dataset involve the identification and temporal localization of the initial state, end state, and state-modifying action in a video. A set of 44 state-changing actions is provided, each demonstrated in approximately 15 videos on average. In total, there are $34,428$ videos with an average duration of 4.6 minutes.

Comparison: The newly introduced Ego4D-OSCA dataset comprises sequential videos featuring actions, some of which may involve state changes, while others may not. In contrast, the ChangeIt dataset consists of single-action clips, therefore one object state change is performed per clip. Ego4D-OSCA offers a wide range of scenarios without imposing any limitations and encompasses video durations spanning from minutes to hours. Conversely, the ChangeIt dataset confines scenarios to irreversible actions, aiming to eliminate instances where two actions return an object from an initial state to the same initial state via an intermediate state. Additionally, it excludes videos exceeding 15 minutes in length.

MOST [48]: The newest dataset related to object states in videos addresses the problem of temporal segmentation of multi-label object states. It includes manually collected instructional videos from YouTube, covering six object categories: apple, egg, flour, shirt, tire, and wire, each with around 10 annotated object states. These states represent common appearances or conditions an object may take. Annotators marked the time intervals when specific object states were visible, resulting in a dataset of 61 fully annotated videos with a total duration of 159.6 minutes. Unlike other datasets, such as ChangeIt, which focuses only on state transitions, MOST captures diverse object states, even if those are not tied to specific actions, offering a comprehensive benchmark for object state recognition.

Comparison: The MOST dataset is designed to assist the recognition of multiple object states for a *single object category per video*. In contrast, our proposed Ego4D-OSCA dataset focuses on anticipating the *state change class of multiple objects within each video*. This means that while MOST aims to recognize objects' current state, Ego4D-OSCA emphasizes predicting what an object's state change will be as a result/effect of the next (near future), yet unobserved, action. Additionally, Ego4D-OSCA covers a larger dataset with 1,498 videos and 478 object classes, compared to MOST's 61 videos and 6 object categories. We see potential in bridging this gap in future work by adding annotations for state change classes to the MOST dataset, which could open up new avenues for research in multi-label object state change anticipation. Additionally, contrary to the proposed Ego4D-OSCA dataset, the MOST dataset does not provide annotations regarding the actions and activities across the video. Such information, which our proposed methodology utilizes, can be crucial for understanding the progression of actions and their effects on objects enabling models to predict future states and state changes more accurately. Without this contextual information, it becomes significantly more challenging to infer the relationships between actions, object interactions, and subsequent state changes, limiting the ability to anticipate how an object might evolve within the dynamic environment of a pro-

cedural activity.

HowToChange [59]: This dataset is generated using a subset of The Food & Entertaining category of the HowTo100M dataset [32], a large-scale dataset of narrated videos. The reasons for selecting that particular subset are: (1) it constitutes one-third of the entire HowTo100M video collection, (2) cooking tasks within this category provide a rich variety of objects, tools, and state changes, making it an excellent testing ground for open-world Object-State-Change (OSC) understanding, and (3) in cooking activities, a single state transition can often be linked to a diverse array of objects, creating opportunities for compositional learning. The dataset provides annotation data related to three OS classes (initial, transitioning, and end state) related to OSC localization. A total of $498,475$ videos and $11,390,287$ ASR transcriptions processed with LLAMA2 reveal the most frequently observed state transitions and the associated objects. This information is utilized to establish an OSC vocabulary, identifying 134 objects, 20 state transitions, and 409 unique OSCs.

Comparison: The HowToChange dataset contains clips that involve a single state-changing action that is not compatible with the requirement for subsequent actions in a single video as in Ego4D-OSCA. On the other hand, it contains novel objects in the test set which sets the dataset a challenging benchmark for OSC analysis.

VSCOS [61]: This dataset comprises $1,905$ video clips of an average duration of 7.4 seconds, capturing various interactions with objects and state changes. These videos encompass 30 action categories and 124 object categories, resulting in 271 valid combinations in total. The state changes in the dataset can be categorized into four prominent groups: Rigid Object Composition and Decomposition (e.g., combine, cut, split, disintegrate, unpackage), Non-rigid Object Transformation (e.g., pour (liquid), crack (egg)), Object Appearance Change (e.g., cook, clean), and Object Articulation (e.g., open, close, twist).

Comparison: Each video in the VSCOS dataset contains a single state-changing action. The test set of this dataset is challenging because it encompasses the following cases: novel objects - seen state changes, seen objects - novel state changes, and novel objects - novel state changes.

VOST [51]: VOST is introduced as a benchmark for video object segmentation, emphasizing intricate object transformations. In contrast to current datasets, VOST introduces scenarios where objects undergo processes such as breaking, tearing, and molding, leading to substantial alterations in their overall appearance. Comprising over 700 high-resolution videos captured in diverse environments, each video in the dataset has an average duration of 21 seconds and is meticulously labelled with instance masks of objects across frames.

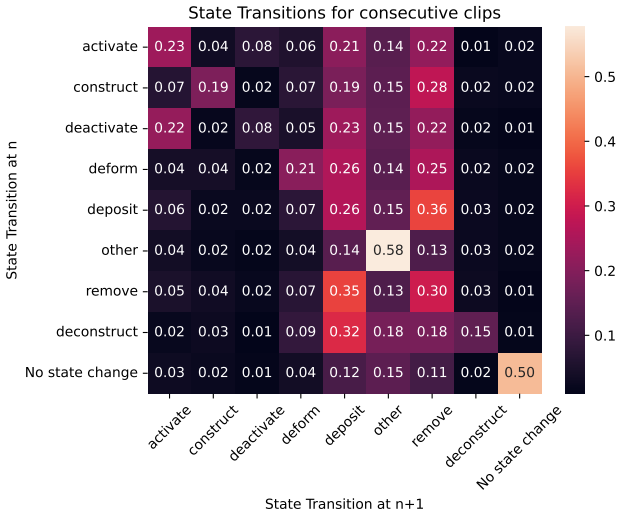

Figure 8. Transition frequencies for all pairs of subsequent object state changes in the Ego4D-OSCA dataset videos.

Comparison: The VOST dataset comprises videos depicting objects undergoing state-changing transformations. While a change in an object's state is demonstrated in each clip of the dataset, the state types are not explicitly labelled. According to the authors, the transformations are indicated by 51 specific verbs such as cut, peel, apply, break, open, scoop, fold, mold, etc. This dataset excludes videos without any transformation. In contrast, Ego4D-OSCA consists of sequential videos that exhibit a series of object state changes alongside videos lacking such changes. Ego4D-OSCA encompasses 117 verbs and 475 objects.

8. Ego4D-OSCA dataset statistics

We focus on the dynamic nature of object interactions in the dataset by extracting statistics for the combinations of object states in conjunction with action verbs and object classes. We investigate how different action verbs are associated with a variety of object states, highlighting the diversity and context-dependence of an action's effects. Additionally, we examine the variability of object states based on the actions performed, emphasizing the challenges imposed for the tasks of action and activity recognition and anticipation, and object state classification and anticipation. This analysis underscores the richness of the dataset and the sophisticated modelling required to accurately interpret and predict object states in various human activity contexts.

8.1. Action verbs vs object states

Each histogram illustrated in Fig. 11 demonstrates the distribution of the occurrences of action verbs in action segments of the proposed dataset associated with an object state

change class, with challenging long-tail distributions. The high variability of action verbs and the state change class associations are observed for the 'activate', 'deactivate', 'construct', 'deconstruct', 'deposit', and 'remove'.

Respectively, in Fig. 12, each histogram shows the (frequency) occurrences distribution of the instances of object classes in action segments associated with an object state change class of the proposed dataset.

In Fig. 13, the histogram provides the number of distinct object state change classes associated with various action verb classes. It can be verified that the majority of action classes are associated with at least three distinct object state classes. For example, actions involving the verb "open" (leftmost label on the x-axis) can lead to any of the state change classes depending on the object, e.g. the "activate" state change occurs when opening a microwave and "deposit" when opening a box. The observed diversity highlights the complexity and context-dependence of state-modifying actions in the dataset capturing a wide range of interactions and their state changes on interacting objects, which makes it valuable for training and testing sophisticated predictive models. This, in turn, indicates the need for elaborate learning models that can cope with the wide range of specific visual and semantic contexts in combination with different object classes involved in each action.

8.2. Objects vs object states

The histogram in Fig. 14, illustrates that certain objects in *Ego4D-OSCA* can appear in up to eight different states, depending on the action performed, which reveals the action-dependent variability of object states within the dataset. This observation underscores the complexity and dynamic nature of object interactions as well as the challenges to be tackled by solutions for classifying and predicting object state changes. The variability in these interactions presents significant challenges also for action recognition, where models need to accurately identify state changes and their transitions induced by subsequent actions, necessitating robust temporal models and comprehensive training data.

8.3. Variability in state changes & activity duration

The histogram in Fig. 15 illustrates the distribution of state transitions observed within the first 100 videos from the dataset. Each bar represents the frequency of specific state transitions between actions within a video, providing insights into the temporal dynamics and complexity of activities performed in the dataset.

The histogram highlights a significant variation in the number of state transitions observed within each video sample, indicating varying levels of complexity and duration in the actions performed. For instance, video 14 (video sample: *1e5bd816-e1dd-43d3-8709-42c83114dc7c*) stands out with 880 object state transitions and a duration of approximately 3 hours. This underscores the intricate nature of the actions captured in the original Ego4D dataset [19], where the extent of state transitions, as presented in the proposed variant (*Ego4D-OSCA*), reflects not only the complexity of the activities but also their temporal duration. Such variability emphasizes the need for comprehensive modelling approaches capable of accommodating diverse activity durations and complexities within the dataset. Moreover, this variability necessitates the consideration of a large action and state history by methods that utilize this information to predict future actions or object states, ensuring robustness and accuracy in forecasting.

Building on the significant diversity in the number of state transitions observed within each video, the transition matrix in Fig. 8 provides further insight into the probabilistic nature of these transitions. For instance, the data indicate that when an object is in the 'activate' state during action n, it frequently transitions to the 'deposit' or 'remove' states in action $n+1$. This suggests that actions involving the activation of objects, typically electrical appliances, are usually followed by actions that involve placing items into or removing items from these objects. Such logical transitions underscore the presence of temporal action ordering and causality in activities, where one action sets the stage for subsequent actions. This pattern highlights the importance of understanding state persistence and transitions, in developing predictive models. Accurately capturing these probabilistic dependencies is essential for models to effectively anticipate future states and actions. These insights reinforce the necessity for models to consider extensive action and state histories to accommodate the intricate and dynamic nature of the actions and activities in the Ego4D dataset, and inherently also in the proposed *Ego4D-OSCA*.

9. On the super-annotation of object state change classes

As stated in the main paper, the original Ego4D dataset does not provide specific state labels for individual video frames; instead, it offers annotations on state changes tied to entire video segments. These annotations include object bounding boxes and classes for seven key frames, centered around the moment of state change in each segment. Building on this, we augment critical frames with state-related labels. Specifically, for each segment, we label the initial frame as pre_X and the final frame as $post\_X$, where X represents the state change. To capture the semantics of these state transitions, we define three pairs of inverse state changes. Each pair reflects that one action reverses the outcome of the other. For instance, if X and Y are inverse changes, then pre_X and $post\_Y$ are considered equivalent, as are pre_Y and $post\_X$. A practical example of this is the pair pre_remove and $post\_deposit$, since "remove" and "deposit" are inverse actions. Table 6 outlines these super-annotated state labels

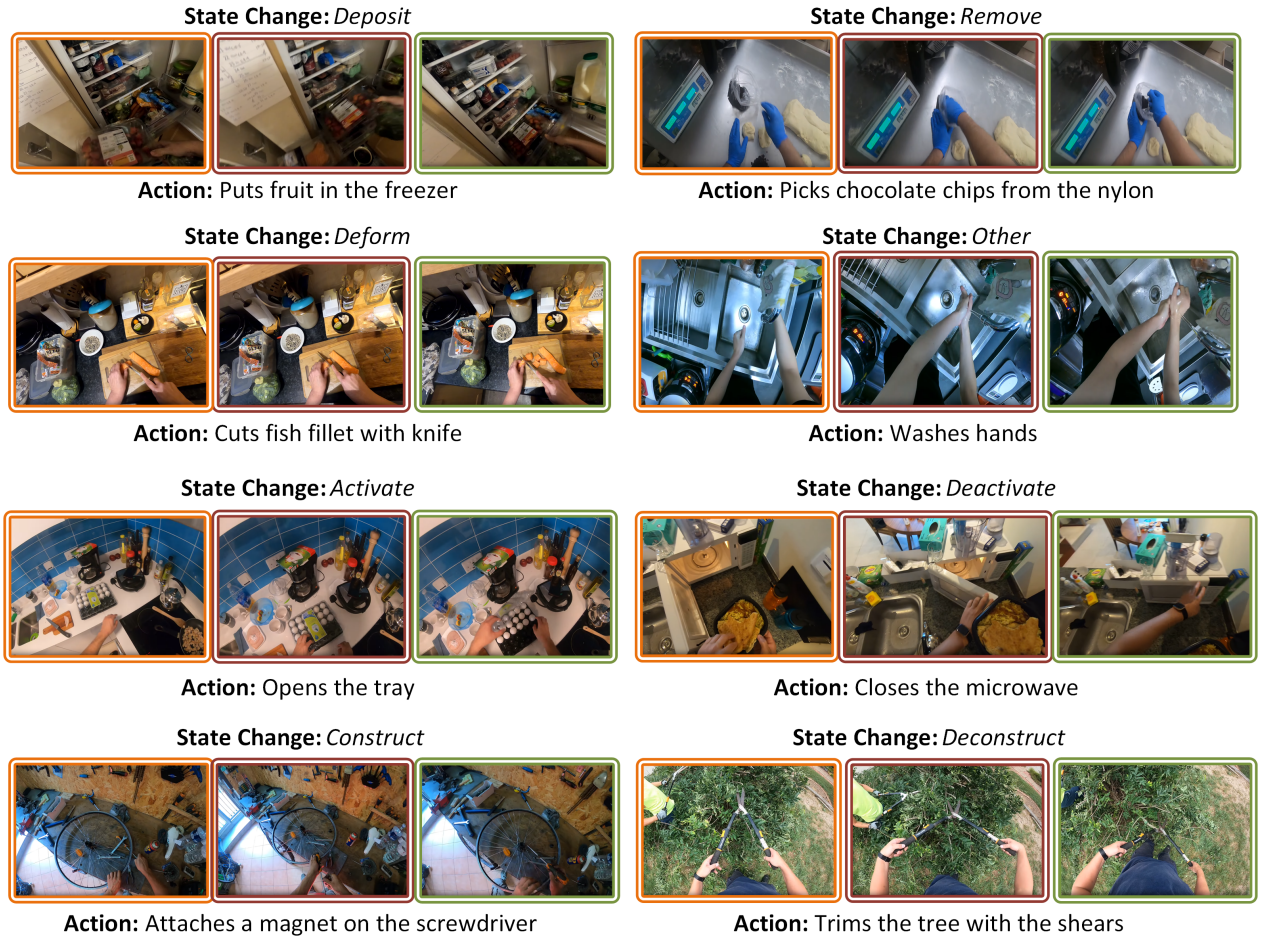

Figure 9. Sample frames of the Ego4D-OSCA dataset depict the initial state, the PNR frame, and the final state for the eight object state change classes (the class 'No object state change' is not included). The variability of visual environments/contexts, actions, and objects associated with the object state changes classes is highlighted.

OSC	activate	deactivate	deposit	remove	construct	deconstruct	deform	other
Pre	pre activate	pre deactivate	pre deposit	pre remove	pre construct	pre deconstruct	pre deform	pre other
Post	post activate	post deactivate	post deposit	post remove	post construct	post deconstruct	post deform	post other

Table 6. The super-annotated state change labels and the corresponding pre-/post-state labels of a video segment, where the state modifying action occurs. The pairs activate-deactivate, deposit-remove, and construct-deconstruct constitute pairs of inverse state change actions. Frame state labels that correspond to the same state are depicted with the same colour.

and their inverse association.

To emphasize the complexity and context variability of state change classes, Figure 9 shows sample images from the Ego-OSCA dataset depicting the object state change visual progression. In those few samples, one may notice the inverse association between different state change stages, as well as the large variability of visual environments and contexts, actions, and objects involved in different classes of object state changes. Finally, as also stated in lines $84-86$ and shown in Fig. 3 of the main paper, it is worth mentioning that motion motifs (as defined by verb primitives) do not necessarily have a one-to-one correspondence with states. As an example, in Fig. 10 we observe that the verb "close" can result to more than one object state change class. This

Figure 10. Sample frames from 3 instances of the "close" action, each involving different contexts and objects from the Ego4D-OSCA dataset, each resulting in various types of state changes.

highlights the fact that the object-related context is as important as the motion motif when defining an action as well as when estimating the anticipated object state change due to the execution of the action. Therefore to address the OSCA task an ideal method should build upon the past and current estimates of object detection and state estimation, as well as action recognition methods in order to robustly estimate the anticipated state of an object in procedural activities.

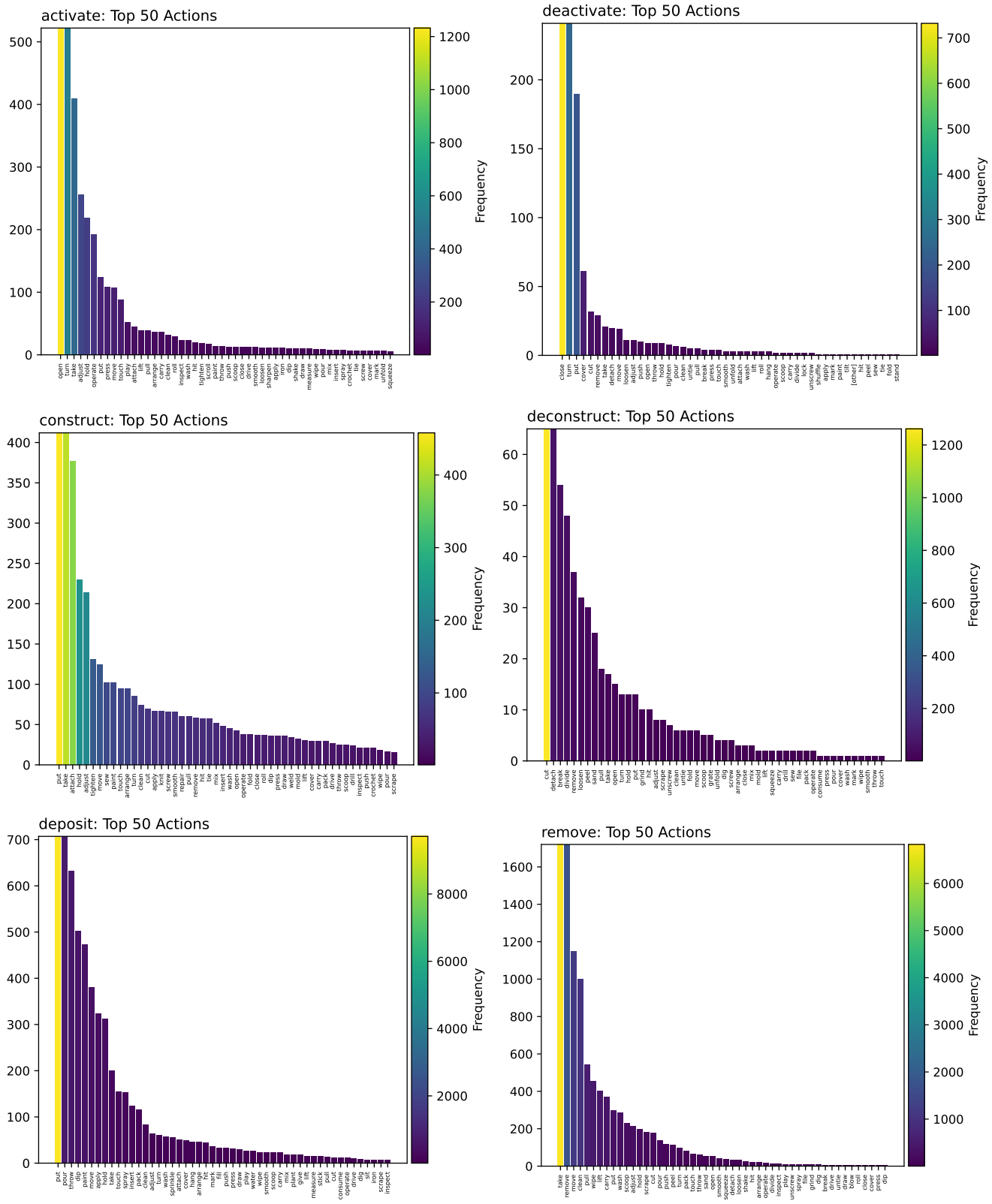

Figure 11. The frequency distribution of the top 50 actions (occurrences of action classes in the dataset action segments) concerning an object state change class is illustrated in each histogram for the classes ‘activate’, ‘deactivate’, ‘construct’, ‘deconstruct’, ‘deposit’, ‘remove’.

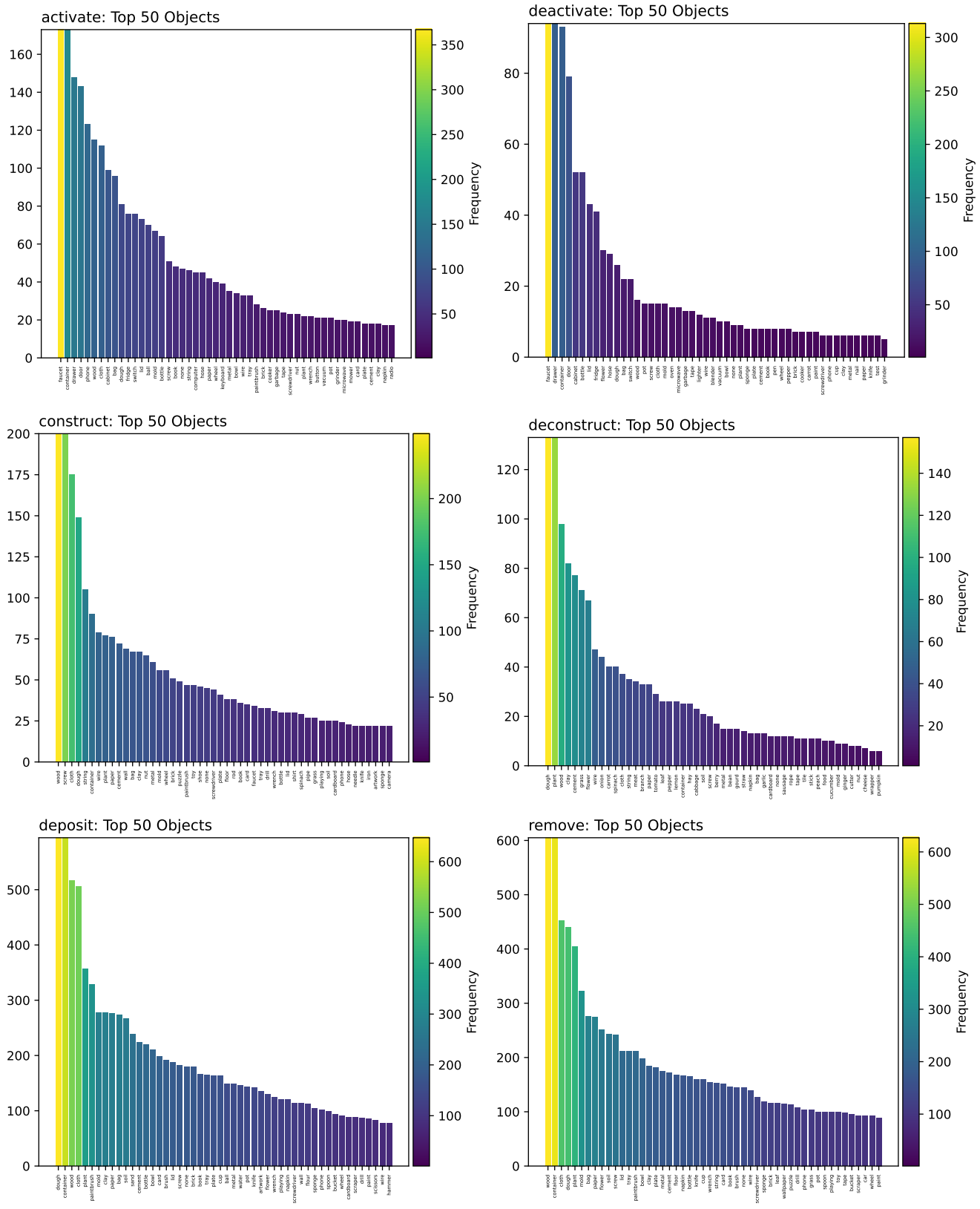

Figure 12. The frequency distribution of the top 50 objects (occurrences of object classes based on the dataset action segments) concerning an object state change class is illustrated in each histogram for the classes ‘activate’, ‘deactivate’, ‘construct’, ‘deconstruct’, ‘deposit’, ‘remove’.

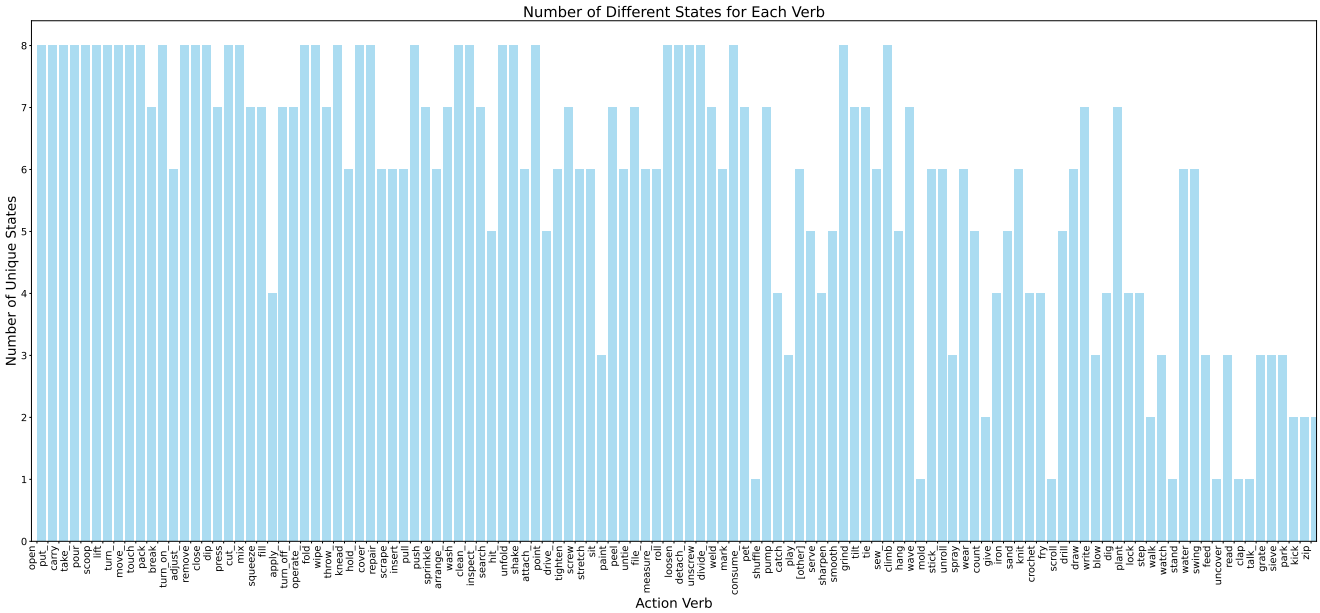

Figure 13. Histogram of object states associated with action verb classes.

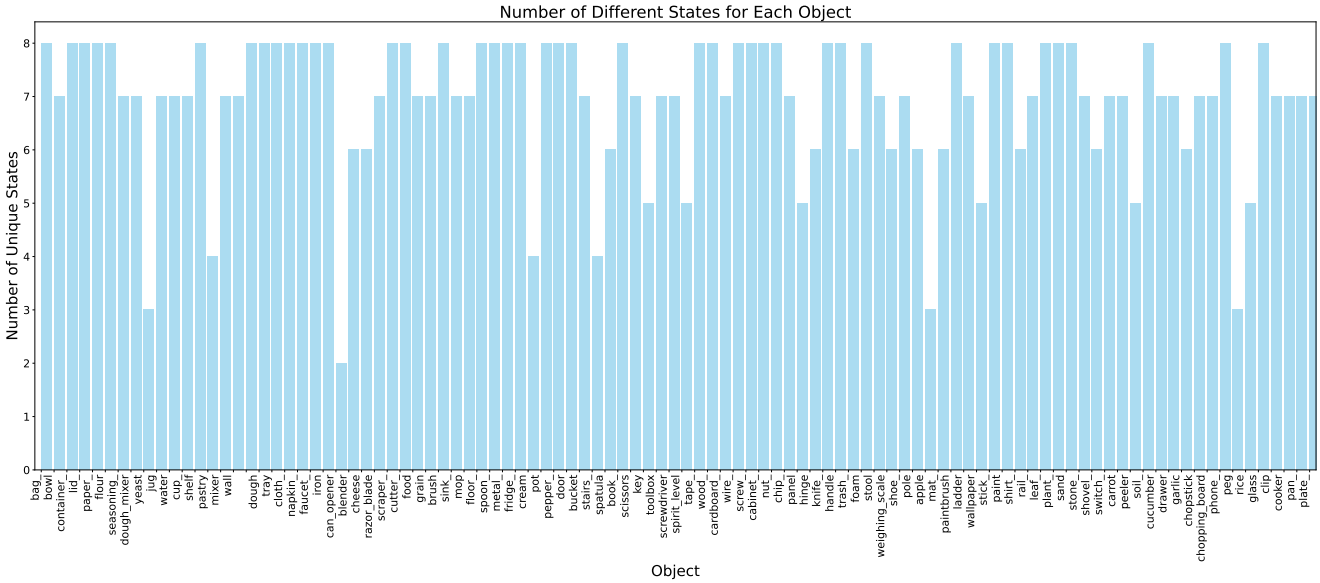

Figure 14. Histogram of object states associated with object classes. For better visualization purposes, we only depict the variability in the states of the first 100 objects.

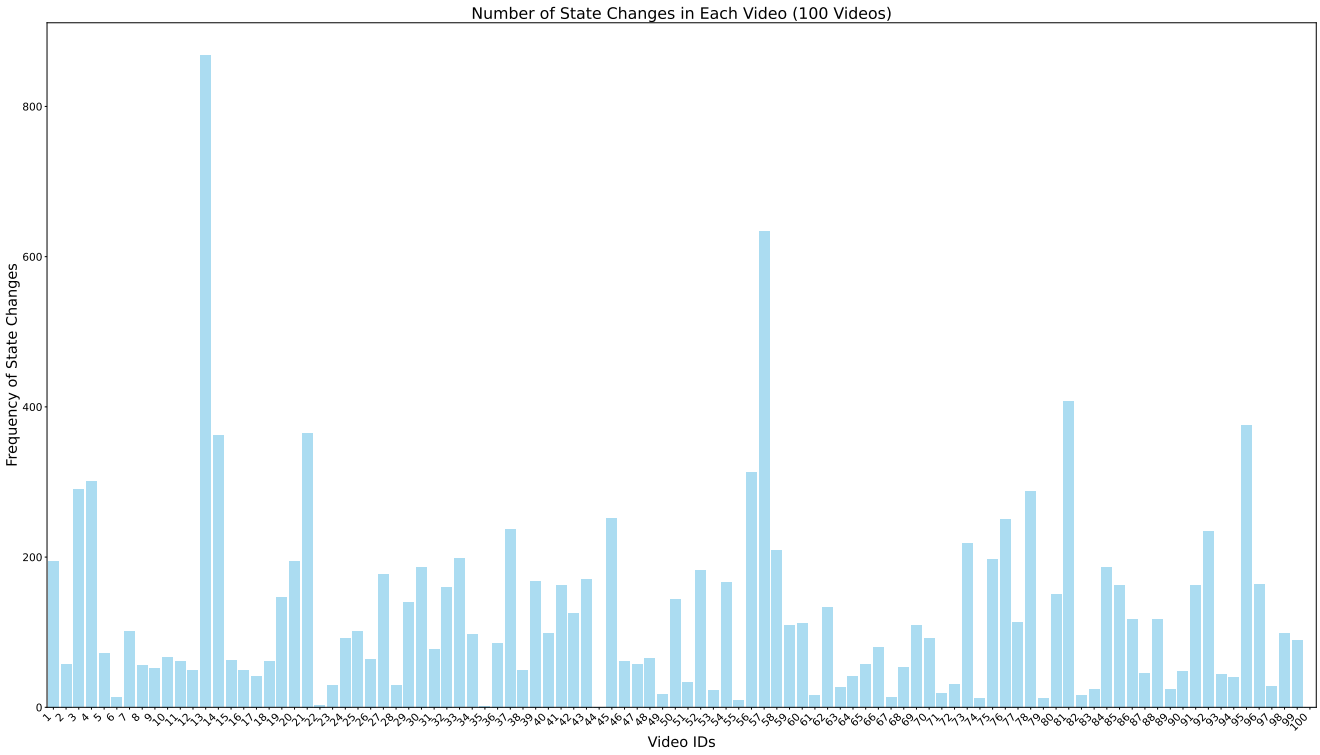

Figure 15. Histogram of the frequency of state transitions in the first 100 videos.