Inferring Past Human Actions in Homes with Abductive Reasoning

Clement Tan[1] Chai Kiat Yeo[1] Cheston Tan[2] Basura Fernando[1,2]
[1]Nanyang Technological University, Singapore
[2]Centre for Frontier AI Research (CFAR), Agency for Science, Technology and Research (A*STAR)
{s190099, asckyeo}@e.ntu.edu.sg, {cheston-tan@i2r, fernando_basura@ihpc}.a-star.edu.sg

Abstract

Abductive reasoning aims to make the most likely inference for a given set of incomplete observations. In this paper, we introduce "Abductive Past Action Inference", a novel research task aimed at identifying the past actions performed by individuals within homes to reach specific states captured in a single image, using abductive inference. The research explores three key abductive inference problems: past action set prediction, past action sequence prediction, and abductive past action verification. We introduce several models tailored for abductive past action inference, including a relational graph neural network, a relational bilinear pooling model, and a relational transformer model. Notably, the newly proposed object-relational bilinear graph encoder-decoder (BiGED) model emerges as the most effective among all methods evaluated, demonstrating good proficiency in handling the intricacies of the Action Genome dataset. The contributions of this research significantly advance the ability of deep learning models to reason about current scene evidence and make highly plausible inferences about past human actions. This advancement enables a deeper understanding of events and behaviors, which can enhance decision-making and improve system capabilities across various real-world applications such as Human-Robot Interaction and Elderly Care and Health Monitoring. Code and data available at https://github.com/LUNAProject22/AAR

1. Introduction

Reasoning is an inherent part of human intelligence as it allows us to draw conclusions and construct explanations from existing knowledge when dealing with an uncertain scenario. One of the reasoning abilities that humans possess is abductive reasoning. Abductive reasoning aims to infer the most compelling explanation for a given set of observed facts based on a logical theory. In this work, we study the new problem of inferring past human actions from visual information using abductive inference. It is an extremely useful tool in our daily life, as we often rely on a set of facts to form the most probable conclusion. In fact, a comprehensive understanding of a situation requires considering both past and future information. The ability to perform abductive reasoning about past human actions is vital for human-robot collaboration AI-assisted accident and crime investigation and assistive robotics. Furthermore, robots working in dynamic environments benefit from understanding previous human actions to better anticipate future behaviour or adapt their actions accordingly. Imagine the scenario where a rescue robot enters an elderly person's house to check on why he or she is not responding to the routine automated phone call. Upon entering the house, the robot observes its surroundings and notices that the *back door is left open* but *nothing else is out of the ordinary*. These observations may form a basis for a rational agent – the elderly might have *opened the door* and *went into the garden*. The robot can immediately make its way to search for him/her in the back garden. This example illustrates how a social rescue robot can utilize observed facts from the scene to infer past human actions, thereby reasoning about the individual's whereabouts and ensuring their safety through abductive reasoning.

In recent years, there have been some great initiatives made in abductive reasoning for computer vision [18, 24, 44]. In particular, [24] generates the description of the hypothesis and the premises in the natural language given a video snapshot of events. Without the generation of a hypothesis description, these methods boil down to dense video captioning. A similar task is also presented in [18] where given an image, the model must perform logical abduction to explain the observation in natural language. The use of natural language queries in these tasks presents challenges related to language understanding, making the abductive reasoning task more complicated.

In contrast to these recent works, we challenge a model to infer multiple human actions that may have occurred in the past from a given image. Based on the visual information from the image, objects such as a person, glass and cabinet, may provide clues from which humans can draw conclusions – see Fig. 1. We term this new task, *Abductive Past Action Inference* and further benchmark how deep learning models perform on this challenging new task. For this task, the models are not only required to decipher the effects of human actions resulting in different environment states but also solve causal chains of action effects, a task that can be challenging even for humans. Furthermore, the task relies on the model's ability to perform abductive reasoning based on factual evidence i.e., determining what actions may have or have not been performed based on visual information in the image. Humans can solve this task by using prior experience (knowledge) about actions and their effects and using reasoning to make logical inferences. Are deep learning models able to perform abductive past action inference by utilizing visual knowledge present in a given image and a learned understanding of the domain? We aim to answer this question in this paper.

Human action can be viewed as an evolution of human-object relationships over time. Therefore, the state of human-object relations in a scene may give away some cues about the actions that may have been executed by the human. We hypothesize that deep learning models are able to perform logical abduction on past actions using visual and semantic information of human-object relations in the scene. As these human-object relations provide substantial visual cues about actions and the effects of past actions, it makes it easier for the models to infer past actions. On the other hand, there is also the duality in which the evidence should support those conclusions (the actions inferred by the model). If a human executed a set of actions $\mathcal{A}$ which resulted in a state whereby a human-object relation set $\mathcal{R}$ is formed as an effect of those executed actions (i.e., $\mathcal{A} \rightarrow \mathcal{R}$), then using the relational information, we can formulate the task by aiming to infer $\mathcal{A}$ from $\mathcal{R}$. Therefore, we argue that human-object relational representations are vital for abductive past action inference and provide further justifications in our experiments.

In this work, our models rely on the human-centric object relation tuples such as (person, glass) and (person, closet) obtainable from a single image at the current point in time to perform abductive past action inference. One can see why these human-centric relations are vital for identifying past actions: the (person, glass) relation may lead to deriving actions such as (person-pouring-water, person-took-glass-from-somewhere) while (person, closet) may imply actions such as (person-opening-closet, person-closing-closet) see – Figure 1. Therefore, we use objects and their relationships in the scene to construct human-centric object relations within each image. These relations are made up of both visual and semantic features of recognized objects. To effectively model relational information, we use bilinear pooling and to model inter-relational reasoning, we use a new relational graph neural network (GNN). We propose a new model called BiGED that uses both bilinear pooling and GNNs to effectively reason over human-object relations and inter-relational information of an image to perform abductive inference on past actions effectively.

Our contributions are summarized as follows: (1) To the best of our knowledge, we are the first to propose the abductive past action inference task, which involves predicting past actions through abductive inference. (2) We benchmark several image, video, vision-language, and object-relational models on this problem, thereby illustrating the importance of human-object relational representations. Additionally, we develop a new relational rule-based inference model which serve as relevant baseline models for the abductive past action inference task. (3) We propose a novel relational bilinear graph dncoder-decoder model (BiGED) to tackle this challenging reasoning problem and show the effectiveness of this new design.

2. Related Work

Our work is different from action recognition [17,19] in a fundamental way. First, in action recognition, the objective is to identify the actions executed in the visible data (e.g., a video or an image in still image action recognition [14]). In action recognition, the models can learn from visual cues what the action looks like and what constitutes an action. In our work, we aim to infer past actions that the model has never seen the human performing. The model only sees visual evidence (e.g. human-object relations) in the scene which is the outcome of executed actions. There are no motion cues or visual patterns of actions that the model can rely on to predict past actions. From a single static image, the machine should infer what actions may have been executed. This is drastically different from classical action recognition and action anticipation tasks.

Abductive past action inference shares some similarity to short-term action anticipation [8, 10] and long-term action anticipation [1]. However, there are several notable differences between the two tasks. Firstly, in abductive past action inference, the goal of the model is to identify the most plausible actions executed by a human in the past based on the current evidence, whereas, in action anticipation, the model learns to predict future action sequences from current observations. The primary distinction lies in abductive past action inference, where observations (evidence) may imply certain past actions, contrasting with action anticipation tasks that predict future actions without certainty. In other words, in abductive past action inference, the evidence and clues indicate possible actions executed in the past that resulted in the evidence or clues. However, in action anticipation, the clues [33], context [13], current actions [12], and knowledge about the task [45] are used to infer probable future actions, but there is no guarantee that the predicted actions will be executed by a human. For instance, observing *a person cleaning a room with a broom* suggests prior actions such as *picking up the broom from somewhere* must have happened among many others. Even if *putting away the broom* is anticipated somewhere in the future, other actions such as *holding the broom* and *opening a window* are also possible. Therefore, while action anticipation addresses the uncertainty of future human behavior, abductive past action inference models can utilize scene evidence (such as objects in the scene) to infer the most likely past actions. Additionally, in abductive past action inference, the uncertainty arises from the fact that several different actions may have resulted in similar states $\mathcal{R}$. In our task, models should comprehend the consequences of each executed action and engage in abductive reasoning to infer the most probable set or sequence of past actions. Another key difference between action anticipation and abductive past action inference is that in action anticipation, predictions made at time t can leverage all past observations. In contrast, abductive past action inference relies solely on present and future information, where new future observations can potentially alter the evidence about past actions, making the inference process more challenging.

Visual Commonsense Reasoning (VCR) [41, 43] and causal video understanding [28, 29] are also related to our work. In VCR [43], given an image, object regions, and a question, the model should answer the question regarding what is happening in the given frame. The model has to also provide justifications for the selected answer in relation to the question. Authors in [27] also studied a similar problem where a dynamic story underlying the input image is generated using commonsense reasoning. In particular, VisualCOMET [27] extends VCR and attempts to generate a set of textual descriptions of events at present, a set of commonsense inferences on events before, a set of commonsense inferences on events after, and a set of commonsense inferences on people's intents at present. In this vein, given the complete visual commonsense graph representing an image, they propose two tasks; (1) generate the rest of the visual commonsense graph that is connected to the current event and (2) generate a complete set of commonsense inferences. In contrast, given an image without any other natural language queries, we recognize visual objects in the scene and how they are related to the human, then use the human-centric object relational representation to infer the most likely actions executed by the human.

Recently, there are machine learning models that can also perform logical reasoning [3,5,15,21,46]. Visual scene graph generation [38] and spatial-temporal scene graph generation [4] are also related to our work. Graph neural networks are also related to our work [35,36,42]. Our work is also related to bilinear pooling methods such as [9, 11].

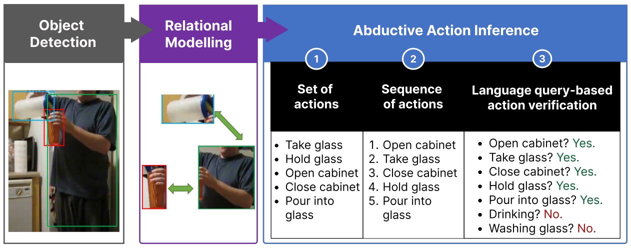

Figure 1. Proposed object-relational approach for abductive past action inference. Models are tasked to: 1) abduct the set of past actions, 2) abduct the sequence of past actions, and 3) perform abductive past action verification.

3. Abductive Past Action Inference

Task: Given a single image, models have to infer past actions executed by humans up to the current moment in time. We name this task Abductive Past Action Inference. Let us denote a human action by $a_i \in A$ where A is the set of all actions and $E_1, E_2, \cdots$ is a collection of evidence from the evidence set $\mathcal{E}$. As the evidence is a result of actions, we can write the logical implication $\mathcal{A} \rightarrow \mathcal{E}$ where $\mathcal{A}$ is the set of actions executed by a human which resulted in a set of evidence $\mathcal{E}$. Then, the task aims to derive 1) the set of past actions, 2) the sequence of past actions that resulted in the current evidence shown in the image, and 3) abductive past action verification. The abductive past action verification is a binary task where the model is given a single image and is required to answer a yes or no to an action query (did the person execute action a_x in the past?).

3.1. Object-Relational Representation Approach

Our primary hypothesis is that human-object relations are essential for abductive past action inference. Therefore, we propose a human-object relational approach for the task. In all three tasks, our general approach is as follows. We make use of detected humans and objects in the image and then generate a representation for human-centric object relations. Then, using these human-centric object relations, we summarize the visual image, and using neural models, we infer the most likely actions executed by the human. The overview of this approach is shown in Figure 1. Next, we first discuss abductive past action set inference, followed by the details of abductive past action verification.

Abductive past action set prediction. Let us denote the object by $o \in O$, the predicate category by $p \in P$, and the human by h. The j^{th} relation R_j is a triplet of the form $\langle h, p, o \rangle$. In the i^{th} image, we observe n number of relations $\mathcal{R}_i = \{R_1, R_2, \cdots R_n\}$ where $\mathcal{R}_i$ is the relation set present in the situation shown in that image. These relations constitute the evidence ($\mathcal{E}$). The relation set $\mathcal{R}_i$ is an effect of a person executing an action set/sequence $\mathcal{A}_i = \{a_1, \cdots a_K\}$.

Therefore, the following association holds.

$$\mathcal{A}_i \rightarrow \mathcal{R}_i \quad (1)$$

However, we do not know which action caused which relation (evidence), as this information is not available. The association reveals that there is a lack of specific knowledge about the exact effects of individual actions and when multiple actions have been executed, the resulting effect is a combined effect of all executed actions. Consequently, the learning mechanism must uncover the probable cause-and-effect relationships concerning actions. Therefore, given $\mathcal{R}$ we aim to perform abductive past action inference to infer the most likely set of actions executed by the human using neural network learning.

We learn this abduction using the deep neural network functions $\phi()$, and $\phi_c()$. The relational model, ϕ takes relation set as input and outputs a summary of relational information as a vector x_r.

$$x_r = \phi(R_1, \cdots, R_n; \theta_\phi) \quad (2)$$

The parameters of the relational model are denoted by θ_ϕ. The linear classifier (ϕ_c) having the parameters θ_c, takes relational information as a vector x_r as input and returns the conditional probability of actions given the relational evidence as follows:

$$P(a_1, \cdots, a_K | R_1, \cdots, R_n) = \phi_c(x_r; \theta_c) \quad (3)$$

The training and inference sets comprise images and corresponding action set $\mathcal{A}_i$. From each image, we extract the relation set $\mathcal{R}_i$. Therefore, the dataset consists of $\mathcal{D} = \bigcup_i \{\mathcal{R}_i, \mathcal{A}_i\}$. Given the training set ($\mathcal{D}$), we learn the model function in Equations 2 and 3 using backpropagation as follows:

$$\theta_\phi^*, \theta_c^* = \texttt{argmin}_{\theta_\phi, \theta_c} \sum_i -log(P(\mathcal{A}_i | \mathcal{R}_i)) \quad (4)$$

where $\theta_\phi^*, \theta_c^*$ are the optimal parameters. As this is a multi-label-multi-class classification problem, we utilize the max-margin multi-label loss from PyTorch where the margin is set to 1, during training.

Abductive past action verification. Abductive verification model $\phi_{ver}()$ takes the evidence $\mathcal{E}$ and the semantic representation of the past action (e.g. textual encoding of the action name) y_a as inputs and outputs a binary classification score indicating if the evidence supports the action or not, i.e. $\phi_{ver}(\mathcal{E}, y_a) \rightarrow [0, 1]$. Specifically, we encode the past action name using the CLIP [31] text encoder to obtain the textual encoding y_a for action class a. Then, we concatenate y_a with x_r and utilize a two-layer MLP to perform binary classification to determine whether action a was executed or not. We use the max-margin loss to train $\phi_{ver}()$. *Note that semantic embedding of action classes (y_a) is not a necessity here. For example, one might learn the class embeddings from scratch* removing the dependency on language or use one-hot-vectors.

3.2. Relational Representation

To obtain the relation representation, we extract features from the human and object regions of each image using a FasterRCNN [32] model with a ResNet101 backbone [16]. Let us denote the human feature by x_h, the object feature by x_o, and the features extracted from taking the union region of both human and object features by x_u. As we do not know the predicate or the relationship label for the relation between x_h and x_o, we use the concatenation of all three visual features x_h, x_o, and x_u as the joint relational visual feature $x_v = [x_h, x_o, x_u]$. Using FasterRCNN, we can also obtain the object and human categories. We use Glove [30] embedding to acquire a semantic representation of each human and object in the image. Let us denote the Glove embedding of the human by y_h and the object by y_o. Then, the semantic representation of the relation is given by $y_s = [y_h, y_o]$. Using both visual and semantic representations, we obtain a joint representation for each human-centric object relation in a given image. Therefore, the default relation representation for a relation $R =< h, p, o >$ is given by $r = [x_v, y_s]$. Note that we do not have access to the predicate class or any information about the predicate. Next, we present several neural and non-neural models that we developed in this paper that uses relational representations for the abductive past action inference task.

The details of abductive past action sequence inference are provided in the supplementary materials (section 4.1). Next, we present our graph neural network model to infer past actions based on relational information.

3.3. GNNED: Relational Graph Neural Network

The graph neural network-based encoder-decoder model summarizes relational information for abductive past action inference. Given the relational data with slight notation abuse, let us denote the relational representations by a $n \times d$ matrix $\mathcal{R} = [r_1, r_2, \cdots, r_n]$, where r_n has d dimensions. In our graph neural network encoder-decoder (GNNED) model, we first project the relational data using a linear function as follows:

$$\mathcal{R}' = \mathcal{R}W_l + b_l \quad (5)$$

where $\mathcal{R}' = [r'_1, r'_2, \cdots, r'_n]$. Then, we construct the affinity matrix $W_A(i, j)$ using Jaccard Vector similarity, where $W_A(i, j)$ shows the similarity/affinity between the i-th relation and the j-th relation in the set. Here, we use Jaccard Vector Similarity which is a smooth and fully differentiable affinity [8]. Note that Jaccard Vector Similarity is bounded

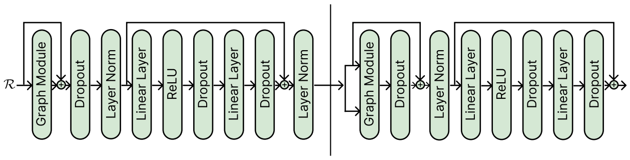

Figure 2. The graph neural network encoder (left) and graph neural network decoder (right) architecture. The residual connections are shown with the + sign.

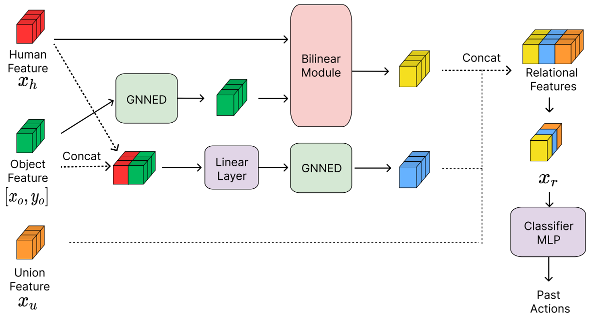

Figure 3. The Bilinear Graph Encoder-Decoder (BiGED) architecture.

by [-1,1]. Thereafter, we obtain the graph-encoded relational representation as follows:

$$G_e = ReLU((W_A\mathcal{R}')W_g + b_g) \tag{6}$$

where W_g and b_g are the weight matrix and bias term respectively. We call equations 5-6 the graph module. Using the graph module as a base model, we develop a graph encoding layer. The relational graph neural network encoder-decoder architecture we proposed is shown in Figure 2. Our graph encoder-decoder model consists of one graph encoder and three graph decoders by default. The graph module is able to model the inter-relations between human-object relations and reasons about them to perform abduction on past actions. The graph encoding layer (left) is very similar to the Transformer encoder layer [39]. The graph encoding layer consists of drop-out layers, layer norm [2], linear layers, and residual connections [16]. The graph decoder layer (right) is also similar to a Transformer decoder layer except for the graph module. Finally, we apply max-pooling at the end of the graph encoder-decoder model to obtain the final image representation x_r.

3.4. RBP: Relational Bilinear Pooling

To effectively model the higher-order relational information between human and object features, we use bilinear pooling. Given the human representation x_h and the object representation x_o, we use bilinear pooling with a weight matrix W_b of size $d \times d \times d$ and linear projection matrices W_{bl}, W_{jb} as follows:

$$o' = ReLU(W_o x_o + b_o) \tag{7}$$
$$h' = ReLU(W_h x_h + b_h) \tag{8}$$
$$r_b = ReLU([h'W_b o'; ([h'; o']W_{bl} + b_{bl})])W_{jb} + b_{jb} \tag{9}$$

where [;] represents the vector concatenation and $h'W_b o'$ is the bilinear pooling operator applied over human and object features. $([h'; o']W_{bl} + b_{bl})$ is the output of concatenated human and object features followed by a linear projection using weight matrix W_{bl} and bias term b_{bl}. In contrast to bilinear pooling, the concatenated linear projection captures direct relational information between human and object features. Then, we concatenate the bilinear pooled vector $(h'W_b o')$ and the output of the linear projection $(([h'; o']W_{bl} + b_{bl}))$. Next, we use ReLU and apply another linear projection $(W_{jb} + b_{jb})$. Finally, we concatenate the overlap feature x_u with the overall model output (r_b) and apply max-pooling across all relational features $([r_b; x_u])$ in the image to obtain x_r.

3.5. BiGED: Bilinear Graph Encoder-Decoder

Finally, to take advantage of both bilinear relational modeling and graph neural network encoder-decoder models, we combine both strategies as shown in Fig 3. The main idea is to replace the projection function in Equation 7 with a graph neural network encoder-decoder model. Let us denote the graph neural network encoder-decoder model by $f_{Ged}()$. Then, equation 7 will be replaced as follows:

$$O' = f_{Ged}(X_o) \tag{10}$$

where X_o is all the object features in the image. Afterward, we apply equation 9 before using bilinear modeling to obtain the relational representation. Note that as there are only one or two humans in the image, we do not use the GNNED to model human semantic features. The inputs to the BiGED model are the visual human features x_h, concatenated visual and semantic object features $[x_o, y_o]$ as well as the union features x_u. Next, we concatenate the human and object features $[x_h, x_o, y_o]$ to obtain a joint feature and then pass it through a linear layer and another GNNED model. The outputs of the bilinear, joint feature-based GNNED models and overlap union feature x_u are concatenated to obtain the final relational representation. Afterward, we use max-pooling to obtain the representation x_r for the image. For all models, we employ a linear classifier to infer past actions using the representation vector x_r.

4. Experiments and Results

4.1. Action Genome Past Action Inference dataset

We extend the Action Genome (AG) dataset [20] and benchmark all models on the AG dataset for the abductive past action inference task. Built upon the Charades dataset [37], the AG dataset contains 9,848 videos with 476,000+ object bounding boxes and 1.72 million visual relationships annotated across 234,000+ frames. *It should be noted that not all video frames in the Charades dataset are used in the AG dataset. Only a handful of keyframes are used in AG, and we follow the same.* The AG dataset does not provide action annotations. To obtain action annotations for images of AG, we leverage the Charades dataset which contains 157 action classes. The process of generating action sets and sequences using images from the Action Genome and action labels from the Charades dataset for the abductive past action inference task is detailed in Section 2 of the supplementary materials.

4.2. Experimental Setup

After obtaining the action annotations of images for a given video, we drop videos having only one image as there are no past images and therefore, no past actions. For the remaining images, we assign action labels from the previous images in two different evaluation setups:

1. Abduct at T: Given a image at time T, we add action labels from all the previous images to the ground truth (including actions from the current image) where $\mathcal{A}_t$ denotes all past actions of the t^{th} image. Therefore, the ground truth action set $\mathcal{A}$ is given by $\mathcal{A} = \overset{T}{\underset{t=1}{\cup}} \mathcal{A}_t$.

2. Abduct last image: Based on the first setup, we add an additional task where the model has to perform inference only on the last image of each video which contains all past actions. If the last image is T', then the action set is $\mathcal{A} = \overset{T'}{\underset{t=1}{\cup}} \mathcal{A}_t$. Note that in the Action Genome dataset, the images are sampled non-homogeneously from the Charades dataset videos. Therefore, the previous image occurs several seconds before the current image. In our abductive past action inference task, the ground truth past action sets are confined to the length of each video. We provide details on the number of images for a set of n past actions in the AG dataset for these setups in the supplementary materials – section 2 and figure 4.

4.3. Evaluation Metrics

We utilize the mean Average Precision (mAP), Recall@K (R@K), and mean Recall@K (mR@K) metrics to evaluate the models for the abductive action set prediction and action verification tasks. Each image contains 8.9 and 8.2 actions for the ***Abduct at*** T and ***Abduct last image*** setups respectively. Therefore, K is set to 10 based on the average number of actions contained in a image. Please see the supplementary material Section 3 for more implementation details. We will also release all our codes and models for future research.

4.4. Baseline Models

We benchmark several publicly available image (Resnet101-2D [16], ViT [6]) and video models (Slow-Fast [7] and Resnet50-3D) using the surrounding 8 frames of a image from the Charades dataset, and Video-Swin-S [26], Mvitv2 [23] and InternVideo [40] models using future K images from Action Genome to explore video based methods. The value of K is set to the minimum possible frame size for each model, with the default being 5 frames. Image models are pre-trained on ImageNet [34] while video models are pre-trained on Kinetics400 [22] dataset and we fine-tune these models on our task. We use a batch size of 32 with a learning rate of 1e-5. As for ViT, we use a batch size of 2042 using A100 GPUs. All video-based methods are fine-tuned end-to-end. We also use CLIP linear-probe and zero-shot to perform abduction using several variants of the CLIP model [31]. The details of all other baseline models (Relational Rule-based inference, Relational MLP, and Relational Transformers) are presented in supplementary material Section 1.

4.5. Human Performance Evaluation

Human performance for the abductive past action set inference and verification tasks in the *Abduct at* T setup is presented in Tables 1 and 3. Performance on the abductive past action sequence inference is provided in the supplementary materials–see Table 1 and Section 4.1. All human experiments for the three sub-problems in the *Abduct at* T setup follow the same procedure. Evaluators are asked to review 100 randomly sampled test images and manually assess all action classes in the Charades dataset without viewing the ground truth. They then select the likely past actions for each image.

4.6. Results: Abductive Past Action Set Prediction

Our results for the abductive past action inference set prediction task are shown in Table 1. These results are obtained based on the *Abduct at* T setup. During training, the model learns from every single image in the video sequence independently. Likewise, during inference, the model predicts the past action set on every single image. The end-to-end trained models such as Slow-Fast [7], ResNet50-3D, Resnet101-2D, and ViT perform poorly as it may be harder for these models to find clues that are needed for the abductive inference task. As there are no direct visual cues to infer previous actions (unlike object recognition or action recognition) from a given image, end-to-end learning becomes harder or near impossible for these models. The Video-Swin-S Transformer model [26] shows promise in end-to-end models due to its use of future context (K future snapshots) and strong video representation capabilities.

On the other hand, multi-modal foundational models such as the CLIP [31] variants are able to obtain better

Model	**mAP**	**R@10**	**mR@10**
Human Performance	–	80.60	82.81
End-to-end training			
ResNet101-2D [16]	9.27	18.63	11.51
ViT B/32 [6]	7.27	16.84	8.82
Resnet50-3D [7]	8.16	16.08	7.83
Slow-Fast [7]	7.91	14.42	7.65
Video-Swin-S [26] - (K= 5)	14.86	34.18	19.05
MvitV2 [23] - (K=16)	14.01	34.38	15.17
InternVideo [40] - (K=8)	12.29	30.72	12.37
Vision-language models			
CLIP-ViT-B/32 (zero-shot) [31]	14.07	14.88	20.88
CLIP-ViT-L/14 (zero-shot) [31]	19.79	21.88	27.77
CLIP-ViT-B/32 (linear probe) [31]	16.16	31.25	16.38
CLIP-ViT-L/14 (linear probe) [31]	22.06	40.18	20.01
Object-relational methods - using GT human/objects			
Relational Rule-based inference	26.27	48.94	36.89
Relational MLP	27.73±0.20	42.50±0.68	25.80±0.61
Relational Self Att. Transformer	33.59±0.17	56.03±0.40	40.04±1.15
Relational Cross Att. Transformer	34.73±0.05	56.89±0.47	40.75±0.57
Relational GNNED	34.38±0.36	57.17±0.35	42.83±0.21
Relational Bilinear Pooling (RBP)	35.55±0.30	59.98±0.68	43.53±0.63
BiGED	**35.75±0.15**	**60.55±0.41**	**44.37±0.21**
BiGED - (K=3)	36.00 ± 0.12	60.17±0.44	42.82± 0.90
BiGED - (K=5)	**37.34 ± 0.21**	**61.16± 0.56**	44.07 ± 0.87
BiGED - (K=7)	36.57 ± 0.38	60.65 ± 0.52	43.12 ± 0.47
Object-relational method - using FasterRCNN labels			
BiGED	**24.13± 0.04**	**43.59± 0.88**	**30.12± 0.23**

Table 1. Abductive past action set inference performance using the proposed methods on the *Abduct at T* setup.

Model	**mAP**	**R@10**	**mR@10**
Relational Rule-based inference	26.18	44.34	33.94
Relational MLP	25.99±0.10	38.79±0.86	23.54±0.72
Relational Self Att. Transformer	30.13±0.11	47.55±0.14	35.05±0.55
Relational Cross Att. Transformer	31.07±0.20	48.33±0.15	35.32±0.50
Relational GNNED	30.95±0.30	48.18±0.17	36.36±0.12
RBP	31.48±0.20	**49.79±0.55**	**36.96±0.36**
BiGED	**31.41±0.15**	49.62±0.64	36.15±0.61
Object-relational method - using FasterRCNN labels			
BiGED	**22.01±0.26**	**37.06±0.52**	**25.01±0.37**

Table 2. Abductive past action set inference performance using the proposed methods on the *Abduct last image* setup.

results than vanilla CNN models on this task perhaps due to the quality of the visual representation. Interestingly, object-relational models such as MLP and rule-based inference obtain decent performance. One might argue that the performance of human-object relational models is attributed to the use of ground truth object labels in the scene. However, when we tried to incorporate ground truth objects using object bounding boxes with red colored boxes as visual prompts in the CLIP [31] model, the performance was poor. The poor performance of the CLIP might be attributed to their training approach, which aims to align overall image features with corresponding text features. During their training, CLIP assumes that the text in the captions accurately describes the visual content of the image. However, when it comes to abductive past action inference, no explicit visual cues are available to indicate the execution of certain actions. We also note that the CLIP model demonstrates reasonable zero-shot performance. This may be because the CLIP model learns better vision features.

We experimented with generative models like ViLA [25] (instruction tuning) and BLIP (question answering). Details are in the supplementary materials sections 1.4 for GPT-3.5 and section 1.5 for ViLA. After instruction tuning on our dataset, ViLA achieved 10.5 mAP, 29.3 R@10, and 19.8 mR@10. We also tested GPT-3.5 with human-object relations as context, yielding 9.98 mAP, 25.22 R@10, and 20.32 mR@10. Due to the challenges of unconstrained text generation, models like BLIP, ViLA, and GPT-3.5 are excluded from the main comparison table.

The results also suggest that the human-object relational representations provide valuable evidence (cues) about what actions may have been executed in contrast to holistic vision representations. Among object-relational models, the MLP model and rule-based inference perform the worst across all three metrics. The rule-based inference does not use any parametric learning and therefore it can only rely on statistics. Interestingly, the rule-based method obtains similar performance to the MLP model indicating the MLP model merely learns from the bias of the dataset.

The relational transformer model improves results over MLP. Furthermore, the relational GNNED performance is comparable to the relational transformers. The transformer variants and GNNED have similar architectural properties and have better relational modeling capacity than the MLP model. These models exploit the interrelation between visual and semantic relational representations to better understand the visual scene. This potentially helps to boost the performance of abductive past action inference.

Surprisingly, Relational Bilinear Pooling (RBP) obtains particularly good results outperforming the transformer and GNNED models. The way relational reasoning is performed in RBP is fundamentally different from transformers and GNNED. The RBP models interactions between the human and object features within a relation more explicitly than the GNNED and Transformer. However, unlike the GNNED or Transformer, RBP is unable to model interactions between relations. Finally, the combination of GNNED and RBP, i.e., BiGED performs even better. This result is not surprising as BiGED takes advantage of better inter and intra-relation modelling. We also experimented with a BiGED model, that takes K future frames starting from frame at T (i.e. Action Genome frames from T to $T + K$) as inputs. The results of this experiment suggests that use of future snapshots helps improve performance.

All object-relational models utilize the ground truth object labels from the AG dataset to obtain semantic representations. We observe a drop in performance when we use predicted objects from the FasterRCNN model. Additionally, FasterRCNN-based object-semantic prediction performs worse than the visual-only BiGED model (Table 4), indicating that incorrect semantics significantly harm performance. Nevertheless, the performance of BiGED with FasterRCNN labels is significantly better than end-to-

Model	**mAP**	**R@10**	**mR@10**
Human Performance	–	92.26	93.71
Relational MLP	26.58± 0.37	41.71± 0.82	25.40± 0.66
Relational Self Att. Transformer	27.94± 0.35	45.72± 1.42	30.12± 2.30
RBP	32.19± 0.44	53.76± 0.89	38.44± 0.67
BiGED	**34.13± 0.39**	**57.39± 0.10**	**41.97± 0.36**

Table 3. Abductive past action verification performance using the proposed methods on the *Abduct at T* setup.

end trained models and vision-language models. Finally, it should be emphasized that human performance on this task is significantly better than any of the modern AI models, highlighting a substantial research gap in developing AI systems capable of effectively performing abductive past action inference.

4.7. Results: Abduction on the Last Image

We evaluated object-relational models on the second setup, where we perform abduction on the last image of each video, using models trained in the previous setup. Due to the variety of possible actions in a video sequence, this setup is more challenging. It should be noted that this is a special case of *abduct at T*. This additional experiment allows us to observe and select the longest time horizon to determine whether the models are still able to abduct actions. Results in Table 2 show lower performance across all models compared to the previous setup, indicating the task's increased difficulty. The MLP model and rule-based inference perform poorly. The GNNED, RBP, and BiGED methods outperform the Transformer model, despite GNNED's similar architecture to the Transformer. BiGED achieves the highest mAP, while RBP excels in R@10 and mR@10.

4.8. Results: Abductive Past Action Verification

We present abductive past action verification results in Table 3 using the object-relational approach. We use the ground truth human and object class names to obtain the semantic representation. As the query is in textual form (i.e. the action class name), we suggest that the abductive past action verification resembles a human-like task. It is easy to answer yes, or no to the question "Did the person execute action a_i in this image to arrive at this state?" Interestingly, the performance of this task is slightly lower than the main results we obtained in Table 1. Even though this task is mentally more straightforward for the human, it seems the task is slightly difficult for the machine as it now has to understand the complexities of human languages.

4.9. Ablation on semantic vs visual

We use both visual and semantic features (Glove embedding of object names) to obtain the relational features – see Section 3.2. We ablate the impact of visual and semantic features on each model on Abductive Past Action Set Prediction (*abduct at T*) and the results are shown in Table 4.

Model	Visual Only	Semantic Only	Both
Rule-based inference	–	26.27	–
MLP	17.82	18.55	27.73
Transformer	21.30	32.81	33.59
Relational GNNED	21.55	32.82	34.38
RBP	22.15	**33.03**	35.55
BiGED	**24.62**	30.25	**35.75**

Table 4. mAP on Abductive Past Action Set Prediction (*Abduct at T*) using visual and semantic features.

While RBP achieves the best performance using ground truth object semantics, BiGED is the best-performing model for visual data alone by a considerable margin, making it the overall best method. We can conclude while semantic features are effective, both visual and semantic features are complementary. **Qualitative Results.** The qualitative results in supplementary material Section 4.4 demonstrate RBP and BiGED infer past actions more accurately. **Generalizability of BiGED.** A visual-only BiGED model trained to infer past actions was evaluated for action recognition at the video level. We obtained 50.5 mAP on the Charades dataset without any tuning. Although not state-of-the-art, these results suggest the value of the abductive past action inference model for general action understanding using only visual features.

5. Discussion & Conclusion

This paper introduces abductive past action inference, a task involving past action set prediction, sequence prediction, and verification, all formulated as closed-set classification tasks. Our experiments show that while deep learning models can perform these tasks to some extent, holistic end-to-end models are ineffective. Large multimodal models like CLIP show promise, but our proposed human-object relational approaches—such as relational graph neural networks, bilinear pooling, and the BiGED model—outperform them, demonstrating the value of object-relational modelling. We find conditional text generation unsuitable for this task due to limited control, and even advanced foundational models fail after instruction tuning. Overall, human-object-centric video representations emerge as the most effective approach, and abductive past action inference may enhance general human action understanding.

Acknowledgment. This research/project is supported by the National Research Foundation, Singapore, under its NRF Fellowship (Award NRF-NRFF14-2022-0001) and this research is partially supported by MOE grant RG100/23. It was also partly funded by an ASTAR CRF award to Cheston Tan and supported by the ASTAR CIS (ACIS) Scholarship awarded to Clement Tan. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of these organizations.

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

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

[3] Le-Wen Cai, Wang-Zhou Dai, Yu-Xuan Huang, Yu-Feng Li, Stephen H Muggleton, and Yuan Jiang. Abductive learning with ground knowledge base. In *IJCAI*, pages 1815–1821, 2021. 3

[4] Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, and Michael Ying Yang. Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16372–16382, 2021. 3

[5] Wang-Zhou Dai, Qiuling Xu, Yang Yu, and Zhi-Hua Zhou. Bridging machine learning and logical reasoning by abductive learning. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6, 7

[7] 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, 7

[8] Basura Fernando and Samitha Herath. Anticipating human actions by correlating past with the future with jaccard similarity measures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13224–13233, 2021. 2, 4

[9] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016. 3

[10] 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. 2

[11] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

[12] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13505–13515, October 2021. 2

[13] 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. 2

[14] Guodong Guo and Alice Lai. A survey on still image based human action recognition. *Pattern Recognition*, 47(10):3343–3361, 2014. 2

[15] Zhongyi Han, Le-Wen Cai, Wang-Zhou Dai, Yu-Xuan Huang, Benzheng Wei, Wei Wang, and Yilong Yin. Abductive subconcept learning. *Science China Information Sciences*, 66(2):122103, 2023. 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 5, 6, 7

[17] Samitha Herath, Mehrtash Harandi, and Fatih Porikli. Going deeper into action recognition: A survey. *Image and vision computing*, 60:4–21, 2017. 2

[18] Jack Hessel, Jena D Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. The abduction of sherlock holmes: A dataset for visual abductive reasoning. *arXiv preprint arXiv:2202.04800*, 2022. 1

[19] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3192–3199, 2013. 2

[20] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10236–10247, 2020. 5

[21] Yang Jin, Linchao Zhu, and Yadong Mu. Complex video action reasoning via learnable markov logic network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3242–3251, 2022. 3

[22] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 6

[23] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4804–4814, 2022. 6, 7

[24] Chen Liang, Wenguan Wang, Tianfei Zhou, and Yi Yang. Visual abductive reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15565–15575, 2022. 1

[25] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models, 2023. 7

[26] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 6, 7

[27] Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. Visualcomet: Reasoning about the dynamic context of a still image. In *European Conference on Computer Vision*, pages 508–524. Springer, 2020. 3

[28] Paritosh Parmar, Eric Peh, Ruirui Chen, Ting En Lam, Yuhan Chen, Elston Tan, and Basura Fernando. Causalchaos! dataset for comprehensive causal action question answering over longer causal chains grounded in dynamic visual scenes. *arXiv preprint arXiv:2404.01299*, 2024. 3

[29] Paritosh Parmar, Eric Peh, and Basura Fernando. Learning to visually connect actions and their effects. *arXiv preprint arXiv:2401.10805*, 2024. 3

[30] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 4

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4, 6, 7

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 4

[33] 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 (WACV)*, pages 6740–6750, January 2024. 2

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014. 6

[35] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008. 3

[36] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pages 593–607. Springer, 2018. 3

[37] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016. 6

[38] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3716–3725, 2020. 3

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[40] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 6, 7

[41] Aming Wu, Linchao Zhu, Yahong Han, and Yi Yang. Connective cognition network for directional visual commonsense reasoning. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[42] Changqian Yu, Yifan Liu, Changxin Gao, Chunhua Shen, and Nong Sang. Representative graph neural network. In *European Conference on Computer Vision*, pages 379–396. Springer, 2020. 3

[43] 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. 3

[44] Hao Zhang, Yeo Keat Ee, and Basura Fernando. Rca: Region conditioned adaptation for visual abductive reasoning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9455–9464, 2024. 1

[45] Qi Zhao, Shijie Wang, Ce Zhang, Changcheng Fu, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. Antgpt: Can large language models help long-term action anticipation from videos? In *ICLR*, 2024. 2

[46] Tianyang Zhong, Yaonai Wei, Li Yang, Zihao Wu, Zhengliang Liu, Xiaozheng Wei, Wenjun Li, Junjie Yao, Chong Ma, Xiang Li, et al. Chatabl: Abductive learning via natural language interaction with chatgpt. *arXiv preprint arXiv:2304.11107*, 2023. 3