Prompt-based Object-centric Video Representation for Action Anticipation

Abstract

This paper focuses on building object-centric representations for action anticipation in videos. Our key motivation is that objects provide important cues to recognize and predict human-object interactions, especially when the predictions are longer term, as an observed "background" object could be used by the human actor in the future. We observe that existing object-based video recognition frameworks either assume the existence of in-domain supervised object detectors or follow a fully weakly-supervised pipeline to infer object locations from action labels. We propose to build object-centric video representations by leveraging visual-language pretrained models. This is achieved by "object prompts", an approach to extract task-specific object-centric representations from general-purpose pretrained models without finetuning. To recognize and predict human-object interactions, we use a Transformer-based neural architecture which allows the "retrieval" of relevant objects for action anticipation at various time scales. We conduct extensive evaluations on the Ego4D Long-term Anticipation, 50Salads, and EGTEA Gaze+ benchmarks. Both quantitative and qualitative results confirm the effectiveness of our proposed object prompts and the overall model.

1. Introduction

Given an egocentric video observation, the action anticipation task [51] is defined as generating an action sequence of the camera-wearer in the form of verb and object pairs. Of particular interest is the long-term action anticipation (LTA) task [20], which aims to anticipate future actions over a long time-horizon. A reliable action anticipation algorithm is crucial for building intelligent agents, as it provides important signals for planning in interactive environments.

This paper aims to build effective object-centric video representations for action anticipation. As illustrated in Figure 1, our key motivation is that a detailed, object-centric understanding of the scene provides visual cues on the goals of the actions and the available tools to be interacted with. While objects have been shown to play a crucial role for action understanding in both humans [49, 58] and machines [62, 40], their impacts on video-based action anticipation is yet to be studied. Among the earlier attempts at object-based video action understanding, one common approach is to leverage object detectors trained on in-domain bounding box annotations [9, 23, 62, 40] on the same or similar datasets. While effective, the in-domain bounding box annotation process is time-consuming and labor-intensive, which makes the corresponding frameworks unlikely to scale to visually more diverse videos and complex, cluttered scenes. The other approach is to leverage generic object proposals [57, 54] or to directly work with image patches [53, 45, 18], and rely on the attention mechanism [55] to pick the salient regions with weak supervision from the action labels. Despite being more flexible, this approach does not incorporate prior knowledge on object locations, and often struggles to "detect" the actual objects, especially when the training data is limited.

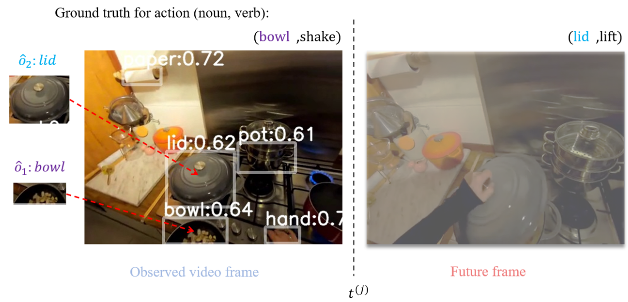

Figure 1: Objects are not only helpful for action recognition (left, *shake a bowl*), they also reveal the possible options for future human-object interactions (right, *lift a lid*). We propose object prompts, which leverage visual-language pretrained models (*e.g.* GLIP [29]) to build object-centric video representations without dataset-specific finetuning.

We propose to address the limitations of the existing approaches by leveraging visual-language models pretrained on large-scale datasets, such as GLIP [29]. We hypothesize that these pretrained models, whose objective is to (contrastively) associate image regions with text descriptions, learn generic object-centric representation that can be transferred to the action anticipation task, without the need to finetune their weights. We identify two challenges in order to investigate whether visual-language pretrained models help anticipation: First, how to properly "query" the pretrained models to retrieve the most relevant objects in a video observation, based on the domain knowledge; Second, how to associate different objects in an often cluttered scene to predict different human-object interactions in a long time-horizon. For the first challenge, we propose *object prompts*, which incorporate the domain knowledge of the target dataset by mapping the action (*e.g*. verb and object pairs) vocabulary into object prompts, which are used to query the visual-language pretrained models. For the second challenge, we propose using a *predictive transformer encoder* (PTE), which is a Transformer encoder network that jointly attends to the motion cues (as encoded by pretrained video ConvNets or Transformers) and the object-centric representation (based on the object prompts), and dynamically associates the motion and object evidence in order to predict future actions at different time steps. Our overall proposed framework is based on the Transformer architecture, and is trained end-to-end with the future action classification objectives.

We conduct thorough experiments on the long-term action anticipation benchmark in the Ego4D dataset [20], and the anticipation benchmarks in the Salads [51] and EGTEA Gaze+ [30] datasets. Our quantitative experiments confirm that the prompt-based object-centric representations can substantially improve the action anticipation performance. Ablation experiments reveal that it is important to incorporate in-domain knowledge when designing the object prompts, and that object-centric representation significantly outperforms image-level counterpart [11]. In addition, qualitative analysis on object attention visualization shows that the model learns to associate the corresponding objects when predicting actions at different time steps.

In summary, our contributions in this paper are threefold: First, we demonstrate the effectiveness of object-centric representation for video action anticipation; Second, we propose "object prompts" to incorporate domain information when querying the pretrained models, and predictive transformer encoder to dynamically associate the object evidence for action anticipation; Finally, we provide extensive quantitative and qualitative analysis of the proposed framework, which achieves competitive performance on three benchmarks. Our implementation along with the pretrained models will be released.

2. Related Work

Object-centric video representation is an active research area for action recognition applications. The motivations include producing a more compact, thus efficient video representation by attending to regions of interest; and enabling compositional generalization to unseen human-object interactions [41] with a structured representation. For example, Wang et al. [57] extract RoI features from an 3D CNN feature maps using off-the-shelf detector, and build a graph neural network on top of RoI features alone. LFB [59] and Object Transformer [60] load off-the-shelf object features from object detectors in video backbones to encode long-term video features. Recently, with the advance of transformer-based architectures, ORViT [23] proposes to crop [22] object regions as a new object tokens and attach them to pixel tokens. ObjectLearner [62] fuses an object-layout stream and a pixels stream using an object-to-pixel transformer. Most of the existing object-based approaches either assume the availability of in-domain bounding box annotations, or apply generic "objectness" criteria (*e.g*. bounding box proposals trained on COCO [32]). We propose object prompts to leverage a pretrained visual-language model to generate task-specific object representations.

Video Transformers. Transformers [14] are now the predominant architectures for video recognition. The vanilla vision transformer (ViT) [14] evenly devides images into non-overlapping tokens, and run multi-head self-attention [56] over the tokens. TimeSFormer [5] and ViViT [3] extends ViT to videos, by introducing cube tokenization and efficient cross-time attention, i.e., axis-based space-time attention [5] or factorized attention [3]. MotionFormer [45] enhance space-time attention by an implicit trajectory attention module. MViT [15, 31] and VideoSwin Transformer [35] re-introduce resolution-pooling as in convolution nets in video transformers for efficiency.

Visual-Language Pretrained Models. We have collectively made huge progress towards building unified learning frameworks for a wide range of tasks, including natural language understanding [12, 47, 7, 34], visual recognition [28, 26, 61, 17], and multimodal perception [24, 52, 36, 19, 2]. As this pretraining-adaptation learning paradigm gains momentum, researchers at Stanford [6] even coined the term "foundation models" to refer to these pretrained neural networks. While the earlier visual-language pretrained models work with image-level [36, 46] or video-level [52, 38] representations, more recent models are object grounded [29, 43, 63, 21, 37]. We explore the benefits of both object-level [29] and image-level [46] for action anticipation.

3. Methods

In this section, we first introduce the long-term action anticipation (LTA) problem formulation and the next action prediction (NAP) formulation. Then, we describe our overall model architecture. Finally, we describe our choices on object detections and representation in detail.

3.1. The Action Anticipation Task

Given the significant applications of action anticipation, numerous benchmarks have been introduced to evaluate

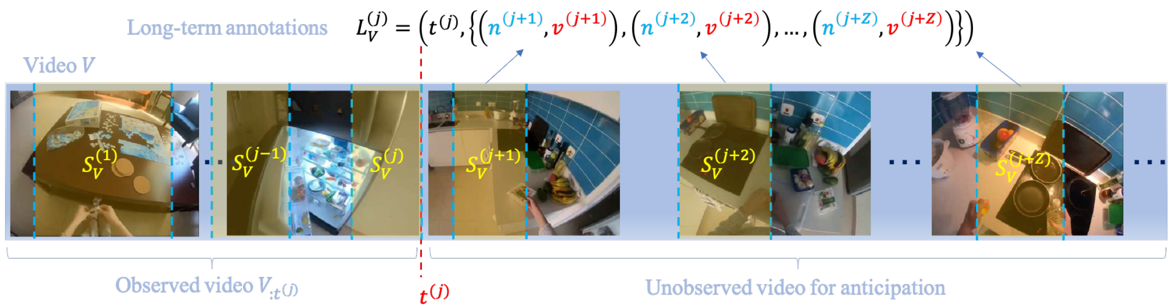

Figure 2: **Illustration of the Action Anticipation task.** In the long-term action anticipation (**LTA**), the learned model is expected to predict a sequence of Z actions, in the form of object and verb pairs, given visual observations up to time $t^{(j)}$ in the video. $t^{(j)}$ is the end time for the j-th labeled segment in the original video. During evaluation, the edit distance between a predicted sequence and the ground truth sequence is computed. To account for uncertainty in action anticipation, the model can predict up to K sequences for each input example. In the next action prediction (**NAP**), the learned model is only expected to predict a set of actions. ($Z = 1$)

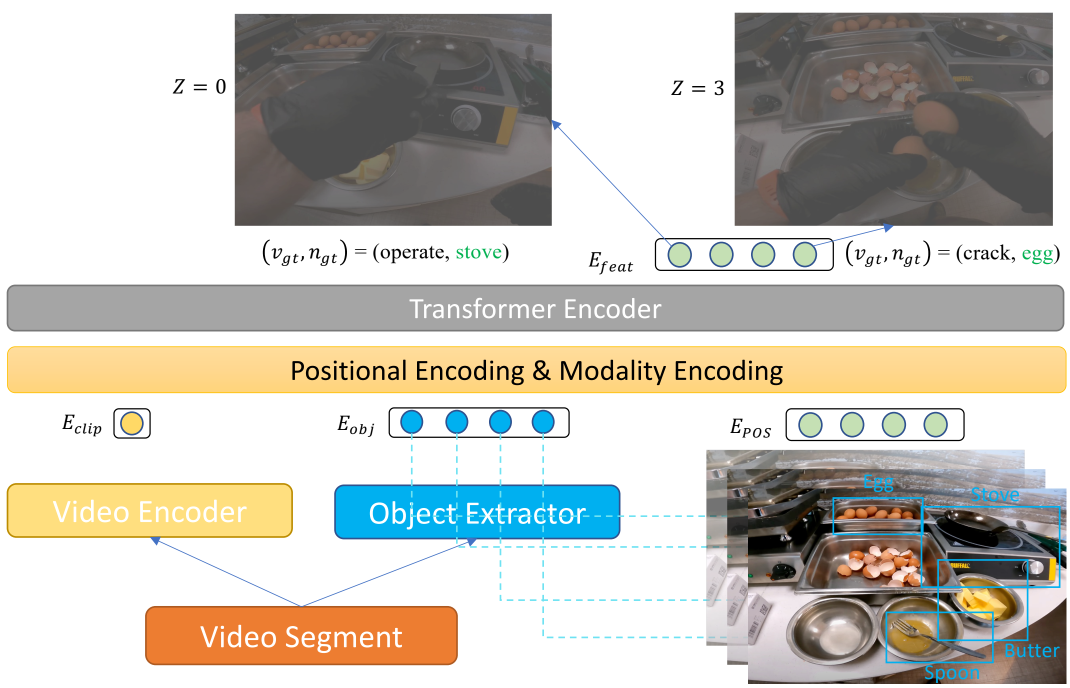

Figure 3: **Illustrations of overall model architecture** when $N_v = 1, N_o = 4, Z = 4$ where N_{seg} is the number of input segments, N_{obj} is the number of objects per segments, Z is the number of future actions we want to predict. Given the features $E_{\text{clip}}, E_{\text{obj}}$ extracted by the encoders, Predictive Transformer Encoder (PTE) generates Z features for future prediction. A decoder network is applied on each generated feature z_i to compute verb and object probabilities.

model performance [48, 27, 44, 9, 10, 54]. Long-term action anticipation **LTA** is problem setup introduced by the Ego4D benchmark [20]. As illustrated in Figure 2, the provided annotations first split a long video V into smaller segments $\{S_V^{(i)}\}$, where $S_V^{(i)}$ is the i-th annotated video segment. Each segment is labeled with its starting time, end time, and action label. The action label is represented as one verb, object pair $(n^{(j)}, v^{(j)})$ for each segment $S_V^{(j)}$. The LTA task is specified by a "stop time" $t^{(j)}$, which denotes the end time for the last observed video segment $S_V^{(j)}$. The learned model is allowed to observe any video frames before $t^{(j)}$ in order to make future predictions $\{(n^{(j+1)}, v^{(j+1)}), \ldots, (n^{(j+Z)}, v^{(j+Z)})\}$, where Z is the number of future steps to predict. For example, suppose a person is frying an egg with a pan in a kitchen where knives, onions, water, and pots are scattered around. In this scenario, the LTA requires the model to predict the person's upcoming actions sequentially, such as picking up the knife, cutting the onion, and drinking water. To account for the uncertainty of future behaviors, the model is allowed to make up to K sets of action predictions for each future step. We follow the standard setup and use $Z = 20$, $K = 5$. More details on the evaluation metric are in Section 4.

Next action prediction (**NAP**) is problem setup which focuses on predicting the action at the next step ($Z = 1$). As shown in Figure 2, the learned model is also allowed to observe any video frames before $t^{(j)}$ but only make one future prediction $\{(n^{(j+1)}, v^{(j+1)})\}$ (*i.e.*, $K = 1$) for video segment $S_V^{(j+1)}$.

3.2. Overall Model Architecture

We now introduce our model architecture as illustrated in Figure 3. We follow the standard experimental setup as used in [20]. Given a video V and the stop time $t^{(j)}$, our model takes a sequence of video segments $\{S_V^{(j-N_\text{v}+1)}, ..., S_V^{(j)}\}$ as input observation, and generate a sequence of actions $\{(n^{(j+1)}, v^{(j+1)}), ..., (n^{(j+Z)}, v^{(j+Z)})\}$ as outputs. N_v is the number of observed video segments, and Z is the number of future steps. Our overall model architecture consists of three modules: (1) a collection of video or object encoders that generate multimodal representations from video segments; (2) an aggregator network which fuses multimodal input representation across space and time; (3) an output decoder which generates action predictions from the aggregated features.

Encoders. We sample N_v clips, each one from each of the N_v input video segments. Then we pass the clips to a video encoder to generate clip-level representations $E_\text{clip} \in \mathbb{R}^{N_\text{v} \times D}$, where D is the encoded embedding size. We also sample N_o objects from the N_v video segments. Then we then use an object encoder to generate object representations $E_\text{obj} \in \mathbb{R}^{N_\text{o} \times D}$. We will discuss our choices in object detectors and object encoders in following sections.

Aggregators. We introduce Predictive Transformer Encoder (PTE), a Transformer-based architecture for the action anticipation task. Given E_clip and E_obj, PTE is used to generate Z features $z_0, z_1, ..., z_{Z-1}$ for future prediction.

PTE has learnable tokens $E_{POS} \in \mathbb{R}^{Z \times D}$. It concatenate E_{clip}, E_{obj} and E_{POS} to form a sequence of length $N_\text{v} + N_\text{o} + Z$, then adds positional and modality encodings to the entire sequence. Finally, it passes the sequence to a vanilla Transformer Encoder [55], then take out features corresponding to the last Z learnable tokens as $z_0, z_1, ..., z_{Z-1}$.

Decoders and Training Objectives. For each z_i generated by PTE, we apply one linear layer on top of it to generate the logits for verb predictions, and one seperate linear layer for noun predictions. We use Softmax Cross-Entropy as the loss function, and assign equal weights to all future steps.

Fusion strategies. Besides early fusion, we also implement late fusion. Leveraging the advatages of PTE, we first only use video features to generate logits (outputs of the decoder) for future actions, then we only use object features to generate the logits. We average the logits from the video and object streams as the final logits.

3.3. Object-based Video Representation

We now describe our choices on object detections and representation. By LTA task definition, object information should provide important cues to predict human actions. Taking inspiration from this, we propose to leverage GLIP [29], a grounded language-image pretrained model to construct object-based video representation. We hypothesize that models pretrained on diverse object detection and phrase grounding datasets like GLIP already encode transferrable object information, and can be accessed using task-specific "prompts" (e.g. common objects used in the target dataset). Thus, it is crucial to develop suitable text prompts to utilize the object-centric information encoded by GLIP for the LTA task.

3.3.1 Leveraging Pretrained Grounding Models

Grounded Language-Image Pre-training (GLIP) [29] proposed a pretraining strategy to build vision-language foundation models that can be further adopted for "zero-shot" object detection. During pretraining, GLIP is agnostic to the choice of object detectors to generate region proposals. In our work, we follow GLIP's setting to use Dynamic Head [8] object detector for images, and the BERT [13] encoder for text inputs. Similar to applying CLIP [46] for zero-shot image classification, the zero-shot object detection can be achieved by querying with "object prompts" (*e.g. "an image region with a cat"*) to the pretrained language-image grounding model.

To retrieve object detector's features corresponding to finally chosen region proposals, we design two approaches. First, we inject a special identifier inside each region proposal data structure in order to gain original object features for each proposal. Second, we crop and feed the object proposal region into a pretrained image encoder to generate object features. Additionally, we also extract the object category level alignment scores and the box location to append to the object-level representation. The second approach allows us to leverage any image-language pretrained model, such as CLIP [46].

3.3.2 Object Prompt Strategy

In order to get object-level features from grounded pretrained models, proper prompts design is necessary. We hy-

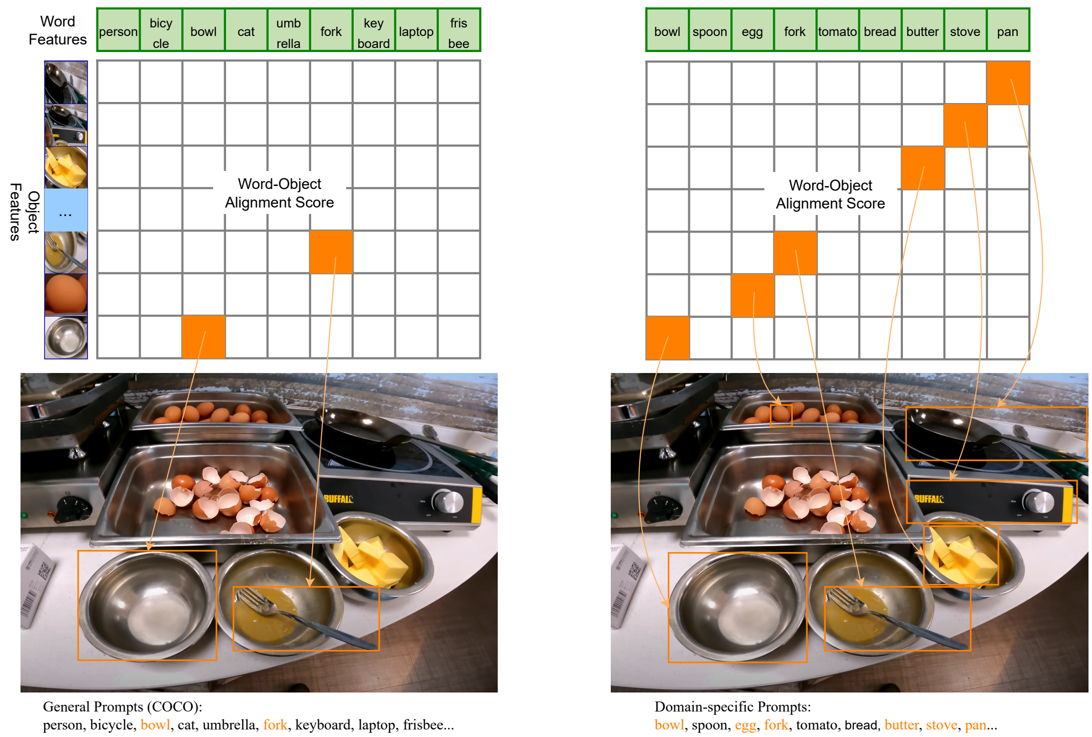

Figure 4: **Domain-specific object prompts are necessary to extract effective object-centric representation from visual-language pretrained models.** Here we illustrate the procedure of aligning words in prompts with object-level features by calculating contrastive scores. For this example in a kitchen scene, on the left we show the list of objects obtained from the COCO dataset classification vocabulary, on the right is the list of objects obtained by our object prompts.

pothesize that the desirable object prompts should incorporate the domain knowledge (*e.g*. common objects appearing in the dataset), and explore different strategies to design the object prompts.

In order to obtain the object prompts, we first define the object vocabulary that contains the object classes to be detected. One simple and intuitive solution is to directly borrow the vocabulary used in the task: For LTA, this refers to the list of objects to be interacted with. We then explore two intuitive approaches to refine the vocabulary: 1) picking the most common object categories based on their frequency in the training data of the target task; 2) using word embedding (*e.g*. word2vec [42]) and K-Means clustering to group similar categories. In addition, we also explore the vocabulary used by the COCO [32] dataset for comparison.

Figure 4 illustrates the importance of having domain-specific object prompts, as provided by the most common objects or word clusters. Figure 5 shows the actual detections by GLIP with different object prompt strategies.

4. Experiment

We conduct extensive experiments and ablation studies to demonstrate the effectiveness of our proposed approach in action anticipation task.

4.1. Experiment Setup

Ego4D [20] contains 3,670 hours of daily life activity egocentric video spanning hundreds of scenarios. We focus on videos under Forecasting subset which contains 1723 clips with 53 scenarios and around 116 hours. In addition, it contains 115 verb as interaction behaviors and 478 noun as objects. We follow the standard train, validation, and test splits from [20] annotations for evaluation.
EGTEA Gaze+ contains 28 hours egocentric videos of cooking activities from 86 unique sessions of 32 subjects. Each video is annotated with interactive actions, spanning from 19 verb as interaction behaviors and 53 noun as objects. We follow the same train test split from annotations to for evaluation
50Salads contains 25 people preparing 2 mixed salads each (50 videos totally). There are on average 17 action classes and 20 action sequence per video. We follow the dataset

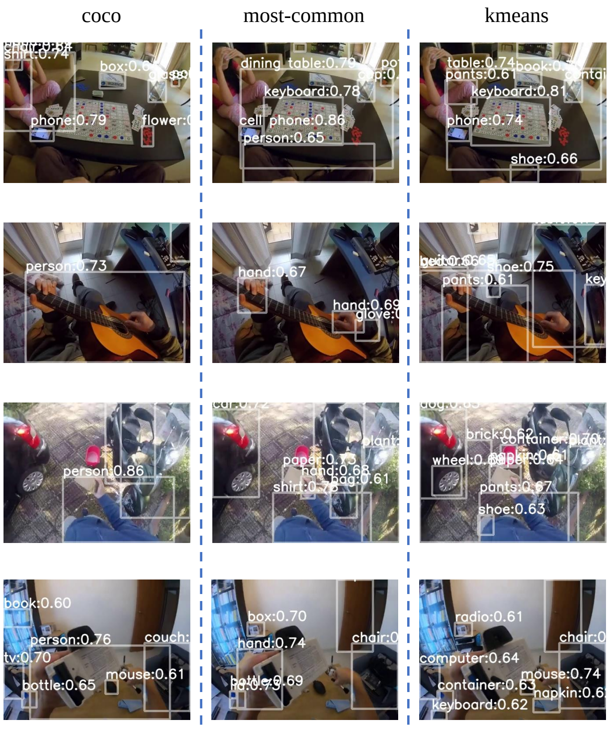

Figure 5: **Example GLIP detection results with different object prompts** on randomly sampled video frames. The K-Means clustering strategy (last column) offers the best precision and recall of relevant objects in all four cases.

standard and perform 5-fold cross-validation in our evaluation.

Metrics. For Ego4D, we follow the standard [20] using edit distance (ED) for evaluation metrics, which is computed as the Damerau-Levenshtein distance over sequences of predictions of verbs, nouns and actions. For the K possible sequences the model predicts, we choose the smallest edit distance between the ground truth and any of the K sequences. Following [20], we set $K = 5$ in our experiments. We report Edit Distance at $Z = 20$ (ED@20) on the test set and Average Edit Distance (AUED) on the training set.

For EGTEA Gaze+ and 50Salads, we report top1 accuracy, class-mean top1 accuracy for Next Action Prediction. Following the long-term action anticipation setting in [16], we report the mean over classes accuracy (MoC@(O,F)), where O is the fraction of observation, F is the fraction of future. In our experiments we report MOC@(0.2, 0.5) which is the hardest anticipation setting in the 50Salads LTA benchmark.

4.2. Ablations

In this section we focus on the LTA benchmark on the Ego4D dataset.

Object prompts. By designing object prompts, we are able to incorporate our domain knowledge into specific tasks. The Ego4D [20] Dataset has 478 nouns which is too large as a vocabulary for GLIP [29]. To handle this, we explore three types of vocabulary, each containing only 80 words as object prompts: "most-common", "kmeans", "coco" in Table 1a. "most-common" contains top 80 frequent nouns. "kmeans" contains top 80 frequent nouns after clustering to remove redundancy. "coco" contains the object categories appearing in the COCO [32] dataset. We also include "random" as a baseline which randomly picks regions in images as bounding boxes. Both "most-common" and "kmeans" outperform "coco" and "random" significantly, showing the effectiveness of our designed object prompts.

Object locations and categories. Locations and categories are important features for action recognition. In Table 1c, we demonstrate that both location and category features provide useful information. After removing location features and category features, we observe a 0.003 (0.4%) and 0.015 (1.9%) raise in noun ED respectively. Due to the long-term nature of the task, object locations are less important than categories. Besides providing specific classes of the objects, a set of category information can also provide the model with high-level scene information *e.g.* "person", "laptop", and "book" might indicate a library. The general information helps the model infer long-term future actions.

Object quantity and quality. Object qualities can be reflected by their confidence scores. Higher scores usually mean better qualities. In Table 1d, we use a threshold on confidence scores to filter objects with low qualities on the dataset level. For objects that have lower scores than the given threshold, we replace the corresponding object features with a learnable padding token. We see the model performs the best when we set threshold to 0.3. In Table 1b, we select different numbers of top-ranking objects on the frame level based on their confidence scores. We see the model performs the best when we use 5 objects per frame. We argue that there is a trade-off between the quantity and quality of objects. It is good to include more detections while bad detections can have a negative influence on long-term action anticipation. Thus, it is important to control the threshold and number of objects.

Temporal modeling. We compare our temporal modeling methods with Ego4D LTA Baseline [20] in Table 2. Using only video modality, we observe significant improvement in both verb and noun AUED when we use PTE. This demonstrates that PTE are more effective in long-term temporal modeling than naive Transformer Encoder.

Fusion with video backbone. We explore the impact of object modalities in Table 2. Compared with PTE+video, PTE+(video+object, early) brings 0.006 (0.8%) and 0.010 (1.3%) improvement on verb and noun respectively. This confirms our hypothesis that object-centric representation

Vocabulary	Verb ↓	Noun ↓
random bbox	0.745	0.945
coco	0.737 (-1.0%)	0.789 (-16.5%)
most common	0.734 (-1.5%)	0.776 (-17.8%)
kmeans	**0.728 (-2.3%)**	**0.771 (-18.4%)**

(a) **Vocabulary**

#obj	Verb ↓	Noun ↓
1	0.735	0.834
3	0.727	0.800
5	**0.728**	**0.771**
10	0.731	0.781

(b) **Number of Objects**

Loc.	Cate.	Verb↓	Noun↓
✗	✗	0.730	0.789
✓	✗	0.730	0.786
✗	✓	0.731	0.774
✓	✓	**0.728**	**0.771**

(c) **Location and Category**

Threshold	Verb ↓	Noun ↓
0.00	0.731	0.781
0.30	**0.728**	**0.771**
0.45	0.728	0.773
0.55	0.731	0.787

(d) **Threshold**

Table 1: **Ablation experiments on object-only models**. We conduct detailed ablation on (1) object vocabulary, (2) number of object per frame, (3) object location and category, (4) detection threshold.

Aggregator	Modality	Fusion	Verb ↓	Noun ↓
Baseline	video	-	0.751	0.766
PTE	video	-	0.713 (-5.1%)	0.753 (-1.7%)
PTE	video+object	early	**0.707(-5.9%)**	**0.743 (-3.0%)**
PTE	video+object	late	0.709 (-5.6%)	0.748 (-2.3%)

Table 2: **Temporal modeling and modality fustion on Ego4D LTA.** PTE is more effective in temporal modeling. Object significantly helps action anticipation.

Model	Modality	AUED(Verb)↓	AUED(Noun)↓
-	video	0.713	0.753
GLIP	video+object	0.707 (-5.9%)	0.743 (-3.0%)
CLIP	video+object	**0.700 (-6.8%)**	**0.717 (-4.8%)**

Table 3: **Additional pretrained models for object representation.** Incorporating CLIP brings additional performance gain over GLIP object embeddings.

helps action anticipation.

Fusion strategy. We compare two fusion strategies in Table 2, namely early fusion where video and object representations are jointed encoded by PTE, or late fusion where the two modalities are encoded separately until the last layer. We observe that allowing joint attention from the input layer generally improves the performance.

Incorporating image pretrained models. While we use GLIP [29] to detect and represent objects, their are many other pretrained models we can use to obtain object representation. In Table 3, we explore additional pretrained models for object representation. We still use GLIP [29] for object detection but use CLIP [46] to represent detected objects. CLIP object embeddings significantly helps video modality, and brings 0.007 (1.0%) and 0.026 (3.5%) improvement compared with GLIP object embeddings. This shows CLIP is more powerful in representing objects and demonstrates the importance of object representation. More pretrained models can be explored, which we leave for future works.

4.3. Qualitative Analysis

In Fig. 6, we show several qualitative examples of object attention weights produced by PTE. We use attention rollout [1] to compute attention weights from Z output action (noun, verb) pairs to previous visual observations and choose top 10 objects which has the relatively highest weight. Comparing with ground truth label shows the model learns to associate the corresponding objects when predicting actions at different time steps.

4.4. Comparison to the State-of-the-art

Table 4 compares our model with previous methods.

Ego4D. We compare our best model (Slowfast + PTE (video + CLIP object, early)) with recent state-of-the-art. We report verb, noun and action AUED on the test set. Our model achieves competitive results. Note that even though HierVL [4] uses additional text annotations from the Ego4D dataset, our model still has a comparable performance.

50Salads. We conduct experiments on both Next Action Prediction and Long-term Action Anticipation following the setting in [16]. For LTA, we predict 50% of the video

Model	Verb ↓	Noun ↓	Action ↓
HierVL* [4]	**0.7239**	**0.7350**	**0.9276**
Brunos (ours)	0.7265	0.7396	0.9290
ICVAE[39]	0.7410	0.7396	0.9304
VCLIP [11]	0.7389	0.7688	0.9412
Baseline [20]	0.7389	0.7800	0.9432

(a) **Ego4D LTA on the test set.*** used additional text annotations from the **Ego4D** dataset.

Model	top1 acc ↑	LTA@ 50% ↑
RNN [16]	30.1	13.49
CNN [16]	29.8	9.87
ActionBanks [50]	40.7	-
Slowfast+PTE (V)	41.0	15.2
Slowfast+PTE (V+O)	**43.8**	**16.9**

(b) **50Salads NAP and LTA**

Model	class-mean acc ↑
I3D-Res50 [25]	34.8
FHOI [33]	36.6
Slowfast+PTE (V)	35.8
Slowfast+PTE (V+O)	**36.8**

(c) **EGTEA Gaze+** top-1 acc.

Table 4: **Comparison with previous works.**

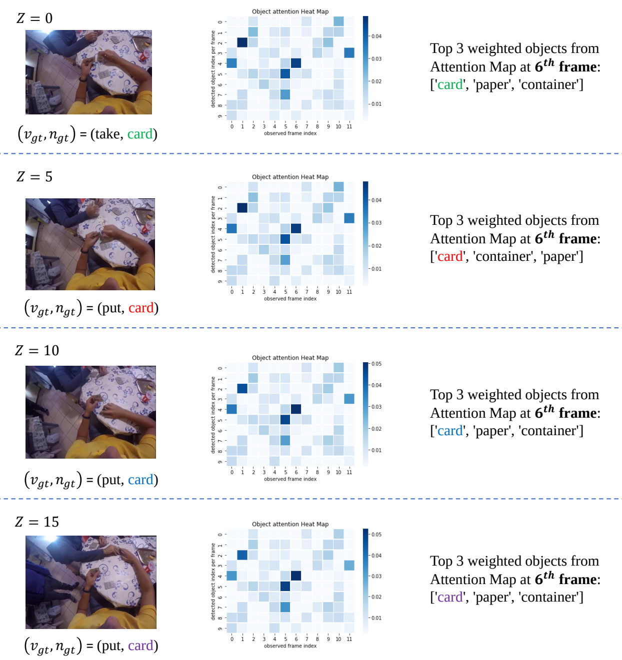

Figure 6: **Visualization of object attention heatmap and retrieved objects** , when $N_{\text{seg}} = 3, N_{\text{img}} = 4, N_{\text{obj}} = 10, Z = 20$. **Left**: Representative video frame correlate to Z^{th} future step for anticipation, where (v_{gt}, n_{gt}) are ground truth action labels at Z^{th} step in (verb, noun) pairs. **Middle**: Normalized object attention heatmap to previous observed visual input at Z^{th} step. **Right**: Top 3 weighted objects from heatmap according to the center observed (6^{th}) frame.

after observing 20% of the video as in [16]. We report top1 action accuracy in Next Action prediction and class-mean top1 action accuracy in Long-term Action anticipation as in [16]. We use PTE as temporal modeling method in video-only models and CLIP object representation in video+object models. Compared with video-only models, video+object model brings 2.8% improvement in Next Action Prediction and 1.7% improvement in LTA@50%. This shows the effectiveness of object modality in both short-term and long-term action anticipation. Besides first-person-view videos, object also helps action anticipation on third-person-view videos

EGTEA Gaze. Table 4c shows results on the next action prediction benchmarks for EGTEA+ Split 1 as in recent work [33]. We use CLIP object representation in video+object models, PTE as temporal modeling and finetune the video backbone. We report top1 accuracy following common practice. By adding object modality on top of video modality, we notice 1% improvement. This shows that object representation also helps human-object interaction modeling on short-term action anticipation.

5. Conclusion

We propose a prompt-based approach to construct object-centric video representation from pretrained visual-language models. We demonstrate the effectiveness of object-centric representation on two action anticipation settings, namely next-action prediction (NAP) and long-term action anticipation (LTA). We propose two modules, “object prompts” which incorporate domain information to query pretrained grounded models, and predictive transformer encoder, which dynamically associates the object evidence for long time-horizon action prediction. Experiments results confirm that both modules improve the action anticipation performance. We report encouraging results on Ego4D Long-term Anticipation, 50Salads, and EGTEA Gaze+ benchmarks.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *ACL*, 2020. 7

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 2

[3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. *arXiv preprint arXiv:2103.15691*, 2021. 2

[4] Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, and Kristen Grauman. Hiervl: Learning hierarchical video-language embeddings, 2023. 7, 8

[5] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021. 2

[6] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 2

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. In *NeurIPS*, 2020. 2

[8] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic head: Unifying object detection heads with attentions. *CoRR*, abs/2106.08322, 2021. 4

[9] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The EPIC-KITCHENS dataset. *CoRR*, abs/1804.02748, 2018. 1, 3

[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision. *CoRR*, abs/2006.13256, 2020. 3

[11] Srijan Das and Michael S Ryoo. Video+ clip baseline for ego4d long-term action anticipation. *arXiv preprint arXiv:2207.00579*, 2022. 2, 8

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. 4

[14] 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. *ICLR*, 2021. 2

[15] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, 2021. 2

[16] Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what? - anticipating temporal occurrences of activities, 2018. 6, 7, 8

[17] Golnaz Ghiasi, Barret Zoph, Ekin D Cubuk, Quoc V Le, and Tsung-Yi Lin. Multi-task self-training for learning general representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8856–8865, 2021. 2

[18] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *CVPR*, 2019. 1

[19] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16102–16112, 2022. 2

[20] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson

Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video, 2021. 1, 2, 3, 4, 5, 6, 8

[21] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 2

[22] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 2

[23] Roei Herzig, Elad Ben-Avraham, Karttikeya Mangalam, Amir Bar, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Object-region video transformers. In *CVPR*, 2022. 1, 2

[24] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021. 2

[25] 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. 8

[26] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018. 2

[27] Kris M. Kitani, Brian D. Ziebart, James Andrew Bagnell, and Martial Hebert. Activity forecasting. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision – ECCV 2012*, pages 201–214, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. 3

[28] Iasonas Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6129–6138, 2017. 2

[29] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. *CoRR*, abs/2112.03857, 2021. 1, 2, 4, 6, 7

[30] 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. 2

[31] 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 *CVPR*, 2022. 2

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2014. 2, 5, 6

[33] Miao Liu, Siyu Tang, Yin Li, and James Rehg. Forecasting human-object interaction: Joint prediction of motor attention and actions in first person video, 2019. 8

[34] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 2

[35] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *CVPR*, 2022. 2

[36] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 2

[37] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022. 2

[38] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 2

[39] Esteve Valls Mascaro, Hyemin Ahn, and Dongheui Lee. Intention-conditioned long-term human egocentric action forecasting @ ego4d challenge 2022. 8

[40] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *CVPR*, 2020. 1

[41] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *CVPR*, 2020. 2

[42] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013. 5

[43] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers. *arXiv preprint arXiv:2205.06230*, 2022. 2

[44] Tushar Nagarajan, Yanghao Li, Christoph Feichtenhofer, and Kristen Grauman. EGO-TOPO: environment affordances from egocentric video. *CoRR*, abs/2001.04583, 2020. 3

[45] Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and João F Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. *NeurIPS*, 2021. 1, 2

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 2, 4, 7

[47] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 2

[48] Nicholas Rhinehart and Kris M. Kitani. Online semantic activity forecasting with DARKO. *CoRR*, abs/1612.07796, 2016. 3

[49] Scott J. Robson and Valerie A. Kuhlmeier. Infants' understanding of object-directed action: An interdisciplinary synthesis. *Frontiers in Psychology*, 7, 2016. 1

[50] Fadime Sener, Dipika Singhania, and Angela Yao. Temporal aggregate representations for long-range video understanding, 2020. 8

[51] Sebastian Stein and Stephen J. McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *UbiComp*, 2013. 1, 2

[52] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, 2019. 2

[53] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Kevin Murphy, Rahul Sukthankar, and Cordelia Schmid. Actor-centric relation network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 318–334, 2018. 1

[54] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Rahul Sukthankar, Kevin Murphy, and Cordelia Schmid. Relational action forecasting. In *CVPR*, 2019. 1, 3

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. 1, 4

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 2

[57] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *ECCV*, 2018. 1, 2

[58] A. Woodward. Infants selectively encode the goal object of an actor's reach. *Cognition*, 69(1):1–34, 1998. 1

[59] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *CVPR*, 2019. 2

[60] Chao-Yuan Wu and Philipp Krahenbuhl. Towards long-form video understanding. In *CVPR*, 2021. 2

[61] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018. 2

[62] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Is an object-centric video representation beneficial for transfer? *arXiv:2207.10075*, 2022. 1, 2

[63] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. 2