
Opening the Vocabulary of Egocentric Actions

Dibyadip Chatterjee¹ Fadime Sener² Shugao Ma² Angela Yao¹

¹National University of Singapore ²Meta Reality Labs Research

{dibyadip, ayao}@comp.nus.edu.sg {famesener, shugao}@meta.com

<https://dibschat.github.io/openvocab-egoAR/>

Abstract

Human actions in egocentric videos often feature hand-object interactions composed of a verb (performed by the hand) applied to an object. Despite their extensive scaling up, egocentric datasets still face two limitations — sparsity of action compositions and a closed set of interacting objects. This paper proposes a novel open vocabulary action recognition task. Given a set of verbs and objects observed during training, the goal is to generalize the verbs to an open vocabulary of actions with seen and novel objects. To this end, we decouple the verb and object predictions via an object-agnostic *verb encoder* and a prompt-based *object encoder*. The prompting leverages CLIP representations to predict an open vocabulary of interacting objects. We create open vocabulary benchmarks on the EPIC-KITCHENS-100 and Assembly101 datasets; whereas closed-action methods fail to generalize, our proposed method is effective. In addition, our object encoder significantly outperforms existing open-vocabulary visual recognition methods in recognizing novel interacting objects.

1 Introduction

Egocentric or *first-person* videos captured from body-mounted cameras often feature the person manipulating objects with one or both hands. An action is thus conveniently defined as the composition of a verb performed by the hand with an interacting object, e.g., the action “*slicing apple*” is defined as “*slice*” and “*apple*”. Egocentric datasets [17, 8, 51] have scaled up by increasing the variety of verbs and objects. Theoretically, the increase in feasible actions should be quadratic, yet actually observed compositions are sparse, e.g., EPIC-KITCHENS-100 [8] has 97 verbs and 300 objects, but only 14% of possible verb-object compositions are observed as actions. Although not all compositions are feasible, e.g., “*eat dishwasher*”, many feasible actions are not observed, e.g., “*serve chicken*”, “*crush ice*”.

We observe that the verbs in egocentric datasets tend to be domain-agnostic. For example, even though EPIC100 is set in the kitchen, its verbs like *pull*, *remove*, and *shake* are also applicable outside of the kitchen. Ideally, we want to learn the verb concepts and generalize the verb to any object. Yet the dataset bounds the type and number of objects and the capture environment, e.g., EPIC100 has cooking utensils and food, whereas Assembly101 [51] has toy vehicle parts. Simply put, the objects restrict the subsequent action space of the dataset.

Existing action recognition models [61, 11, 43, 48, 22] are designed for a closed set of actions. A few works have studied open-set action recognition [4, 65], but they simply reject actions not seen during training. Inspired by the recent success of open vocabulary object detection [63, 18, 68], a more practical objective would be to provide novel actions as text inputs during inference. To that end, we address this problem of *open vocabulary action recognition* where a video model trained on a *base* vocabulary of actions needs to recognize *novel* actions outside of that vocabulary.

We tackle open vocabulary action recognition as two problems – (1) generalizing verbs to novel objects and (2) recognizing novel objects from egocentric videos. Note that this differs from a zero-shot setup, where some actions are not observed during training. These unobserved actions are not novel because the vocabulary is known and fixed during training, hence closed. More recently, [25, 57]

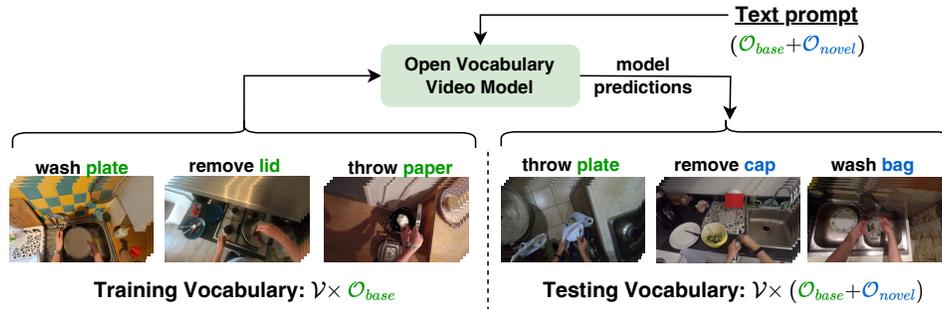


Figure 1: Open vocabulary action recognition for verb-object composed actions. During training, the model observes a predefined vocabulary of verbs and objects. During testing, the model is presented with seen verbs with any object (described by a text prompt). Objects in **green** are base classes (O_{base}) seen during training; those in **blue** are novel classes (O_{novel}) and unknown during training.

proposed recognizing an open vocabulary of verbs by leveraging large-scale vision-language models (VLMs) [46]. Contrary to their task, we are interested in the scenario of generalizing verbs to an open vocabulary of objects.

Generalizing verbs to novel object categories is challenging when the classes are naturally long-tailed [31, 44]. Models trained on such imbalanced datasets are biased towards head verb-object composed actions (when a verb is frequently seen with a particular object). Furthermore, these models suffer from static bias [7, 28] and overfit to the appearance of fixed frames without learning the transformation of objects over time [35]. To alleviate these biases, we decouple the verb and object predictions via two separately trained encoders. We propose an **Object Agnostic Pretraining** (OAP) for the verb encoder. OAP is based on a supervised contrastive loss [26], to which we add a set of *guiding augmentations*. These augmentations are specially designed for egocentric videos and allow the verb encoder to learn *object-agnostic* representations to improve transfer to novel objects. To our knowledge, this work is the first to explore contrastive learning of video backbones end-to-end on egocentric videos. In contrast, the augmentations in existing video contrastive learning [12, 45] are designed for Kinetics [6] styled videos where actions can be predicted from a single frame or a short clip sampled anywhere from the video. Such augmentations are ineffective for egocentric footage (Sec. C of Supplementary).

To recognize novel objects, we design an open-vocabulary object encoder that leverages a pretrained CLIP [46] model. We are specifically interested in recognizing *active* objects – those that undergo a state change from the actor’s interactions. Active objects are not necessarily *handheld* [64, 9] e.g., while “*beating egg*”, the actor holds a *whisk* and a *bowl* in his hand, but the *egg* is the active object undergoing a state change. Recognizing active objects in egocentric videos is challenging even under closed settings due to the many distractor objects in natural environments. To this end, we propose **Active Object Prompting** (AOP) that guides the CLIP model to understand which object is active. Given a frame containing a whisk, bowl, egg batter, hands, and other distractors, AOP aims at learning the context *i.e.* “*beating egg*” by generating a set of learnable verb-conditioned prompts. Conditioning the prompts on the verb features generated by OAP helps the CLIP recognize the active object.

Egocentric datasets [8, 51] focus on a closed set evaluation of actions where a handful of objects are unseen. However, for a realistic open vocabulary evaluation, a variety of novel objects is expected to be encountered for the first time during inference. To solve this, we create open vocabulary benchmarks on EPIC100 and Assembly101 by repurposing their original split. Our contributions can thus be summarized as: (i) an object agnostic pretraining of video backbones to improve its verb generalization to novel objects, (ii) an active object prompting of CLIP to recognize novel active objects and (iii) open vocabulary action recognition benchmarks on EPIC100 and Assembly101.

2 Related Works

Egocentric Action Recognition. Egocentric computer vision has garnered huge interest in recent years, and massive efforts at scaling up have resulted in the development of datasets like Ego4D [17],

EPIC-KITCHENS-100 [8] and Assembly101 [51]. It has also aided the development of sophisticated video recognition architectures [11, 3, 43, 58]. Recent works [48, 22, 33] focus on learning more object-centric representations by explicitly modeling object transformations over time with off-the-shelf detectors [50] and trackers [5]. However, these models are restricted to a closed set of predefined actions and do not scale to the diversity of objects in the real world.

Contrastive Learning in Videos. Videos provide unique opportunities for contrastive learning [19, 41] due to its multimodal and highly redundant nature. Most popular methods follow the self-supervised instance discrimination task [59] where a single positive of an anchor video is contrasted against a set of negatives. This positive can be an augmented view of the RGB stream [12, 45, 49] or sampled from a complementary modality like optical flow [40, 60], text [14, 36, 62, 21, 30], audio [2, 27, 34, 15] or even all three at once [1, 53]. Contrary to instance discrimination, recent works [26, 20] propose supervised contrastive learning. They show that sampling multiple positives from the same class as the anchor outperforms instance discrimination. Another line of research uses multiple positives to combat noisy anchor-positive alignment [37]. In this paper, we propose a set of lightweight augmentations to train video backbones for verb prediction. These augmentations result in a noisy alignment due to the camera motion inherent in egocentric videos. Thus, we leverage multiple positives by modifying [37] to combat it.

Learning for an Open vocabulary. The remarkable zero-shot capabilities of large language models (LLMs) [10, 47] and vision language models (VLMs) [46, 23] have widened the possibility of recognition and detection tasks in the visual domain. Of particular interest is the open vocabulary setting, which can recognize novel classes unknown during training. By presenting the novel classes as a textual phrase, it becomes possible to probe the huge corpora of knowledge in LLMs and VLMs. It differs from the traditional zero-shot setting where the novel classes are assumed to be known during training [63] and hence is restricted to the closed vocabulary of predefined classes. Most works for open vocabulary recognition are limited to the image domain [63, 67, 66, 18, 68, 39]. Recently, some works [56, 32, 25, 57] have adapted VLMs for video tasks but primarily focus on zero-shot verb recognition. In contrast, we are interested in a practical hand-object interaction scenario of generalizing verbs performed by hands to any object.

Prompting VLMs. Large Scale Vision Language Models (VLMs) like CLIP [46] and ALIGN [23] are trained on an enormous corpus of web data and contain semantically grounded knowledge of images. Prompting [24] refers to designing textual instructions to steer the VLM towards some desired output. Manually engineering discrete text prompts is laborious and requires domain expertise. One alternative is fine-tuning the VLM encoder for the task at hand; this approach incurs significant computation and is also at risk of catastrophic forgetting [29]. The more widely adopted approach is simply learning the prompts by backpropagating the task-specific loss while freezing the VLM encoders [67, 66, 13]. Such an approach is parameter-efficient since only a few prompt embeddings are learned while keeping the encoders frozen. In this paper, we are interested in such parameter-efficient prompt learning to steer the VLM toward active object recognition.

3 Preliminaries

Consider a video dataset consisting of verbs \mathcal{V} applied to a base set of objects \mathcal{O}_{base} . The possible set of base actions is $\mathcal{A}_{base} = \mathcal{V} \times \mathcal{O}_{base} = \{(v, o) \mid v \in \mathcal{V}, o \in \mathcal{O}_{base}\}$. During training, only a sparse subset of actions are observed ¹ *i.e.* $\mathcal{A}_{train} \subset \mathcal{A}_{base}$. The goal is to learn a model $f : \mathbb{X}_{train} \rightarrow \mathcal{A}_{base}$ where \mathbb{X}_{train} is the set of videos seen during training. During inference, along with base objects, we also encounter novel objects \mathcal{O}_{novel} that form novel actions $\mathcal{A}_{novel} = \{(v, o) \mid v \in \mathcal{V}, o \in \mathcal{O}_{novel}\}$. We want our model f to predict both base and novel actions *i.e.* $\mathcal{A}_{test} = \mathcal{A}_{base} \cup \mathcal{A}_{novel}$. Conventional action recognition follows a closed vocabulary evaluation where $\mathcal{A}_{test} = \mathcal{A}_{base}$.

4 Method

We train two encoders – one for predicting verbs and another for predicting an open vocabulary of objects as depicted by the pink and yellow blocks in Fig. 2a. During inference, predictions from the verb and object encoders are composed to generate the action prediction.

¹not all unobserved actions are feasible

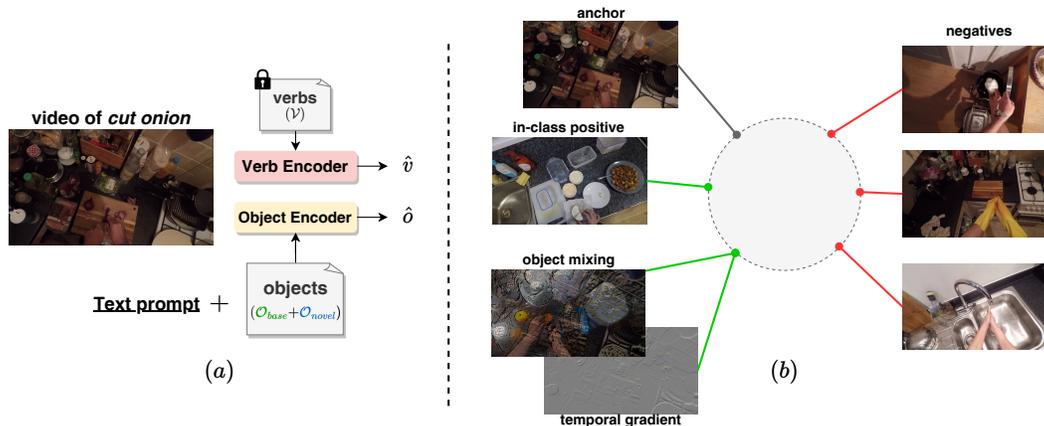


Figure 2: (a) **Framework.** We train separate encoders for predicting verb \hat{v} and object \hat{o} . The verb encoder operates on a closed set of verbs, whereas the object encoder, trained on \mathcal{O}_{base} , can predict any object described by a text prompt. (b) **Object Agnostic Pretraining of the Verb Encoder.** We design an object mixing augmentation to facilitate learning object-agnostic verb representations. Multiple positives are sampled for each anchor video, including our proposed object mixing (mixed with “cut orange”), temporal gradients of the anchor video, and in-class positives *i.e.* videos featuring the same verb (“cut cheese”). The two augmentations corresponding to the same point in the embedding space illustrate that the anchor only needs to minimize its distance from a softmax weighted average of the two (softmax over the similarity of the anchor with the two).

Verb Encoder. Encoding verbs requires spatiotemporal modeling, so we use a video backbone as the verb encoder. We want to learn object-agnostic verb representations to encourage generalization to novel objects. We propose Object Agnostic Pretraining (OAP) (described in Sec. 4.1), which is a general pretraining strategy that can be applied to any video backbone and does not add parameters or increase the model size.

Object Encoder. We aim to recognize an open vocabulary of *active objects*. Even with a closed vocabulary, finding the active object amongst the many other objects in a cluttered scene is difficult. As such, state-of-the-art open vocabulary object detectors [18, 68] perform poorly. One option for support is to spatially locate a region of interaction in the frame; we use a pretrained hand-object interaction (HOI) detector, 100DOH [52] as the base of our object encoder. It provides bounding boxes of the left and right hands and the corresponding object(s)-in-contact in a class-agnostic fashion. Note that the object bounding boxes may contain more than one object (Fig. 3). Thus, we propose an Active Object Prompting (AOP) strategy (described in Sec. 4.2) by leveraging CLIP [46] to generate active object labels from the interaction region for both base and novel classes.

4.1 Object Agnostic Pretraining (OAP)

OAP aims to pull instances with the same verb as the anchor close in the embedding space (irrespective of its object class) and contrast them against instances with different verbs. We randomly sample a batch $B = \{(x_1, y_1), \dots, (x_i, y_i), \dots, (x_N, y_N)\}$ where y_i is the verb class for the raw video clip $x_i \in \mathbb{X}_{train}$ and N is the batch size. We first use

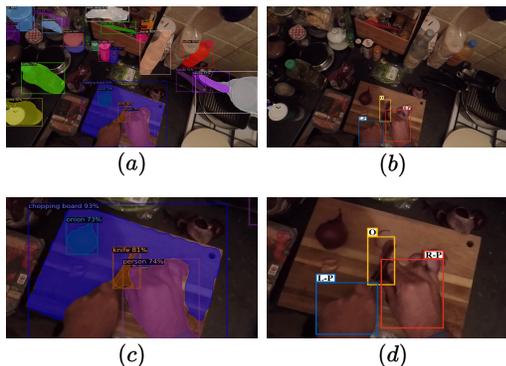


Figure 3: Example frame of “cut onion” from EPIC. (a) DETIC [68] output with EPIC noun vocabulary. (b) HOI Detector [52] output. (c) & (d) are crops of (a) & (b) respectively for better visualization. The object-in-contact crop **O** alongside onion also captures knife.

random photometric augmentations like Gaussian blur and color jittering to create two views of each anchor, thus doubling the batch size to $2N$. Each view is then fed to the verb encoder f_θ followed by a projection head g_θ which generates features for the whole batch $\{z_1, z_2, \dots, z_{2N}\}$. These features are then normalized to the unit hypersphere. The verb encoder is trained with the standard supervised contrastive loss [26]. Videos sampled from B with the same verb as the anchor are considered positives. For each anchor z_i we have the bag of positives $\mathcal{P}_i = \{z_j | y_j = y_i, \forall j \in [1, 2N]\}$. We use the loss \mathcal{L}_{out} with temperature parameter τ_{out} :

$$\mathcal{L}_{out}^{(i)} = -\frac{1}{|\mathcal{P}_i|} \sum_{j \in \mathcal{P}_i} \log \frac{\exp(z_i \cdot z_j / \tau_{out})}{\sum_{k \in B \setminus i} \exp(z_i \cdot z_k / \tau_{out})}. \quad (1)$$

Guiding Augmentations. Sampling in-class positives is still bounded by the verb-object compositions in the training set. Our intuition is that as long as a verb is being performed in a video, exchanging objects and the presence of other distractor objects in the background should not affect the outcome of the verb encoder. To reduce the reliance on such objects for verb prediction, we propose a set of guiding augmentations to diversify the set of positives. The augmentations are lightweight and allow new positives to be generated online during training.

To achieve this, we first mask static information (objects in the background) from a frame of the video clip. The most obvious choice is to use *temporal gradient* that eliminates any static information and focuses primarily on the region with non-zero motion vectors, which we refer to as the active region². In hand-object interactions, this active region is expected to contain the active object. Furthermore, we exploit the compositional nature of actions and propose a new *object mixing* augmentation. The idea is simple – we combine two clips with the same verb but different objects to generate new object mixed clips (Fig. 2b). Formally, given two clips from the batch, x_i and x_j , having the same verb, we get the object mixed clips x'_i and x'_j .

$$\begin{aligned} x'_i &= \alpha(\nabla x_i \odot x_i) + (1 - \alpha)[(\nabla x_j)^{-1} \odot x_j] \\ x'_j &= \alpha(\nabla x_j \odot x_j) + (1 - \alpha)[(\nabla x_i)^{-1} \odot x_i] \end{aligned} \quad (2)$$

where ∇x_i is the first-order temporal gradient of x_i (which masks static regions) and \odot is the element-wise matrix multiplication; $(\nabla x_j)^{-1}$ refers to the inverted gradient image of x_j and α is the mixing hyperparameter. To summarize, masking aims at minimizing the impact of distractor objects outside of the active region, and mixing prevents the model from being biased towards low-level object information, e.g., color, shape, and hence to the observed verb-object compositions.

Learning from Noisy Augmentations. The basis of our guiding augmentations is the temporal gradient. In egocentric videos, the camera (ego) motion will result in non-zero gradients on static parts of the scene in addition to the true active region. While it is feasible to specially design expensive augmentations that account for camera motion [55, 54]; instead, we propose to adjust the contrastive loss so that it can learn from these noisy but lightweight augmentations. In \mathcal{L}_{out} , all the positives for an anchor contribute equally to the loss; this is desirable for in-class positives where the condition for being a positive is well-defined and fixed, but undesirable for the noisy guiding augmentations. Now, if we move the summation inside the *log*, the *logit* for each anchor becomes the average cosine similarity of *all* the guiding augmentations. The contribution of each augmentation is thus dependent on its similarity with the anchor, where noisy augmentations having low cosine similarity will have less contribution to the gradient. With a sufficiently large bag of augmentations, this is expected to reduce the noise implicitly. Such a loss is formulated as \mathcal{L}_{in} in [26]. If \mathcal{G}_i is our bag of guiding augmentations, our \mathcal{L}_{in} loss is as follows:

$$\mathcal{L}_{in}^{(i)} = -\log \frac{\frac{1}{|\mathcal{G}_i|} \sum_{j \in \mathcal{G}_i} \exp(z_i \cdot z_j / \tau_{in})}{\sum_{k \in B \setminus i} \exp(z_i \cdot z_k / \tau_{in})} \quad (3)$$

where τ_{in} is the temperature parameter. This loss is also similar to MIL-NCE [37] for $\tau_{in} = 1$, which is used for learning from noisy video-text pairs. In our case, a lower softmax temperature (< 1)

²active regions may still contain static objects, e.g., for the action of “cutting onions”, the cutting board which is present in the active region is static. This object is not a distractor and is relevant for action prediction.

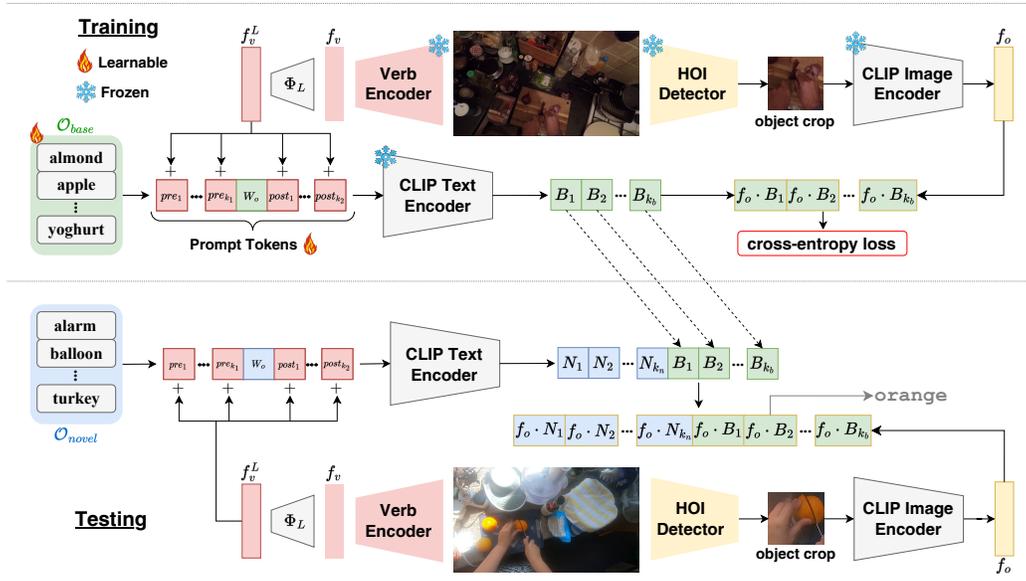


Figure 4: **(Top) Parameter-efficient training for Active Object Prompting.** Keeping the CLIP encoders frozen, the base vocabulary initialized with the CLIP-pretrained word embeddings is fine-tuned on the training set. A set of prefix and postfix prompt vectors conditioned on the verb representation are also learned. **(Bottom) Testing with Active Object Prompting.** The fine-tuned embeddings are used for the base objects, whereas the CLIP-pretrained embeddings are used for novel ones. The learned prefix and postfix prompts are shared among all object classes.

results in better performance. Please refer to the Supplementary for theoretical insights into the loss formulation (Sec. D) and illustrations of guiding augmentations (Sec. F). Thus, the overall loss function is $\mathcal{L} = \mathcal{L}_{out} + \lambda \mathcal{L}_{in}$, where λ is a weighting hyperparameter.

4.2 Active Object Prompting (AOP)

This section discusses leveraging CLIP to predict the active object from a crop with multiple objects. We assume no prior knowledge of the set of objects observed during inference *i.e.* any object is expected to be predicted by our model as long as it is described by a textual phrase (open vocabulary).

The overall framework of AOP is shown in Fig. 4. First, the given set of object classes \mathcal{O} are converted into text prompts by prepending and appending each class with a set of prefix and suffix tokens. For object o_i with word embedding W_{o_i} , a prompt text t_i can be defined as $t_i = \{pre_1, \dots, pre_{k_1}, W_{o_i}, post_1, \dots, post_{k_2}\}$, where pre_j are k_1 prefix tokens, $post_j$ are k_2 postfix tokens. The prefix and postfix prompts, together known as context prompts [67], are essential in guiding a pretrained CLIP to solve a desired task. The context prompt tokens are shared across all classes and can be fixed or learned.

Fixed prompt tokens must be manually engineered by humans, which is laborious and requires domain expertise. Also, prompting CLIP, which operates in a continuous space with discrete human-readable text prompts, limits its usage. Hence, we propose to learn our prompts via two prompting strategies – (i) verb-conditioned context prompt generation and (ii) fine-tuning base object vocabulary.

Verb-Conditioned Context Prompt Generation. As shown in Fig. 3, given a crop containing onion, knife cutting board and hands, we want the CLIP to predict onion as the active object by understanding the context *i.e.* the action of *cut onion*. To this end, we propose verb-conditioned context prompts. Given the video x , we add the verb representation of x learned using OAP to the context prompts as shown in Fig. 4 *i.e.* $pre_j = pre_j + \Phi_L(f_\theta(x))$ and $post_j = post_j + \Phi_L(f_\theta(x))$, where f_θ is the pretrained verb encoder and Φ_L is a projection network used to project the visual feature to the language domain.

Fine-tuning the Base Object Vocabulary. Egocentric videos often capture many different objects in the same environment. Semantic variability of the objects can be very fine-grained e.g. “boom” vs. “arm” in Assembly101 [51]. Such fine granularity is not fully captured by CLIP pretrained on data crawled from the web. We treat this as a vocabulary learning problem with the goal of making fine-grained object categories discriminative in the continuous CLIP feature space. To this end, we fine-tune the pretrained object embeddings W_{o_i} for all base objects seen during training *i.e.* $\forall o_i \in \mathcal{O}_{base}$. This simple fine-tuning allows CLIP to adjust the object embeddings in its continuous space to make them discriminative for optimal performance.

The CLIP-text encoder is kept frozen for both the prompting strategies, making them parameter-efficient. During testing (Fig. 4 bottom), we use the fine-tuned embeddings for the base objects and the original pretrained embeddings for the novel objects. The verb-conditioned context prompts are unique for every instance.

5 Experiments

Datasets. We conduct experiments on three datasets: **EPIC100** [8] features 100 hours of egocentric footage of daily kitchen activities annotated with 97 verbs and 300 interacting objects. **Assembly101** [51] is a multi-view procedural activity dataset featuring 4321 videos of people assembling and disassembling 101 take-apart toy vehicles and is annotated with 24 verbs and 90 objects. We use the egocentric view $e3$ (see [51]), which best captures the hands and interacting objects. **Something-Else** [35], is a human-object interaction subset of SSv2 [16] with 174 verb phrases *e.g.* “picking something up”. Something-Else is not strictly egocentric but shares interesting properties such as diverse hand-object interactions and active objects undergoing state change. We use the compositional split, where all verb phrases are seen during training but applied to novel objects during testing. Interacting objects per video are not annotated; hence, it is unsuitable for our open vocabulary action recognition task but is an ideal testbed for evaluating our verb encoder and OAP with other baselines.

Implementation Details. We use S3D [61] as the verb encoder for the majority of our experiments and ablations as it is a popular architecture for video contrastive learning [20, 37]. For all datasets, the verb encoder is pretrained with the OAP contrastive scheme for 300 epochs and then fine-tuned for verb classification for another 100 epochs. To see how OAP scales with larger backbones, we also evaluate Mformer-B [43] on Something-Else. For the hand-object interaction support, we use the pretrained 100DOH [52] detector without fine-tuning on the target dataset. For the CLIP Image Encoder, we experiment with ViT-B/16 and the larger ViT-L/14 backbones. Keeping both the CLIP image and text encoders frozen, prompts are learned using AOP for 20 epochs. Unless otherwise stated, we choose k_1 and k_2 , the number of learnable prefix and postfix prompts, to be 16 each. For additional implementation details, please refer to Sec. B of the Supplementary.

Table 1: Proposed open vocabulary benchmark splits.

Dataset	\mathcal{V}	\mathcal{O}_{base}	\mathcal{O}_{novel}	#train seg	#test seg.
EPIC100-OV	97	190	110	63k	13.8k
Assembly101-OV	24	60	30	56.2k	15k

5.1 Open Vocabulary Benchmark

Since the original action recognition split of EPIC100 and Assembly101 are unsuitable for open vocabulary (OV) evaluation, we create OV benchmarks EPIC100-OV and Assembly101-OV, respectively, by dividing the complete set of objects into base and novel categories. The base objects are seen during training, whereas the novel objects are kept only for testing. We also incorporate base actions in our test set to evaluate the closed-world performance of the models. The benchmark statistics are provided in Tab. 1. We provide a more detailed explanation of the OV splits in Sec. A of the Supplementary. We use Top-1 Accuracy [8, 51] as our evaluation metric and report results for both closed and novel classes. Additionally, we report the harmonic mean (HM) [38] of the closed and novel accuracies as a consolidated metric for comparing different methods.

Table 2: Compositional Action Recognition on Something-Else [35]. Extra Module refers to if any additional parameterized module has been used on top of the video backbone.

Model	Extra Module	Top-1 (%)	Top-5 (%)
I3D [6]	✗	46.8	72.2
Slowfast [11]	✗	52.2	80.3
Slowfast+IRN [33]	✓	52.9	80.8
STRG,STIN+OIE+NL [35]	✓	56.2	81.3
S3D [61]	✗	49.3	77.3
OAP (S3D)	✗	53.7	80.5
Mformer-B [43]	✗	60.2	85.8
OAP (Mformer-B)	✗	61.8	86.2

Table 3: Comparison of methods for open-vocab object recognition on EPIC100-OV. † Zero-shot evaluation.

Model	Object Top-1 (%)		
	Base	Novel	HM
S3D†	52.2	1.6	3.1
CLIP (ViT-B/16)†	20.2	0.6	1.2
CLIP (ViT-B/16)	7.9	16.3	10.6
DETIc [68]	12.0	2.3	3.9
C2 (ViT-B/16) [25]	19.1	9.9	10.3
C4 (ViT-B/16) [25]	39.9	7.3	12.3
AOP (CLIP ViT-B/16)	41.9	19.2	26.3
AOP (CLIP ViT-L/14)	47.8	22.6	30.7

5.2 Object-Agnostic Pretraining (OAP) vs State-of-the-Art

How effective is OAP in generalizing verbs to novel objects? Tab. 2 shows OAP applied to video backbones S3D [61] and Mformer-B [43] on the Something-Else compositional split. S3D with OAP outperforms Slowfast [11] by 1.5% in Top-1 acc. even though Slowfast is a stronger video backbone. It also outperforms, by 0.8%, the sophisticated Slowfast+IRN [33], which adds an extra transformer to SlowFast to explicitly model the hand-object interactions. Finally, S3D with OAP is still competitive with respect to STRG,STIN+OIE+NL [35], which is an ensemble of various video models, despite being a single backbone without any extra parameters or compute. Mformer-B is a state-of-the-art video architecture comfortably outperforming all other models; adding OAP improves its Top-1 performance by 1.6%, demonstrating that our pretraining is robust to the scale of the backbone.

5.3 Active Object Prompting (AOP) vs. State-of-the-Art

Tab. 3 compares AOP with similar open vocabulary baselines. The first two rows with † refer to zero-shot baselines of S3D and CLIP, where training is done using cross-entropy over all classes (base+novel). CLIP† refers to the linear probing of CLIP’s frozen visual encoder. Both models overfit to the base classes and fail to generalize to novel ones. DETIC [68], a state-of-the-art open vocabulary object detector, improves novel object performance over S3D marginally by 1.3%. It is non-trivial for DETIC to recognize which object is active from a set of per-frame object detections *e.g.* for the action “*beating egg*”, a whole egg is not present in any frame. C4 [25] adapts CLIP for video tasks by adding a transformer layer on top of per-frame CLIP features; again, it overfits to the base classes. Finally, our AOP using both CLIP backbones significantly outperforms all other baselines – AOP improves over the closest competitor (C4) having the same backbone in ViT-B/16 by 14.0% in HM.

5.4 Open Vocabulary Action Recognition

Tab. 4 presents action recognition on open vocabulary splits of EPIC and Assembly101. S3D predicts verbs and objects using two heads but share the same backbone. For a fair comparison with our decoupled verb and object encoders, we also decouple the S3D training, one for predicting verbs and another for objects (reported as 2×S3D). As mentioned in Sec. 5.3, the S3D baseline predicts novel objects in a zero-shot fashion where the novel categories are already assumed to be known during training. Both S3D baselines fail to predict novel objects on both benchmarks. OAP improves verb generalization to novel objects by 1.3% and 4.8% over S3D on EPIC100-OV and Assembly101-OV, respectively. Additionally, the closed-world verb prediction is also enhanced by 1.6% and 7.0% over S3D on EPIC100-OV and Assembly101-OV, respectively. AOP significantly improves novel object prediction performance in both benchmarks. The novel object prediction performance on Assembly101 could be further enhanced by adopting an HOI detector that is robust to domain shifts in the target dataset since Assembly101 egocentric videos are monochrome.

Table 4: Open Vocabulary Action Recognition Results.

Dataset	Model	Closed Top-1 (%)			Novel Top-1 (%)			HM Top-1 (%)		
		Verb	Object	Action	Verb	Object	Action	Verb	Object	Action
EPIC100-OV	S3D	62.5	50.8	37.6	40.1	1.7	0	48.8	3.3	0
	2×S3D	61.5	52.2	36.7	37.9	1.6	0.1	46.9	3.1	0.2
	OAP + AOP	64.1	47.8	35.9	41.4	22.6	11.2	50.3	30.7	17.0
Assembly101-OV	S3D	50.6	34.6	19.9	40.3	0	0	44.9	0	0
	2×S3D	50.7	37.9	20.2	41.8	0	0	45.8	0	0
	OAP + AOP	57.6	30.1	21.6	45.1	6.7	3.7	50.5	11.0	6.6

Table 5: Effect of guiding augmentations for **OAP** on Something-Else.

Loss	Temporal Gradients	Object Mixing	Top-1 (%)	Top-5 (%)
Supervised	✗	✗	49.3	77.3
\mathcal{L}_{out}	✗	✗	50.1	78.1
\mathcal{L}_{out}	✓	✓	52.1	79.3
$\mathcal{L}_{out} + \lambda\mathcal{L}_{in}$	✓	✗	50.3	78.4
$\mathcal{L}_{out} + \lambda\mathcal{L}_{in}$	✓	✓	53.7	80.5

5.5 Ablations

Guiding Augmentations. We study the effect of using guiding augmentations on Something-Else in Table 5. Using \mathcal{L}_{out} already improves performance over supervised training, consistent with the findings in supervised contrastive learning [26, 20]. Introducing verb-specific guidance via temporal gradients and especially our proposed object mixing augmentations boosts the performance further by 3.6%. This demonstrates that OAP with guiding augmentations enhances the generalization of a backbone like S3D without requiring additional parameters.

Active Object Prompting. We thoroughly ablate various learnable prompting strategies in Tab. 6 on EPIC100-OV. ViT-B/16 has been used as the image encoder for all experiments here. For the rows where the context prompts are not learnable, we use the fixed prompt “A picture of a <object>.”. Introducing the HOI detector to generate active object crops improves HM by 3.1% over using full images. Our AOP mainly consists of two proposals – fine-tuning base object vocabulary and conditioning context prompts on verb representations from the video. Most of the gain in Tab. 6 comes from these two proposals. Fine-tuning \mathcal{O}_{base} is expected to improve closed set performance, but we also find 2% improvement over novel categories. This alludes to the fact that learning object names decreases false positives when tasked to predict a novel category. Verb-conditioned prompting brings further improvement of 3.9% in HM. In particular, it boosts the performance over base categories, which is consistent with the findings of conditional prompting learning [66, 67].

Next, we analyze using temporal modeling on top of the CLIP features. Similar to C4 [25], we add a single-layer transformer with eight heads over the per-frame CLIP features. As expected, the model overfits to the base categories seen during training, only achieving 4.8% over novel categories.

Table 6: **AOP ablations on EPIC100-OV.** ViT-B/16 is used as the image encoder. *AOP-Base* and *AOP-Novel* refer to the best-performing models on base and novel object categories.

	HOI Detector	Learnable Context Prompts	Verb Conditioned	Finetune \mathcal{O}_{base}	Temporal modelling	Base	Novel	HM
	✗	✗	✗	✗	✗	5.4	12.1	7.5
	✓	✗	✗	✗	✗	7.9	16.3	10.6
	✓	✗	✗	✓	✗	20.2	18.3	19.2
	✓	✓	✗	✓	✗	21.6	16.6	18.8
<i>AOP-Novel</i>	✓	✓	✓	✓	✗	29.6	18.4	22.7
	✓	✓	✗	✗	✓	46.9	4.8	8.7
<i>AOP-Base</i>	✓	✓	✓	✓	✓	51.0	11.9	19.3
<i>AOP-Base + AOP-Novel</i>	✓	✓	✓	✓	✓	41.9	19.2	26.3

Finally, we ensemble the predictions of the best-performing model on base (*AOP-Base*) and novel (*AOP-Novel*) categories to obtain the best HM.

5.6 Limitations and Outlook

In this paper, we assume that every verb-object composition is feasible. The compositions could be pruned to a feasible subset to reduce the action search space. Obtaining feasibilities, and distinguishing them from common-sense settings, however, is non-trivial, e.g., “*cut thermometer*”, albeit feasible, is unlikely to be observed. Our verb-conditioned prompting of CLIP provides a weak feasibility check. However, CLIP is well known to have a strong object bias [42] over verbs, and it may still yield infeasible active object predictions. One direction to improve the feasibility of actions is to learn object affordances; we leave this for future work.

6 Conclusion

This paper proposed a new task of open vocabulary action recognition for verb-object composed actions and a novel approach to address it. We decouple the verb and object predictions. The verb encoder is trained with an object-agnostic pretraining strategy to improve its generalization to novel objects. Our object encoder leverages a pretrained CLIP model using active object prompting. This guides the CLIP to predict an open vocabulary of interacting objects. Our experiments demonstrate that the proposed approach can recognize hand-object interactions even when the specific actions or objects are not in the training data.

Acknowledgements: This research/project is supported by the National Research Foundation, Singapore and DSO National Laboratories under its AI Singapore Programme (AISG Award No: AISG2-RP-2020-016). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

- [1] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *Advances in Neural Information Processing Systems*, 33:25–37, 2020.
- [2] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*, pages 435–451, 2018.
- [3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.
- [4] Wentao Bao, Qi Yu, and Yu Kong. Evidential deep learning for open set action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13349–13358, 2021.
- [5] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016.
- [6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.
- [7] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. Why can’t i dance in the mall? learning to mitigate scene bias in action recognition. *Advances in Neural Information Processing Systems*, 32, 2019.

- [8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, 130(1):33–55, 2022.
- [9] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. *Advances in Neural Information Processing Systems*, 35:13745–13758, 2022.
- [10] 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.
- [11] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.
- [12] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3299–3309, 2021.
- [13] Kaifeng Gao, Long Chen, Hanwang Zhang, Jun Xiao, and Qianru Sun. Compositional prompt tuning with motion cues for open-vocabulary video relation detection. *arXiv preprint arXiv:2302.00268*, 2023.
- [14] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. Coot: Co-operative hierarchical transformer for video-text representation learning. *Advances in neural information processing systems*, 33:22605–22618, 2020.
- [15] Yuan Gong, Andrew Rouditchenko, Alexander H Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James R Glass. Contrastive audio-visual masked autoencoder. In *The Eleventh International Conference on Learning Representations*, 2022.
- [16] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haanel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The "something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.
- [17] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.
- [18] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2022.
- [19] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.
- [20] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *Advances in Neural Information Processing Systems*, 33:5679–5690, 2020.
- [21] Tengda Han, Weidi Xie, and Andrew Zisserman. Temporal alignment networks for long-term video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2906–2916, 2022.
- [22] Roei Herzig, Elad Ben-Avraham, Karttikeya Mangalam, Amir Bar, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Object-region video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3148–3159, 2022.

- [23] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.
- [24] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020.
- [25] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 105–124. Springer, 2022.
- [26] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschiot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.
- [27] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *Advances in Neural Information Processing Systems*, 31, 2018.
- [28] Matthew Kowal, Mennatullah Siam, Md Amirul Islam, Neil DB Bruce, Richard P Wildes, and Konstantinos G Derpanis. A deeper dive into what deep spatiotemporal networks encode: Quantifying static vs. dynamic information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13999–14009, 2022.
- [29] Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022.
- [30] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z XU, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022.
- [31] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2537–2546, 2019.
- [32] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022.
- [33] Jian Ma and Dima Damen. Hand-object interaction reasoning. In *2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE, 2022.
- [34] Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Active contrastive learning of audio-visual video representations. In *International Conference on Learning Representations*, 2021.
- [35] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1049–1059, 2020.
- [36] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2021.
- [37] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020.

- [38] Tushar Nagarajan and Kristen Grauman. Attributes as operators: factorizing unseen attribute-object compositions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 169–185, 2018.
- [39] Nihal V. Nayak, Peilin Yu, and Stephen Bach. Learning to compose soft prompts for compositional zero-shot learning. In *The Eleventh International Conference on Learning Representations*, 2023.
- [40] Jingcheng Ni, Nan Zhou, Jie Qin, Qian Wu, Junqi Liu, Boxun Li, and Di Huang. Motion sensitive contrastive learning for self-supervised video representation. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 457–474. Springer, 2022.
- [41] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- [42] Jae Sung Park, Sheng Shen, Ali Farhadi, Trevor Darrell, Yejin Choi, and Anna Rohrbach. Exposing the limits of video-text models through contrast sets. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3574–3586, 2022.
- [43] 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. *Advances in neural information processing systems*, 34:12493–12506, 2021.
- [44] Toby Perrett, Saptarshi Sinha, Tilo Burghardt, Majid Mirmehdi, and Dima Damen. Use your head: Improving long-tail video recognition. *arXiv preprint arXiv:2304.01143*, 2023.
- [45] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2021.
- [46] 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.
- [47] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [48] Abhinav Rai, Fadime Sener, and Angela Yao. Transformed rois for capturing visual transformations in videos. *Computer Vision and Image Understanding*, 224:103558, 2022.
- [49] Adria Recasens, Pauline Luc, Jean-Baptiste Alayrac, Luyu Wang, Florian Strub, Corentin Tallec, Mateusz Malinowski, Viorica Pătrăucean, Florent Altché, Michal Valko, et al. Broaden your views for self-supervised video learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1255–1265, 2021.
- [50] 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.
- [51] Fadime Sener, Dibiyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhanian, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022.
- [52] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020.

- [53] Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogério S Feris, David Harwath, James Glass, and Hilde Kuehne. Everything at once-multi-modal fusion transformer for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20020–20029, 2022.
- [54] Vadim Tschernezki, Diane Larlus, and Andrea Vedaldi. Neurdifff: Segmenting 3d objects that move in egocentric videos. In *2021 International Conference on 3D Vision (3DV)*, pages 910–919. IEEE, 2021.
- [55] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013.
- [56] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021.
- [57] Syed Talal Wasim, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Vita-clip: Video and text adaptive clip via multimodal prompting. *arXiv preprint arXiv:2304.03307*, 2023.
- [58] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13587–13597, 2022.
- [59] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.
- [60] Fanyi Xiao, Joseph Tighe, and Davide Modolo. Maclr: Motion-aware contrastive learning of representations for videos. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 353–370. Springer, 2022.
- [61] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018.
- [62] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021.
- [63] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021.
- [64] Lingzhi Zhang, Shenghao Zhou, Simon Stent, and Jianbo Shi. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 127–145. Springer, 2022.
- [65] Chen Zhao, Dawei Du, Anthony Hoogs, and Christopher Funk. Open set action recognition via multi-label evidential learning. *arXiv preprint arXiv:2303.12698*, 2023.
- [66] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022.
- [67] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.
- [68] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 350–368. Springer, 2022.