Rethinking Image-to-Video Adaptation: An Object-centric Perspective

Rui Qian[1], Shuangrui Ding[1], and Dahua Lin[1,2,3⋆]

[1] The Chinese University of Hong Kong, Hong Kong, China
[2] Shanghai Artificial Intelligence Laboratory, Shanghai, China
[3] Centre for Perceptual and Interactive Intelligence (CPII), Hong Kong, China
{qr021,ds023,dhlin}@ie.cuhk.edu.hk

Abstract. Image-to-video adaptation seeks to efficiently adapt image models for use in the video domain. Instead of finetuning the entire image backbone, many image-to-video adaptation paradigms use lightweight adapters for temporal modeling on top of the spatial module. However, these attempts are subject to limitations in efficiency and interpretability. In this paper, we propose a novel and efficient image-to-video adaptation strategy from the object-centric perspective. Inspired by human perception, which identifies objects as key components for video understanding, we integrate a proxy task of object discovery into image-to-video transfer learning. Specifically, we adopt slot attention with learnable queries to distill each frame into a compact set of object tokens. These object-centric tokens are then processed through object-time interaction layers to model object state changes across time. Integrated with two novel object-level losses, we demonstrate the feasibility of performing efficient temporal reasoning solely on the compressed object-centric representations for video downstream tasks. Our method achieves state-of-the-art performance with fewer tunable parameters, only 5% of fully finetuned models and 50% of efficient tuning methods, on action recognition benchmarks. In addition, our model performs favorably in zero-shot video object segmentation without further retraining or object annotations, proving the effectiveness of object-centric video understanding.

Keywords: Image-to-video Adaptation · Object-centric Learning

1 Introduction

Recently, large-scale image pre-training has revolutionized the computer vision community, with the large foundation models dominating various image tasks [34, 56, 69, 78, 87, 94]. The remarkable success of these models can be attributed to the enormous amounts of image data [28, 56] or image-text pairs [34, 69, 78]. However, when we extend this discussion to the video domain, it becomes non-trivial. Acquiring large video datasets and managing high computational

⋆ Corresponding author. Email: dhlin@ie.cuhk.edu.hk

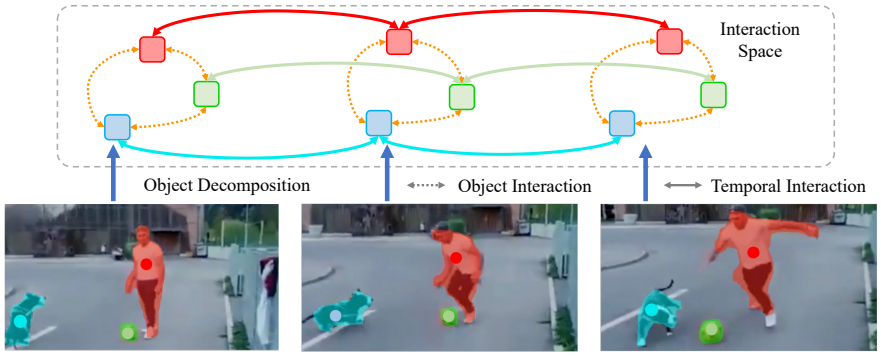


Fig. 1: Illustration of our proposed object-centric video understanding pipeline for image-to-video adaptation. Specifically, we first parse each frame into several object components (represented in different colors) to form an interaction space. Then we respectively establish inter-object interactions within each frame (represented by dotted arrows), and temporal state changes of individual objects (depicted by solid arrows).

costs make training video foundation models from scratch a significant challenge. To circumvent these difficulties, the community has developed an innovative concept dubbed image-to-video adaptation. This approach aims to leverage the power of pre-trained image models and apply them to video data, enhancing both performance and efficiency. The image-to-video adaptation establishes a connection between effective methodologies in image models and the realm of video understanding, thereby opening new avenues for research and exploration.

Motivated by parameter-efficient transfer learning in natural language processing [31,32,41,89], a series of recent studies [36,46,47,55,57,86] have adopted an efficient tuning approach for image-to-video adaptation. These approaches maintain the image pre-trained model in a frozen state and only finetune a minimal set of additional parameters. For example, [47, 55, 57, 86] insert learnable spatio-temporal modules into Transformer encoder blocks. However, this formulation encounters two vital limitations. Firstly, tuning the in-between parameters in the shallow layers still requires a considerable amount of GPU memory during gradient backpropagation. Secondly, the introduction of basic temporal modules, like attention layers, makes temporal reasoning on highly redundant features and lacks the essential inductive bias to effectively benefit the image-to-video adaptation, thereby presenting challenges in interpreting the training's success.

To tackle these challenges, we propose to transfer object knowledge from frozen image pre-trained models to videos for more efficient image-to-video adaptation. In essence, our approach entails two main steps: decoupling each frame into distinct object components, and subsequently identifying object state changes across time to facilitate comprehensive video comprehension as illustrated in Fig. 1. Specifically, we first pass the video frames through the image pre-trained model to obtain frame-wise features. Then, inspired by recent works on object discovery [51,82,83], we employ slot attention [51] with learnable queries [85] to

parse each frame into a compact set of object tokens, serving as compressed representations for each frame. Thereafter, we feed these object tokens into object-time interaction layers to model the object state changes, which vividly depict the temporal dynamics in the video. Finally, we apply a linear head to the latent features representing these object state changes to produce the video-level prediction for action recognition. Unlike the existing works on object-centric representations, our method does not rely on object annotations [30, 80, 92], extra detectors [88] or additional modalities [91]. Instead, we develop two simple object-level losses to assist in distilling effective object-centric representations from image foundation models and establishing meaningful object state changes.

In this way, our method effectively mitigates the limitations of the previous work. Firstly, our method exhibits a notable enhancement in memory efficiency compared to conventional approaches. By exclusively finetuning the additional parameters on top of the image pre-trained model, we successfully reduce the memory required for gradient backpropagation. Additionally, compressing the video frames into object tokens significantly diminishes the computation requirements for temporal modeling, making our method more accessible and scalable. Secondly, by utilizing object-centric reasoning in the form of learnable object tokens and establishing the temporal state changes of individual objects, we inject strong inductive bias into the learning process that goes beyond basic temporal modules like attention layers. We demonstrate that our model can localize the specific objects in the video as a side-product, providing an insightful interpretation of the image-to-video adaptation.

To summarize, our contributions are as follows. (1) We propose an efficient object-centric method to adapt image pre-trained models to the video domain. By parsing video frames into object tokens and establishing object state changes across time, we enhance temporal perception with reduced computation redundancy. (2) We achieve state-of-the-art performance on action recognition, presenting higher efficiency than existing efficient tuning counterparts. (3) Our method effectively discovers different objects in videos without object-level annotations and achieves robust zero-shot video object segmentation.

2 Related Work

Video action recognition is a fundamental problem in computer vision. Compared to image tasks, the extra temporal relationships between different frames are crucial for video understanding. To this end, prevalent methods build on existing image models [20, 29, 49], and make architectural modifications to enable temporal modeling [2, 10, 15, 16, 19, 21, 45, 50, 62, 64, 65, 74, 75]. These works either fully finetune the whole model [2, 5, 10, 21] or directly train on large-scale video data from scratch [66, 73], requiring huge computation cost. More recently, some works propose to recognize actions from the perspective of state changes [1, 54, 59, 71, 79]. This novel idea regards actions as transformations that make changes to the objects and environments. However, most of these works directly model the state changes on the representations of the entire scene, where

different semantics and objects are highly entangled. It is non-trivial to accurately perceive the state changes of different objects. In this work, we explicitly decouple different objects from the video scene, and respectively establish the state changes on individual objects to assist action recognition.

Object-centric video representation learning aims at establishing robust representations for different objects to assist video understanding. To achieve this goal, a series of works borrow the idea from object discovery and employ iterative slot attention [51] to decompose different objects [3, 6, 12, 17, 18, 37, 39, 63, 85]. However, they are specifically trained for segmentation and can only deal with low-level segmentation tasks and lack the ability for high-level understanding. Another line of works uses the object-related features as a complement to the original video representations to enhance object awareness [24, 30, 53, 92, 93]. These methods achieve promising results on multiple action benchmarks [25, 26, 53], but still have two major limitations. First, they depend on bounding box annotations [30,80,92], additional detectors or modalities [88,91] to identify different objects, restricting the scalability to unseen data. Second, the identified objects provide plentiful cues for video analysis, yet these methods still fuse the disentangled object representations with the original video features, imposing limitations on efficiency. In contrast, our process is free of object annotations or handcrafted priors but only adopts simple contrastive loss to distill general object knowledge from image foundation models and decouple object components in video frames. Then we directly model temporal state changes on compact object representations, enabling more intuitive and efficient video understanding.

Efficient tuning becomes a tending direction with the development of large foundation models [7,14,69,94]. It is proposed in natural language processing to reduce the trainable parameters and improve efficiency when transferring pre-trained models to downstream tasks [31, 32, 41, 89]. Common strategies include inserting task-specific adapters into Transformer encoders [31, 60, 61], attaching prompt tokens to the input [42, 48, 67, 70] and learning low-rank approximations [32]. Recently, the efficient tuning is increasingly explored in computer vision [4, 11, 22, 35, 44, 72, 95, 96]. By freezing the image pre-trained model, [11, 72] tune adapters with learnable parameters for efficient adaptation. [4, 22, 35, 44] develop prompt tuning to adapt the frozen image model to various image tasks. Later, [36,46,47,55,57,58,68,77,86] take one step further to adapt image models to videos with temporal dynamics. In order to equip the model with temporal perception ability, [47, 57, 86] insert spatio-temporal modules into Transformer encoder blocks, but the tunable parameters in shallow Transformer layers require huge GPU memory in gradient backpropagation. [58, 68] require two pathways to process spatial and temporal information. [36, 46] attach spatio-temporal attention or motion decoders to the frame-wise features, but there exists high redundancy in attention computation [5,73]. Contrarily, our method retains the backbone architecture without modifications, distills object knowledge from image foundation models, and captures temporal state changes of discovered objects for video downstream tasks.

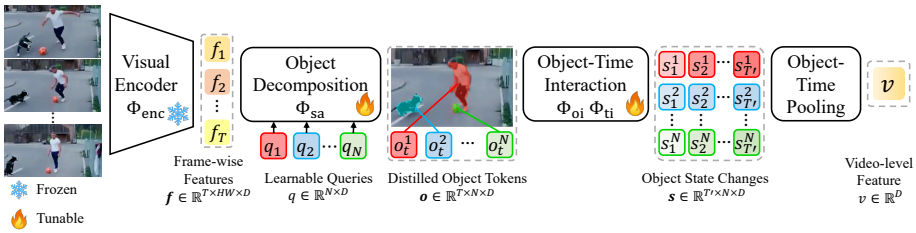


Fig. 2: An overview of our object-centric image-to-video adaptation framework. We use a frozen image pre-trained model to extract frame-wise features and pass them through a lightweight temporal fusion block. Then we employ slot attention with learnable queries to decompose each frame into a compact set of object tokens. Thereafter, we develop object-time interaction layers to establish inter-object interactions and build temporal state changes of individual objects, which are then pooled into the video-level feature for action recognition.

3 Method

The overall architecture of our object-centric image-to-video adaptation is shown in Fig. 2. We use a frozen pre-trained Vision Transformer [20] as the backbone to extract frame-wise features from the video clip. Then we apply slot attention with learnable query tokens to identify different object components and pass the object tokens through object-time interaction layers to model the temporal state changes of individual objects. The obtained object state change vectors encode rich semantics and essential temporal dynamics in the given video, and are subsequently pooled into the final video-level feature for downstream tasks.

3.1 Object Decomposition

To start with, given a video clip $\boldsymbol{x} = \{x_1, x_2, \cdots, x_T\}$ consisting of T frames, we use a frozen image pre-trained Vision Transformer Φ_{enc} as the backbone to extract frame-wise features:

$$p_t, f_t = \Phi_{\text{enc}}(x_t), \quad p_t \in \mathbb{R}^D, \quad f_t \in \mathbb{R}^{HW\times D}, \tag{1}$$

where p_t and f_t denote the output `cls` token and feature map from the last Transformer block. H, W, D respectively denote height, width, and channel dimension. These image pre-trained features encode rich but highly-entangled semantics. And the next step is to parse these frame features into different object components which are crucial for video perception and understanding [24, 30, 80, 92].

Inspired by recent works on unsupervised object discovery and segmentation [51, 82, 83], we develop a variant of slot attention on top of the feature map f_t for object decomposition in each frame. To be specific, instead of random sampling from a prior distribution, we use N learnable query tokens $q \in \mathbb{R}^{N\times D}$ to initialize the slot vectors $S \in \mathbb{R}^{N\times D}$. The intuition of this design is to encourage each query token to capture specific semantics or concepts [33, 83] and produce

temporally coherent object decompositions, which provide essential inductive bias for temporal reasoning in later stages. Then at each slot attention iteration, following [51], we use three learnable linear transformations to project the slot vectors S and frame features f_t into `query`, `key` and `value`, i.e., $Q \in \mathbb{R}^{N\times D}$, $K \in \mathbb{R}^{HW\times D}$ and $V \in \mathbb{R}^{HW\times D}$. After that, we compute the attention matrix $\tilde{A} \in \mathbb{R}^{N\times HW}$ with softmax normalization along the slot dimension as follows:

$$\tilde{A}_{i,j} := \frac{e^{A_{i,j}}}{\sum_{l=1}^{N} e^{A_{l,j}}}, \quad \text{where} \quad A := \frac{1}{\sqrt{D}} QK^T. \tag{2}$$

We use the weighted mean to aggregate the `value` and pass them through Gated Recurrent Unit (GRU) to update the slot vectors:

$$S := \mathrm{GRU}(\texttt{inputs} = MV, \texttt{states} = S), \quad \text{where} \quad M_{i,j} := \frac{\tilde{A}_{i,j}}{\sum_{l=1}^{HW} \tilde{A}_{i,l}}. \tag{3}$$

We iterate the routing process three times, and take the final slot attention weights as potential object segmentation masks. The slot vectors from the final iteration are treated as the object tokens that distill object knowledge from the pre-trained image features. Note that we perform slot attention Φ_{sa} on each frame in parallel, and denote the object tokens of each time stamp as:

$$\boldsymbol{o} = \{o_1 \oplus o_2 \oplus ... \oplus o_T\} \in \mathbb{R}^{T\times N\times D}, \tag{4}$$

where $o_t = \Phi_{\mathrm{sa}}(q, \tilde{f}_t) \in \mathbb{R}^{N\times D}$, and $\oplus$ denotes concatenation operation along time dimension. Since this attention mechanism introduces competition among slots, the obtained object tokens o_t are expected to take over distinct semantic parts and separate each frame into N different components.

Surprisingly, we find that the distilled tokens and corresponding attention matrices perform remarkably well in *zero-shot* video object segmentation tasks. They are even comparable to some pixel-level supervised counterparts, as demonstrated in Section 4.3. This result verifies that our method effectively extracts robust object-centric features for temporal reasoning in the later stage. Note that since videos often contain partial object semantics, there tend to be some slots that carry background information and interact with the foreground object tokens for comprehensive video analysis.

3.2 Object-Time Interaction

After obtaining the object tokens that reveal potential object semantics, we use them as compressed representations of video frames and further model the temporal dynamics from an object-centric perspective [30, 93]. Our intuition is that in continuous frames, the object tokens, $o_t^n \in \mathbb{R}^D, t = 1, 2, ..., T$, with the same index n tend to represent the same objects across time as illustrated in Fig. 1. Hence, it is feasible to temporally track each object and conclude the

object state changes based on the learned object tokens, thereby assisting the process of video understanding.

To achieve this goal, we first apply self-attention to the object tokens within the same frame to introduce inter-object interactions at each time stamp:

$$\tilde{o}_t = \Phi_{\text{oi}}(o_t) \in \mathbb{R}^{N \times D}, \tag{5}$$

where Φ_{oi} is the object interaction module instantiated by a standard Transformer encoder layer consisting of linear projections, self-attention calculation and a feed forward network. The resulting vector $\tilde{o}_t^n$ not only encodes the attribute of object n but also captures its relationships with contextual objects, serving as a representation for the latent state of object n at time stamp t.

Considering that the temporal state changes are a concrete manifestation of the effects of actions, they serve as an important cue for analyzing the events occurring in a video and assisting video understanding [1,59,71,79]. Hence, based on the vectors representing latent object states, we take one step further to look into multiple time stamps and explicitly establish the state changes of each object. Without loss of generality, we take the object with index n for illustration. Given the object states at different time stamps, we define a temporal interval δ to sample the initial and final states, i.e., $\tilde{o}_t^n$ as the initial state and $\tilde{o}_{t+\delta}^n$ as the final state. We concatenate them along channel dimension and pass them through the temporal interaction layer Φ_{ti} to model the object state change between time t and $t+\delta$:

$$s_t^n = \Phi_{\text{ti}}\left(\left[\tilde{o}_t^n \oplus \tilde{o}_{t+\delta}^n\right]\right) \in \mathbb{R}^D, \quad t = \{1, 2, ..., T-\delta\}, \tag{6}$$

where $\oplus$ denotes channel-wise concatenation. We instantiate Φ_{ti} as a multi-layer perceptron of channel dimension $2D$-D-D. In this way, for each object, we obtain $T' = T - \delta$ state change vectors. The state change matrix covering all objects and time stamps can be denoted as $\boldsymbol{s} \in \mathbb{R}^{T' \times N \times D}$.

Since the state change is a continuous process within a video sequence, it is feasible to conclude the overall state change of each object throughout the entire video by simple temporal average pooling on the established matrix $\boldsymbol{s}$. Thereafter, we further apply average pooling on object dimension to aggregate the state change information of all objects as the final video-level representation, $v \in \mathbb{R}^D$, for downstream action recognition.

Discussion. Unlike the existing works that use object features as complements to original video features [24, 30, 53, 92, 93] or model state changes on the highly entangled scene representations [1, 59, 71, 79], we perform temporal reasoning solely on distilled object tokens. Our object-centric state change modeling reduces redundancy, mitigates ambiguity and results in more fine-grained cues for video understanding. And comparing with the prevalent image-to-video adaptation methods, they introduce dense temporal fusion operations, e.g., temporal convolutions [46, 47, 57] and attention [46, 47, 58, 68, 86], to mix multi-frame features for action recognition. In contrast, we leverage the object state changes to perceive temporal dynamics, which closely resemble human logic and provide an effective inductive bias to the model. To be specific, it does not require dense

fusion or traversal of all frames. Instead, we respectively model state changes on separated objects across sparsely sampled frames, enabling more efficient scene parsing and temporal perception as verified in Section 4.4.

3.3 Training

In training, we freeze the parameters in the image pre-trained model Φ_{enc}, and tune parameters that are responsible for object decomposition and temporal perception, i.e., Φ_{temp}, Φ_{oi} and Φ_{ti}. Besides the standard cross-entropy loss for action classification, we introduce two object-level losses to further facilitate object decomposition and temporal state change modeling.

Object Distillation Loss. Considering that the image pre-trained foundation models tend to encode rich semantics in the given scene [9, 56, 69], the output `cls` token p_t is expected to encompass the majority of the semantic components in frame t. Hence, we use a contrastive grounding loss [23, 27, 90] to align the learned object tokens o_t and the `cls` token p_t, encouraging o_t to cover all objects in frame t. Particularly, we first define the correspondence score between the o_t and p_t as:

$$c(o_t, p_t) = \sum_{n=1}^{N} \frac{e^{\langle o_t^n, p_t \rangle / \tau}}{\sum_{l=1}^{N} e^{\langle o_t^l, p_t \rangle / \tau}} \langle o_t^n, p_t \rangle \,, \tag{7}$$

where $\langle o_t^n, p_t \rangle = \frac{o_t^n p_t}{\|o_t^n\| \|p_t\|}$ denotes cosine similarity calculation, τ is the temperature hyper-parameter. This correspondence score adaptively aggregates each object component and avoids penalizing unrelated object semantics. Thereafter, we randomly sample `cls` token vectors p' from other videos within the mini-batch to formulate a negative sample pool $\mathcal{N}$ and apply the contrastive loss to maximize the agreement between the aligned object and `cls` token pairs:

$$\mathcal{L}_{obj} = -\sum_{t=1}^{T} \log \frac{e^{c(o_t, p_t)/\tau}}{e^{c(o_t, p_t)/\tau} + \sum_{p' \sim \mathcal{N}} e^{c(o_t, p')/\tau}}. \tag{8}$$

This self-supervised alignment objective encourages the object tokens to cover diverse semantics and distill object knowledge from the image pre-trained model.

Temporal Reasoning Loss. In our architecture, we analyze temporal dynamics from the perspective of object state changes. Therefore, it is crucial to formulate a learning objective that guides the temporal interaction module Φ_{ti} to accurately capture the state changes between different time stamps. To achieve this goal, our intuition is that the same objects undergoing the same type of actions should possess consistent state changes. Conversely, different actions will result in distinct state changes for the same object, and different objects will exhibit diverse state changes under the same action. Based on this, we develop a margin loss to guide object state change modeling. In detail, given s_t^n as query, we define the videos within the mini-batch that have the same action labels with the current video to form positive pool $\mathcal{P}$, and those of different action labels as

negative pool $\mathcal{N}$. We sample the state change vectors of the same object and the same action, i.e., $\{\hat{s}_t^n | \hat{s} \in \mathcal{P}\}$, as positive pairs, which should be aligned with s_t^n. As for the negatives, we respectively sample the state change vectors of the same object but different actions, i.e., $\{s_t'^n | s' \in \mathcal{N}\}$, and those from the same video but of different objects, i.e., $\{s_t^m, m \neq n\}$, as *hard negatives* for discrimination. The margin loss with a margin hyper-parameter λ is formulated as:

$$\begin{aligned}\mathcal{L}_{temp} = \sum_{t=1}^{T'} \sum_{n=1}^{N} \{ \sum_{\hat{s} \in \mathcal{P}} \| s_t^n - \hat{s}_t^n \|_2 \\ + \sum_{s' \in \mathcal{N}} \max(\lambda - \| s_t^n - s_t'^n \|_2, 0) + \sum_{\substack{m=1 \\ m \neq n}}^{N} \max(\lambda - \| s_t^n - s_t^m \|_2, 0) \},\end{aligned} \tag{9}$$

where $\|\cdot\|_2$ denotes $\mathcal{L}_2$ distance. This term enforces the model to discriminate diverse object states and capture unique dynamic effects of each action category.

Overall Objectives. Besides the above object-level losses, for action recognition tasks, we apply a linear classification head on the video-level representation v with a supervised classification loss $\mathcal{L}_{cls}$ in the form of cross-entropy. In this way, we formulate the overall loss function as:

$$\mathcal{L} = \mathcal{L}_{obj} + \mathcal{L}_{temp} + \mathcal{L}_{cls}. \tag{10}$$

4 Experiments

4.1 Implementation Details

We train and evaluate on three popular action recognition datasets: Kinetics-400 (K-400) [38], Something-Something-v2 (SSv2) [25], and Epic-Kitchens-100 (EK100) [13]. Additionally, we assess zero-shot video object segmentation on challenging multiple object segmentation dataset, DAVIS-2017-UVOS [8]. We use the frozen CLIP [69] pre-trained Vision Transformer [20] as Φ_{enc} for frame-wise feature extraction. In training, we crop each frame into 224×224, set the number of query tokens $N = 8$, and the time interval for object state change $\delta = T/4$ in default. We use AdamW optimizer [52] with an initial learning rate 5×10^{-4} and mini-batch size 512 to update the tunable parameters.

4.2 Video Action Recognition

In this section, we present the comparisons between our method and recent state-of-the-art on K-400, SSv2 and EK100 in terms of the number of tunable parameters, inference GFLOPs and action recognition accuracy.

K-400. We first show the results on K-400 in Table 1 and conclude three observations. (1) Compared to fully finetuned models, our method achieves superior performance with significantly fewer tunable parameters. For example, our method of ViT-L/14 with only 21M tunable parameters outperforms all

Table 1: Results on Kinetics-400. We report the pre-train model initialization, inference GFLOPs, the number of tunable parameters (M), Top-1 and Top-5 accuracy. The 'Views' indicates Frame Number × Temporal Clips × Spatial Crops.

Method	Pre-train	Views	GFLOPs	Tunable	Top-1	Top-5
Full finetuning						
TimeSformer-L [5]	IN-21K	$64 \times 1 \times 3$	7140	121	80.7	94.7
VideoSwin-L [50]	IN-21K	$32 \times 4 \times 3$	7248	197	83.1	95.9
MTV-L [84]	JFT	$32 \times 4 \times 3$	18050	876	84.3	96.3
PromptCLIP A7 [36]	CLIP	$16 \times 5 \times 1$	-	-	76.8	93.5
ActionCLIP [77]	CLIP	$32 \times 10 \times 3$	16890	142	83.8	97.1
X-CLIP-L/14 [55]	CLIP	$8 \times 4 \times 3$	7890	420	87.1	97.6
Efficient tuning						
EVL ViT-L/14 [46]	CLIP	$32 \times 3 \times 1$	8088	59	87.3	-
ST-Adapter ViT-B/16 [57]	CLIP	$32 \times 3 \times 1$	1821	7.2	82.7	96.2
STAN-self-B/16 [47]	CLIP	$16 \times 3 \times 1$	1187	-	84.9	96.8
AIM ViT-B/16 [86]	CLIP	$32 \times 3 \times 1$	2428	11	84.7	96.7
DualPath ViT-B/16 [58]	CLIP	$32 \times 3 \times 1$	710	10	85.4	97.1
DiST ViT-B/16 [68]	CLIP	$32 \times 3 \times 1$	1950	-	85.0	97.0
DiST ViT-L/14 [68]	CLIP	$32 \times 3 \times 1$	8490	-	88.0	97.9
Ours ViT-B/16	CLIP	$8 \times 3 \times 1$	281	6.5	85.8	97.3
Ours ViT-B/16	CLIP	$16 \times 3 \times 1$	562	6.5	86.1	97.5
Ours ViT-B/16	CLIP	$32 \times 3 \times 1$	1123	6.5	86.2	97.5
Ours ViT-L/14	CLIP	$32 \times 3 \times 1$	5068	21	**88.5**	**98.0**

the fully finetuned approaches that require over 100M tunable parameters. (2) Our method with only a visual encoder achieves better results than the multi-modal methods, ActionCLIP [77] and X-CLIP [55], which employ an extra text branch to guide the finetuning process. (3) Among the efficient tuning methods [46, 47, 57, 58, 68, 86], EVL [46] attaches a motion decoder with multiple temporal convolution and attention layers for temporal modeling. AIM [86], ST-Adapter [57] and STAN [47] insert temporal blocks into intermediate CLIP encoder layers. DualPath [58] and DiST [68] introduce an additional temporal pathway to capture temporal dynamics. Our formulation models temporal relationships on decomposed objects in the form of state changes, which reduces computation redundancy but attains superior performance. For example, on ViT-B/16, our method with only an 8-frame input yields superior results to other approaches that employ a 32-frame input. On ViT-L/14, we reach the state-of-the-art results with 40% fewer GFLOPs (5068 vs 8490) than DiST [68].

SSv2. When extending to SSv2 which requires stronger temporal modeling, the observations are consistent. As depicted in Table 2, our object-centric formulation achieves state-of-the-art results and presents higher efficiency than both full finetuning and efficient tuning methods. In particular, our method with ViT-L/14 surpasses the advanced MViTv2-L [43] and DiST [68] with significantly reduced GFLOPs (5068 vs 8484, 8490). Moreover, on ViT-B/16, our method

Table 2: Results on Something-Something-v2. K-400†/K-600† indicates the model is pre-trained on both IN-21K and K-400/K-600.

Method	Pre-train	Views	GFLOPs	Tunable	Top-1	Top-5
Full finetuning						
TimeSformer-L [5]	IN-21K	$64 \times 1 \times 1$	7140	121	62.4	-
MTV-B [84]	IN-21K	$32 \times 4 \times 3$	4790	310	67.6	90.4
ViViT-L/16×2 [2]	K-400†	$16 \times 4 \times 3$	11892	311	65.4	89.8
VideoSwin-B [50]	K-400†	$32 \times 1 \times 1$	963	89	69.6	92.7
MViTv2-L (312↑) [43]	K-400†	$32 \times 1 \times 3$	8484	213	73.3	94.1
Efficient tuning						
EVL ViT-L/14 [46]	CLIP	$32 \times 1 \times 3$	9641	175	66.7	-
ST-Adapter ViT-B/16 [57]	CLIP	$32 \times 3 \times 1$	1955	7.2	69.5	92.6
STAN-self-B/16 [47]	CLIP	$16 \times 3 \times 1$	1376	-	69.5	92.7
AIM ViT-B/16 [86]	CLIP	$32 \times 1 \times 3$	2496	14	69.1	92.2
DualPath ViT-B/16 [58]	CLIP	$32 \times 1 \times 3$	716	13	70.3	92.9
DiST ViT-B/16 [68]	CLIP	$32 \times 3 \times 1$	1950	-	70.9	92.1
DiST ViT-L/14 [68]	CLIP	$32 \times 3 \times 1$	8490	-	73.1	93.2
Ours ViT-B/16	CLIP	$8 \times 1 \times 3$	281	6.5	71.2	93.1
Ours ViT-B/16	CLIP	$16 \times 1 \times 3$	562	6.5	71.6	93.3
Ours ViT-B/16	CLIP	$32 \times 1 \times 3$	1123	6.5	71.8	93.4
Ours ViT-L/14	CLIP	$32 \times 1 \times 3$	5068	21	**73.6**	**94.3**

with only 8 frames as input outperforms others using 32 frames. It is non-trivial to achieve this on such a motion-heavy dataset, which in return reveals the significance of our temporal modeling in the form of object state changes. On the one hand, it enables accurate temporal perception with sparser sampling. On the other hand, our temporal modeling on individual objects eliminates the need for dense temporal fusion, substantially reducing computational redundancy.

EK100. On challenging egocentric EK100 that requires long-term object interactions, our method showcases substantial advantages over existing efficient tunning counterparts [46,57,68] in Table 3. The underlying reasons are two-fold. First, our method involves explicit object decomposition, which helps to capture the primary objects in egocentric videos and contributes to the significant improvements in nouns. Second, our temporal modeling on individual objects reduces the interference of camera motions and enforces the model to focus on the dynamics of each object, thus facilitating verb perception.

4.3 Zero-shot Video Object Segmentation

To verify whether our method understands videos in an object-centric way and interpret why the model works, we evaluate unsupervised video object segmentation in addition to action recognition. Particularly, we use the frozen ViT-B/16 backbone and adopt slot attention weights $\{\tilde{A}_i \in \mathbb{R}^{HW}\}_{i=1}^{N}$ in Eq. 2 as object masks candidate without bells and whistles. We report $\mathcal{J}\&\mathcal{F}$ on challenging

multiple object discovery benchmark, DAVIS-2017-UVOS [8]. Note that we directly apply the model trained on K-400 or SSv2 to video object segmentation in a *zero-shot* manner.

Table 3: Results on Epic-Kitchens-100 Verb, Noun and Action. All compared methods take an input of $8 \times 1 \times 3$.

Method	Verb	Noun	Action
EVL ViT-B/16 [46]	62.7	51.0	37.7
ST-Adapter ViT-B/16 [57]	67.7	55.0	-
DiST ViT-B/16 [68]	69.5	58.1	45.8
DiST ViT-L/14 [68]	70.7	61.6	48.9
Ours ViT-B/16	71.1	60.9	49.0
Ours ViT-L/14	**73.2**	**64.5**	**51.7**

Table 4: Results on multiple object segmentation on DAVIS-2017-UVOS.

Model	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
OCLR [82]	39.6	38.2	41.1
SMTC [63]	40.5	36.4	44.6
BA [17]	43.9	39.2	48.6
RVOS [76]	41.2	36.8	45.7
Ours SSv2	41.0	37.6	44.4
Ours K-400	**44.7**	**42.3**	**47.1**
Ours K-400 CRF	**49.3**	**46.5**	**52.1**

Quantitative Results. We present the comparison with recent state-of-the-arts without object annotations [63, 81, 82] as well as a fully supervised baseline [76] (denoted in grey) in Table 4. Among the compared works without mask annotations, OCLR [82] takes optical flow as input to distinguish motion patterns of diverse objects. SMTC [63] designs two-stage slot attention to produce temporally consistent object segmentations. BA [17] also utilizes image foundation models like DION [9] to provide spatio-temporal correspondence and use clustering to predict the object segmentation masks. On the one hand, since K-400 covers more diverse scenes and objects than SSv2, it leads to better generalization ability in this zero-shot segmentation scenario. On the other hand, our formulation trained on K-400 achieves superior results to these counterparts, even surpassing the simple supervised baseline RVOS [76]. This demonstrates the feasibility of transferring object knowledge from image foundation models to the video domain, and our object-centric temporal reasoning further facilitates temporally coherent object decomposition. However, the gap to the state-of-the-art supervised method is still large because the mask annotations lead to much more accurate segmentation boundaries. A straightforward way is to apply Conditional Random Field (CRF) [40] for refinement, which leads to 4.6 points improvement on $\mathcal{J}\&\mathcal{F}$ score. It is worth further exploration in the future.

Qualitative Results. We also visualize some examples of the generated segmentation masks in Fig. 3, where each column presents the frames within the same video and the same color denotes the masks generated by the same slot query. Generally, our model discriminates different object semantics in videos. Each distilled object token can capture specific semantics or concepts that generalize across diverse time stamps and scenes. For example, the token marked in red focuses on humans, the blue one emphasizes quadruped animals, and the yellow one identifies two-wheeled vehicles. This temporally coherent object

decomposition provides an effective inductive bias for subsequent temporal reasoning, leading to more efficient temporal modeling.

4.4 Ablation Study

In this section, we use the CLIP ViT-B/16 as Φ_{enc} with an input of 16 frames for all ablative experiments unless otherwise specified. More details and extra ablation studies are presented in Supplementary Materials.

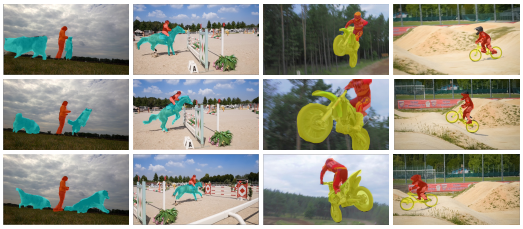


Fig. 3: Visualization of object decomposition in the form of segmentation. Each column presents three frames in a video. The same color denotes the objects identified by the same object token.

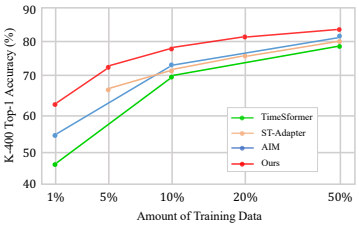


Fig. 4: Comparison on training data efficiency. Our method demonstrates advantages, especially with a small amount of training data.

Training Losses. We explore the effectiveness of our newly introduced object distillation loss $\mathcal{L}_{obj}$ and temporal reasoning loss $\mathcal{L}_{temp}$ in Table 5. We observe that $\mathcal{L}_{obj}$ leads to dramatic improvements on all three datasets, K-400, SSv2, and DAVIS-2017-UVOS. This is because $\mathcal{L}_{obj}$ offers rich semantic knowledge reference from the image foundation model to facilitate object identification. And better object decomposition further enhances high-level action recognition. As for $\mathcal{L}_{temp}$, it guides the model to capture the state changes on each object. Without this guidance, relying solely on the action classification loss is insufficient to direct the temporal module to perceive representative temporal dynamics, thus resulting in a drastic performance drop on the motion-heavy dataset SSv2. And notably, this temporal loss also leads to improvements on DAVIS, which verifies that the temporal perception in return bolsters object decomposition.

Temporal Sampling Strategy. We also study the effect of different temporal sampling strategies in training. Specifically, we vary the number of frames per clip T and the time interval for object state change modeling δ. In Table 6, we display the Top-1 accuracy on K-400 and SSv2, as well as the computation GFLOPs excluding ViT backbone for more intuitive comparison. An exciting phenomenon is that even on SSv2 that requires strong temporal modeling, we achieve promising results with very sparse sampling, i.e., only 4 frames as input with significantly reduced computation. Our method is generally robust to the temporal sampling hyper-parameters, demonstrating the effectiveness of our temporal reasoning on compressed object tokens.

Table 5: Ablation studies on training loss items $\mathcal{L}_{obj}$ and $\mathcal{L}_{temp}$. Note that $\mathcal{L}_{cls}$ always applied in training.

$\mathcal{L}_{obj}$	$\mathcal{L}_{temp}$	K-400	SSv2	DAVIS
✗	✗	70.4	44.8	27.3
✓	✗	80.3	55.2	39.5
✗	✓	73.5	52.1	30.5
✓	✓	86.1	71.6	44.7

Table 6: Ablation studies on the number of frames T and sampling interval δ.

T	δ	GFLOPs	K-400	SSv2
4	1	4.2	85.4	70.7
8	2	8.4	85.8	71.2
8	4	8.2	85.8	71.1
16	2	18.0	86.1	71.6
16	4	17.8	86.1	71.6

Training Data Efficiency. Fig. 4 presents K-400 action recognition performance with various amounts of training data. For fair comparison with fully fine-tuned TimeSformer [5] as well as the efficient tuning methods, ST-Adapter [57] and AIM [86], we use CLIP ViT-B/16 with an input of 8 frames. Our method exhibits superior data efficiency to existing works. And the advantage becomes more significant with less training data, demonstrating the robust capacity of our object-centric formulation to transfer image pre-trained knowledge to videos.

5 Conclusion

In this work, we propose an object-centric formulation to adapt the image pre-trained models to the video domain. Specifically, we employ a frozen image pre-trained model to extract frame-wise features and use slot attention with learnable queries to parse each frame into a compact set of object tokens. Then we apply object-time interaction to these object tokens to explicitly establish object state changes across time span. By using the distilled object tokens as the compressed representations of video frames, we have reduced computation costs and improved temporal reasoning capabilities especially under sparse sampling scenarios. Our method demonstrates state-of-the-art performance across action recognition benchmarks while realizing robust zero-shot video object segmentation, signifying a promising direction for future research in video analysis.

Limitations and Social Negative Impact. Our work demonstrates that the object-centric formulation serves as a reliable compression of video frames, which reduces redundancy and enables efficient temporal perception. But in some applications like a video QA system, it would be better to distill useful information according to specific questions. Incorporating questions as a condition to slot attention might address this problem. And there is a potential negative impact in privacy concerns. This work involves processing large amounts of visual data, which could include sensitive information and raise privacy concerns.

Acknowledgements

This work is partially supported by the National Key R&D Program of China (2022ZD0160201), Shanghai Artificial Intelligence Laboratory, the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Commission (ITC)'s InnoHK.

References

1. Alayrac, J.B., Miech, A., Laptev, I., Sivic, J., et al.: Multi-task learning of object states and state-modifying actions from web videos. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
2. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6836–6846 (2021)
3. Aydemir, G., Xie, W., Guney, F.: Self-supervised object-centric learning for videos. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=919tWtJPXe
4. Bahng, H., Jahanian, A., Sankaranarayanan, S., Isola, P.: Visual prompting: Modifying pixel space to adapt pre-trained models. arXiv preprint arXiv:2203.17274 (2022)
5. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: ICML. vol. 2, p. 4 (2021)
6. Besbinar, B., Frossard, P.: Self-supervision by prediction for object discovery in videos. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 1509–1513. IEEE (2021)
7. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
8. Caelles, S., Pont-Tuset, J., Perazzi, F., Montes, A., Maninis, K.K., Van Gool, L.: The 2019 davis challenge on vos: Unsupervised multi-object segmentation. arXiv preprint arXiv:1905.00737 (2019)
9. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9650–9660 (2021)
10. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
11. Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J., Luo, P.: Adaptformer: Adapting vision transformers for scalable visual recognition. Advances in Neural Information Processing Systems **35**, 16664–16678 (2022)
12. Crawford, E., Pineau, J.: Exploiting spatial invariance for scalable unsupervised object tracking. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 3684–3692 (2020)
13. Damen, D., Doughty, H., Farinella, G.M., Furnari, A., Kazakos, E., Ma, J., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Rescaling egocentric vision. arXiv preprint arXiv:2006.13256 (2020)
14. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

15. Ding, S., Li, M., Yang, T., Qian, R., Xu, H., Chen, Q., Wang, J., Xiong, H.: Motion-aware contrastive video representation learning via foreground-background merging. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9716–9726 (2022)
16. Ding, S., Qian, R., Xiong, H.: Dual contrastive learning for spatio-temporal representation. In: Proceedings of the 30th ACM international conference on multimedia. pp. 5649–5658 (2022)
17. Ding, S., Qian, R., Xu, H., Lin, D., Xiong, H.: Betrayed by attention: A simple yet effective approach for self-supervised video object segmentation. arXiv preprint arXiv:2311.17893 (2023)
18. Ding, S., Xie, W., Chen, Y., Qian, R., Zhang, X., Xiong, H., Tian, Q.: Motion-inductive self-supervised object discovery in videos. arXiv preprint arXiv:2210.00221 (2022)
19. Ding, S., Zhao, P., Zhang, X., Qian, R., Xiong, H., Tian, Q.: Prune spatio-temporal tokens by semantic-aware temporal accumulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16945–16956 (2023)
20. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
21. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6202–6211 (2019)
22. Feng, C., Zhong, Y., Jie, Z., Chu, X., Ren, H., Wei, X., Xie, W., Ma, L.: Promptdet: Towards open-vocabulary detection using uncurated images. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX. pp. 701–717. Springer (2022)
23. Ghiasi, G., Gu, X., Cui, Y., Lin, T.Y.: Scaling open-vocabulary image segmentation with image-level labels. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI. pp. 540–557. Springer (2022)
24. Girdhar, R., Carreira, J., Doersch, C., Zisserman, A.: Video action transformer network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 244–253 (2019)
25. Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The" something something" video database for learning and evaluating visual common sense. In: Proceedings of the IEEE international conference on computer vision. pp. 5842–5850 (2017)
26. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., et al.: Ava: A video dataset of spatio-temporally localized atomic visual actions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6047–6056 (2018)
27. Gupta, T., Vahdat, A., Chechik, G., Yang, X., Kautz, J., Hoiem, D.: Contrastive learning for weakly supervised phrase grounding. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III. pp. 752–768. Springer (2020)
28. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022)

29. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
30. Herzig, R., Ben-Avraham, E., Mangalam, K., Bar, A., Chechik, G., Rohrbach, A., Darrell, T., Globerson, A.: Object-region video transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3148–3159 (2022)
31. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International Conference on Machine Learning. pp. 2790–2799. PMLR (2019)
32. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
33. Jia, B., Liu, Y., Huang, S.: Unsupervised object-centric learning with bi-level optimized query slot attention. arXiv preprint arXiv:2210.08990 (2022)
34. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning. pp. 4904–4916. PMLR (2021)
35. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII. pp. 709–727. Springer (2022)
36. Ju, C., Han, T., Zheng, K., Zhang, Y., Xie, W.: 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. pp. 105–124. Springer (2022)
37. Kabra, R., Zoran, D., Erdogan, G., Matthey, L., Creswell, A., Botvinick, M., Lerchner, A., Burgess, C.: Simone: View-invariant, temporally-abstracted object representations via unsupervised video decomposition. Advances in Neural Information Processing Systems **34**, 20146–20159 (2021)
38. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
39. Kipf, T., Elsayed, G.F., Mahendran, A., Stone, A., Sabour, S., Heigold, G., Jonschkowski, R., Dosovitskiy, A., Greff, K.: Conditional object-centric learning from video. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=aD7uesX1GF_
40. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001)
41. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. arXiv preprint arXiv:2104.08691 (2021)
42. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:2101.00190 (2021)
43. Li, Y., Wu, C.Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., Feichtenhofer, C.: Mvitv2: Improved multiscale vision transformers for classification and detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4804–4814 (2022)
44. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. arXiv preprint arXiv:2210.04150 (2022)

45. Lin, Z., Wu, Y.F., Peri, S.V., Sun, W., Singh, G., Deng, F., Jiang, J., Ahn, S.: Space: Unsupervised object-oriented scene representation via spatial attention and decomposition. arXiv preprint arXiv:2001.02407 (2020)
46. Lin, Z., Geng, S., Zhang, R., Gao, P., De Melo, G., Wang, X., Dai, J., Qiao, Y., Li, H.: Frozen clip models are efficient video learners. In: European Conference on Computer Vision. pp. 388–404. Springer (2022)
47. Liu, R., Huang, J., Li, G., Feng, J., Wu, X., Li, T.H.: Revisiting temporal modeling for clip-based image-to-video knowledge transferring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6555–6564 (June 2023)
48. Liu, X., Ji, K., Fu, Y., Tam, W.L., Du, Z., Yang, Z., Tang, J.: P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. arXiv preprint arXiv:2110.07602 (2021)
49. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
50. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3202–3211 (2022)
51. Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., Kipf, T.: Object-centric learning with slot attention. Advances in Neural Information Processing Systems **33**, 11525–11538 (2020)
52. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018)
53. Materzynska, J., Xiao, T., Herzig, R., Xu, H., Wang, X., Darrell, T.: Something-else: Compositional action recognition with spatial-temporal interaction networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1049–1059 (2020)
54. Nagarajan, T., Grauman, K.: Attributes as operators: factorizing unseen attribute-object compositions. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 169–185 (2018)
55. Ni, B., Peng, H., Chen, M., Zhang, S., Meng, G., Fu, J., Xiang, S., Ling, H.: Expanding language-image pretrained models for general video recognition. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV. pp. 1–18. Springer (2022)
56. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
57. Pan, J., Lin, Z., Zhu, X., Shao, J., Li, H.: St-adapter: Parameter-efficient image-to-video transfer learning for action recognition. arXiv preprint arXiv:2206.13559 (2022)
58. Park, J., Lee, J., Sohn, K.: Dual-path adaptation from image to video transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2203–2213 (June 2023)
59. Peh, E., Parmar, P., Fernando, B.: Learning to visually connect actions and their effects. arXiv preprint arXiv:2401.10805 (2024)
60. Pfeiffer, J., Kamath, A., Rücklé, A., Cho, K., Gurevych, I.: Adapterfusion: Non-destructive task composition for transfer learning. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 487–503 (2021)

61. Pfeiffer, J., Rücklé, A., Poth, C., Kamath, A., Vulić, I., Ruder, S., Cho, K., Gurevych, I.: Adapterhub: A framework for adapting transformers. arXiv preprint arXiv:2007.07779 (2020)
62. Qian, R., Ding, S., Liu, X., Lin, D.: Static and dynamic concepts for self-supervised video representation learning. In: European Conference on Computer Vision. pp. 145–164. Springer (2022)
63. Qian, R., Ding, S., Liu, X., Lin, D.: Semantics meets temporal correspondence: Self-supervised object-centric learning in videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16675–16687 (2023)
64. Qian, R., Dong, X., Zhang, P., Zang, Y., Ding, S., Lin, D., Wang, J.: Streaming long video understanding with large language models. arXiv preprint arXiv:2405.16009 (2024)
65. Qian, R., Li, Y., Liu, H., See, J., Ding, S., Liu, X., Li, D., Lin, W.: Enhancing self-supervised video representation learning via multi-level feature optimization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7990–8001 (2021)
66. Qian, R., Meng, T., Gong, B., Yang, M.H., Wang, H., Belongie, S., Cui, Y.: Spatiotemporal contrastive video representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6964–6974 (2021)
67. Qin, G., Eisner, J.: Learning how to ask: Querying lms with mixtures of soft prompts. arXiv preprint arXiv:2104.06599 (2021)
68. Qing, Z., Zhang, S., Huang, Z., Zhang, Y., Gao, C., Zhao, D., Sang, N.: Disentangling spatial and temporal learning for efficient image-to-video transfer learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 13934–13944 (October 2023)
69. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
70. Shin, T., Razeghi, Y., Logan IV, R.L., Wallace, E., Singh, S.: Autoprompt: Eliciting knowledge from language models with automatically generated prompts. arXiv preprint arXiv:2010.15980 (2020)
71. Souček, T., Alayrac, J.B., Miech, A., Laptev, I., Sivic, J.: Look for the change: Learning object states and state-modifying actions from untrimmed web videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13956–13966 (2022)
72. Sung, Y.L., Cho, J., Bansal, M.: Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5227–5237 (2022)
73. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in neural information processing systems **35**, 10078–10093 (2022)
74. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 4489–4497 (2015)
75. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018)

76. Ventura, C., Bellver, M., Girbau, A., Salvador, A., Marques, F., Giro-i Nieto, X.: Rvos: End-to-end recurrent network for video object segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5277–5286 (2019)
77. Wang, M., Xing, J., Liu, Y.: Actionclip: A new paradigm for video action recognition. arXiv preprint arXiv:2109.08472 (2021)
78. Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., et al.: Image as a foreign language: Beit pretraining for vision and vision-language tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19175–19186 (2023)
79. Wang, X., Farhadi, A., Gupta, A.: Actions˜ transformations. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 2658–2667 (2016)
80. Wang, X., Gupta, A.: Videos as space-time region graphs. In: Proceedings of the European conference on computer vision (ECCV). pp. 399–417 (2018)
81. Wang, X., Misra, I., Zeng, Z., Girdhar, R., Darrell, T.: Videocutler: Surprisingly simple unsupervised video instance segmentation. arXiv preprint arXiv:2308.14710 (2023)
82. Xie, J., Xie, W., Zisserman, A.: Segmenting moving objects via an object-centric layered representation. In: Advances in Neural Information Processing Systems (2022)
83. Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., Wang, X.: Groupvit: Semantic segmentation emerges from text supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18134–18144 (2022)
84. Yan, S., Xiong, X., Arnab, A., Lu, Z., Zhang, M., Sun, C., Schmid, C.: Multiview transformers for video recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3333–3343 (2022)
85. Yang, C., Lamdouar, H., Lu, E., Zisserman, A., Xie, W.: Self-supervised video object segmentation by motion grouping. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7177–7188 (October 2021)
86. Yang, T., Zhu, Y., Xie, Y., Zhang, A., Chen, C., Li, M.: AIM: Adapting image models for efficient video action recognition. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=CIoSZ_HKHS7
87. Yuan, L., Chen, D., Chen, Y.L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al.: Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432 (2021)
88. Zadaianchuk, A., Kleindessner, M., Zhu, Y., Locatello, F., Brox, T.: Unsupervised semantic segmentation with self-supervised object-centric representations. arXiv preprint arXiv:2207.05027 (2022)
89. Zaken, E.B., Ravfogel, S., Goldberg, Y.: Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. arXiv preprint arXiv:2106.10199 (2021)
90. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-vocabulary object detection using captions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14393–14402 (2021)
91. Zhang, C., Fu, C., Wang, S., Agarwal, N., Lee, K., Choi, C., Sun, C.: Object-centric video representation for long-term action anticipation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6751–6761 (2024)

92. Zhang, C., Gupta, A., Zisserman, A.: Is an object-centric video representation beneficial for transfer? In: Proceedings of the Asian Conference on Computer Vision. pp. 1976–1994 (2022)
93. Zhang, Y., Tokmakov, P., Hebert, M., Schmid, C.: A structured model for action detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9975–9984 (2019)
94. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832 (2021)
95. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16816–16825 (2022)
96. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)