Video Action Recognition with Attentive Semantic Units

Yifei Chen[1]   Dapeng Chen[1†]   Ruijin Liu[2⋆]   Hao Li[3⋆]   Wei Peng[1]

[1] IIRC, Huawei   [2] Xi'an Jiaotong University   [3] Xiamen University
chenyifei14@huawei.com   chendapeng8@huawei.com

Abstract

Visual-Language Models (VLMs) have significantly advanced video action recognition. Supervised by the semantics of action labels, recent works adapt the visual branch of VLMs to learn video representations. Despite the effectiveness proved by these works, we believe that the potential of VLMs has yet to be fully harnessed. In light of this, we exploit the semantic units (SU) hiding behind the action labels and leverage their correlations with fine-grained items in frames for more accurate action recognition. SUs are entities extracted from the language descriptions of the entire action set, including body parts, objects, scenes, and motions. To further enhance the alignments between visual contents and the SUs, we introduce a multi-region attention module (MRA) to the visual branch of the VLM. The MRA allows the perception of region-aware visual features beyond the original global feature. Our method adaptively attends to and selects relevant SUs with visual features of frames. With a cross-modal decoder, the selected SUs serve to decode spatiotemporal video representations. In summary, the SUs as the medium can boost discriminative ability and transferability. Specifically, in fully-supervised learning, our method achieved 87.8% top-1 accuracy on Kinetics-400. In K=2 few-shot experiments, our method surpassed the previous state-of-the-art by +7.1% and +15.0% on HMDB-51 and UCF-101, respectively.

1. Introduction

Video action recognition is the fundamental task toward intelligent video understanding. Facilitated by deep learning, great advancements have been made in designing end-to-end network architecture, including two-stream networks [50, 54, 67], 3D convolutional neural networks (3D-CNN) [9, 16, 17, 23, 46, 52, 53, 59] and transformer-based

networks [4, 14, 35, 44, 62]. The visual-languages models (VLM)[47, 25] have achieved great success in the last few years. Taking inspiration from this, recent works[55, 26, 41] started exploiting video representation learning under semantic supervision. In particular, these works propose a paradigm that trains video representations to align with action-name text embeddings, which provide richer semantic and contextual information than conventional one-hot labels. Besides showing competitive performance in closed-set action recognition scenarios, adopting VLM demonstrates great learning effectiveness toward recognizing unseen or unfamiliar categories compared to methods that do not utilize text embeddings.

Despite the achieved improvements, we believe that relying solely on the semantic information of action labels is far from enough to fully harness the advantages of the VLMs. This is because action labels are abstract concepts, and directly mapping a video clip to them may confront a noticeable gap. We propose alleviating the problem by relating the action labels with some common entity, ranging from static scenes to moving body parts. For example, "playing golf" can be associated with people, golf equipment, and grass, which could provide fine-grained and specific guidance for visual alignments. We refer to these action-related basic entities as *semantic units* (SU). In practice, SUs are extracted from discriminative language descriptions of the whole action set, including body parts, objects, scenes, and motions. Our basic idea is to use visual features to select related SUs and then use SUs to guide the decoding of spatiotemporal video features. Compared with the previous approaches, utilizing SUs can bring two advantages: (1) The text corresponding to semantic units can better explain which factors determine an action. (2) The re-usability of semantic units can greatly alleviate learning difficulty in the zero-shot/few-shot scenario.

Another concern arises when we explore fine-grained visual-language correlations through SUs: the visual branch of current VLMs produces frame-level representations, which may hinder the sensitive correspondences to SUs. To

---

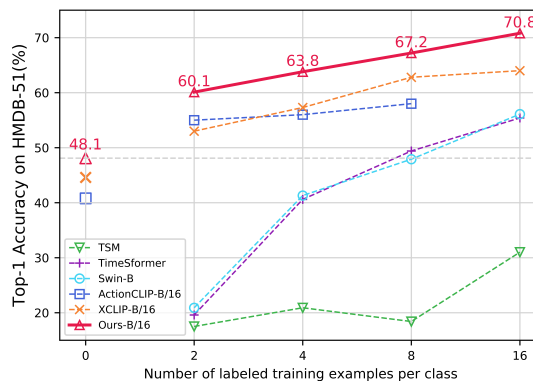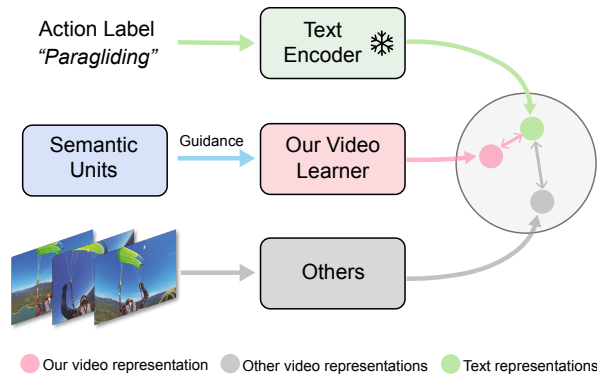† Corresponding author. ⋆ Ruijin Liu and Hao Li were interns at Huawei during the project.

Figure 1. Left: Compared with conventional video learners that encode frames directly, our method utilizes attentive semantic units (SU) to guide the representation learning of videos. The learned representation contains rich semantics and thus can be effectively adapted to the target representation given by the text representation of the action name. **Right:** With the utilization of semantic units (SU), our method can achieve impressive results on zero/few-shot learning. With the model trained on Kinetics-400, we achieve 48.1% and 70.8% top-1 accuracy in zero-shot and K=16 few-shot experiments, which is 3.5% and 6.6% higher than the second best methods.

leverage the local visual appearance, we further introduce multi-region attention (MRA) modules upon the pre-trained visual encoders. MRA spatially perceives sub-regions of each image and enhances the visual features with region-aware info, therefore enabling visual features to select SUs globally and locally. The final video representation is produced by a cross-modal video decoder with the selected SUs as the queries and the enhanced visual features from frames as the keys and values. The cross-modal decoder is a stack of cross-attention and 1D temporal convolution hybrid modules. Cross-attention makes the representation focus on the most critical visual appearance by attentive semantic units, while the temporal convolution exploits the motion cue for video representation. In summary, our contributions are threefold: (1) We utilize semantic units to guide the representation learning of actions. Because the semantic units are fine-grained and reusable, the learned representation can improve both the discriminative ability and the transferability. (2) We introduce multi-region attention (MRA) to the visual branch of VLM, which originally provided global representations. MRA perceives region-aware features for each image, enabling sensitive correspondence to multiple fine-grained semantic cues behind actions. (3) We propose a cross-modal decoder that generates the final video representation. This module utilizes attentive semantic units to highlight critical visual appearances while exploiting motion cues.

2. Related Work

Network Architectures in Action Recognition With the significant advances in deep learning, the current video recognition approaches aim to learn effective spatial-temporal representation by well-designed network architectures. The two-stream networks apply 2D CNNs in video recognition [67, 54, 50], where they combine the color im-

age and optical flow to obtain inter-frame cues. The availability of larger video classification datasets ,such as Kinetics [27], facilitates the training of 3D CNNs [16, 53, 46, 52, 59, 17, 16]. 3D CNN factorizes the convolution across spatial and temporal dimensions and can capture temporal variation among the local neighborhood. The recent transformer-based approaches advance the state-of-the-art previously set by 3D CNNs, including ViViT [3], Timesformer [4], Video Swin Transformer [35], and Multi-view Transformer [62]. Because of the self-attention modules, the transformer-based approaches can capture the long-term spatial-temporal dependencies, leading to better performance. Our method also employs Transformers as backbones: a VIT-like network structure to extract the frame-and-region-level representations and a decoder with semantic-unit attention.

Vision-Language Model in Action Recognition Vision-language models (VLM) are making rapid progress and impressive performance on multiple downstream tasks, including detection [22], segmentation [57], caption [39], summarization [40], etc. There is also a trend to extend VLMs to video action recognition. VideoCLIP [60] replaces image-text with video-text pairs to formulate a video-language pretraining for the downstream action recognition task. However, such an "training from scratch" strategy is data-hungry and computationally expensive. Another pipeline is to finetune upon the available vision-language pre-trained model such as CLIP [47] and ALIGN [25]. ActionCLIP [55] firstly proposes a "pre-train, prompt and finetune" framework. Prompting-CLIP [26] adapts different downstream tasks by prompting learning. X-CLIP [41] introduces a cross-frame attention mechanism and utilizes the video content embedding to generate the text prompts, which is used to enhance the text embedding. To improve training efficiency, EVL [34] and ST-Adapter [43] fix the CLIP backbone and add a lightweight network module
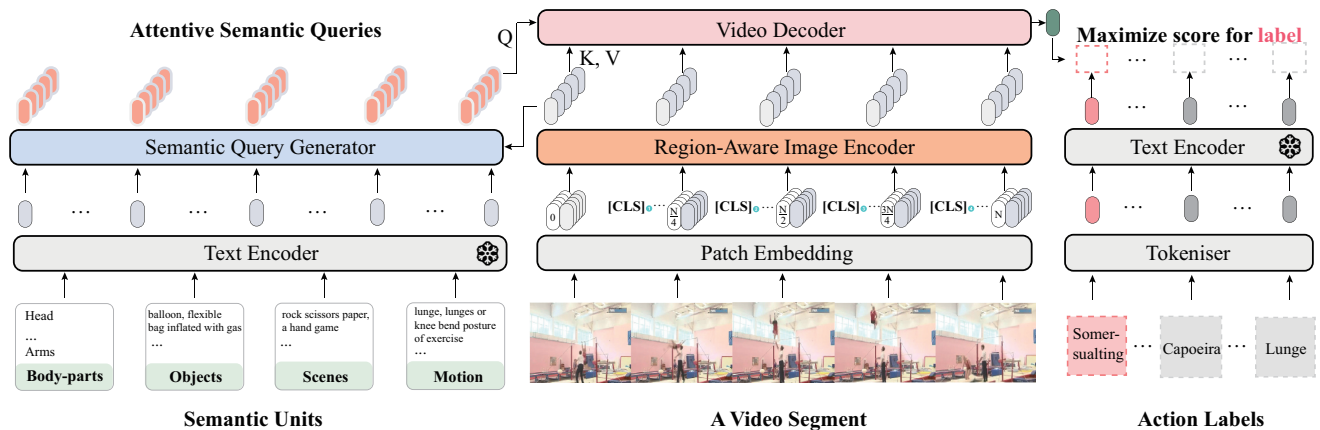
Figure 2. An overview of our framework. We input semantic units and a video segment and output a video representation(in dark green). The video representation is supervised by the goal of maximizing the similarity score with the text feature of the correct action label.

to exploit the spatiotemporal information. Unlike all the above methods that leverage semantics by direct mappings, *i.e.* from video embedding to text features of action label names, our approach further utilizes the capability of pretrained VLMs by introducing fine-grained semantic units to guide the learning of video representations.

Human-Object Interaction in Action Recognition The interactions between the object and the human are essential for action recognition. It was first studied for still images to resolve the ambiguities where two actions have almost identical poses. Towards action recognition tasks, the HOI methods develop a fine-grained manner based on bounding boxes of objects, humans, and scenes [12, 15, 21, 32]. For example, Fang et al. [15] explicitly consider the body parts and their pairwise correlations, then fuse the body representations with the scene and objects. The idea of HOI is then applied to the video data by considering the motion cues. Something-else [37] mainly works on the dynamic interactions between hands and objects. ActionGenome [24] represents an action by a spatiotemporal scene graph, describing the person-object relationships. AKU [36] infer actions from body part movements with the assistance of visual-driven semantic knowledge mining. Our inspiration for exploring action-related entities comes from the above works. However, in contrast to their reliance on detectors for localization and recognition, our approach efficiently leverages text features of semantic units to select the body parts/instances in frames.

3. Methodology

The overall framework of our method is illustrated in Fig. 2. Our method aims to learn more transferable and expressive video representation for action recognition by involving "semantic units" (SU) for guidance. We generate and embed semantic units (Section 3.1), then incorporate the semantic units into the network architecture. The overall architecture has three main modules for video represen-

tation (Section 3.2). The training details are introduced in Section 3.3.

3.1. Semantic Units

Inspired by cognitive science [29, 64], we believe an effective video representation should describe the visual appearance and temporal variation of the body, object, and scenes. These factors can be represented by languages and have correspondences in visual appearance. We call them "semantic units".

Semantic units are in the form of discriminative descriptions. We select nouns or phrases from action names. Then we utilize WordNet [38], Wikipedia, and a few manual modifications to generate discriminative descriptions to represent the context of the selected texts. *e.g.*, the word polo is enhanced as polo, a game played on horseback with a long-handled mallet to be distinguished from the meaning polo shirt. Before generation, we remove the stop word in action names via NLTK [5]. On the other hand, we add body-related semantic entities that are sufficiently discriminative by themselves, including head, arms, hands, hips, legs, and feet. They occupy the majority of human activities and are shared by most actions. An illustration of producing semantic units are provided in *supplementary materials*.

All the semantic units are collected in the set $\mathcal{S} = \{s_k\}_{i=1}^K$. We manually categorized them into "body parts", "objects", and "scenes". For a minority of valid words that solely depend on body motions, such as sidekick and squat, we categorize them as motion units. Therefore,

$$\mathcal{S} = \{\mathcal{S}^{body}, \mathcal{S}^{object}, \mathcal{S}^{scene}, \mathcal{S}^{motion}\}. \quad (1)$$

We employ the text encoder of CLIP to encode every semantic unit and its elaborative description to feature vectors. The feature vectors of all the semantic units are concatenated to be $\mathbf{S} \in \mathbb{R}^{K \times d}$, where K is the number of semantic units, and d is the dimension of the text feature.
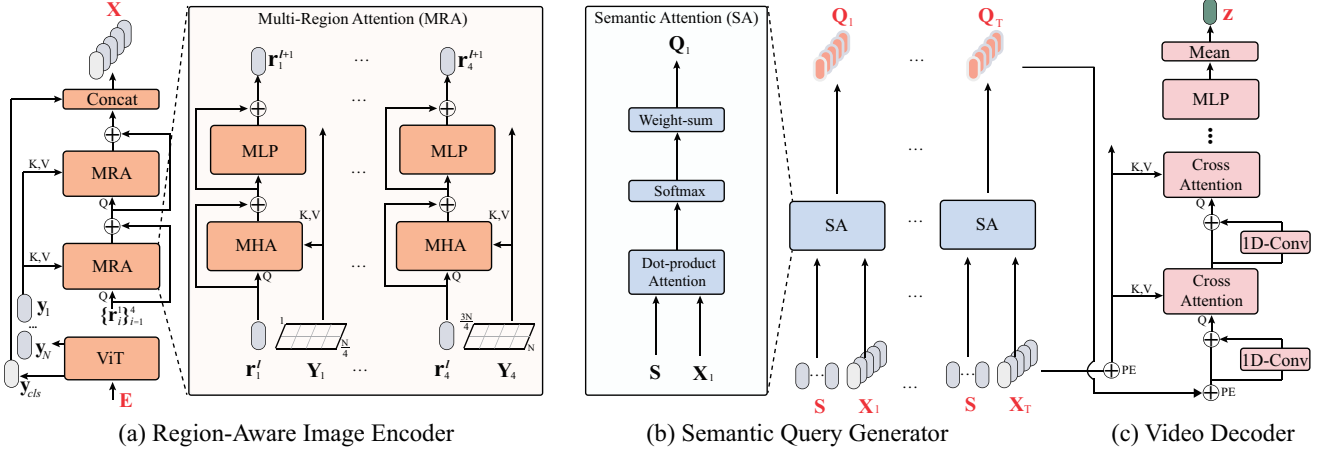
Figure 3. Detailed Network Components. Red variables denote the input or output of the region-aware image encoder, semantic query generator, and video decoder. (a) shows the region-aware image encoder containing a ViT and two multi-region attention (MRA) modules, each of which is shared for four regions. (b) shows the semantic query generator that contains a parameter-free semantic attention (SA) module, and (c) shows the video decoder, which is a stack of hybrid modules composed of cross-attention and 1D temporal convolution.

3.2. Network Architecture

Given a video segment with T frames, we extract the visual representations $\mathbf{X}$ for each frame by a region-aware image encoder, and utilize a semantic query generator to produce the queries $\mathbf{Q}$. The visual representations and queries of all the frames, $\widehat{\mathbf{X}}$ and $\widehat{\mathbf{Q}}$, are fed to the video decoder, which outputs the video representation $\mathbf{z}$.

Region-Aware Image Encoder The region-aware image encoder, as shown in Fig. 3a, takes the video frame as input and provides the frame-level and region-level representations. The encoder consists of two parts: a ViT initialized by CLIP is used to generate frame-level representations; Two consecutive multi-region attention (MRA) modules are used to generate the region-level representations.

For a frame $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, we divide the image into N non-overlapping patches following ViT. The patches are then embedded into vectors $\{\mathbf{e}_i\}_{i=1}^N$. A learnable embedding $\mathbf{e}_{cls}$ is prepended to the sequence of the embedded patches, called $[class]$ token. The embeddings are concatenated to form a matrix, which is further added with the position embeddings $\mathbf{E}_{pos}$ to obtain $\mathbf{E} \in \mathbb{R}^{(N+1) \times d}$:

$$\mathbf{E} = [\mathbf{e}_{cls}^\top, \mathbf{e}_1^\top, \mathbf{e}_2^\top, ..., \mathbf{e}_N^\top]^\top + \mathbf{E}_{pos}. \quad (2)$$

We feed $\mathbf{E}$ to ViT, producing the visual representations $\mathbf{Y}$:

$$\mathbf{Y} = [\mathbf{y}_{cls}^\top, \mathbf{y}_1^\top, \mathbf{y}_2^\top, ..., \mathbf{y}_N^\top]^\top, \quad (3)$$

where $\mathbf{y}_{cls} \in \mathbb{R}^d$ is the frame-level representation according to ViT, and $(\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N)$, obtained by averaging the patch embeddings of the last two layers of the ViT, are the representations for the corresponding patches. Inspired by strategy from the re-ID community [10, 13], the N patch representations are split into groups based on their affiliation to the four vertically arranged equal-size regions, forming $\{\mathbf{Y}_i\}_{i=1}^4$. The procedure is illustrated in Fig. 4. Utilized

as the keys and values, patch representations are further fed into an L-layer cross-attention (MRA) module, where each group $\mathbf{Y}_i$ of them serves the generation of the corresponding region-level representation:

$$\mathbf{r}_i^l = \mathrm{MRA}^l(\mathbf{r}_i^{l-1}, \mathbf{Y}_i), \quad l = 1, ..., L. \quad (4)$$

In particular, l denotes the index of MRA and $L = 2$, $\mathbf{r}_i^1$ are learnable embedding for region i with positional embeddings added, and $\mathbf{r}_i^L$ is region-level representation we finally obtain. MRA consists of Multi-Head Attention (MHA) and Multi-Layer Perceptron (MLP), thus can be represented by:

$$\begin{aligned} \mathrm{MRA}(a, b) &= c + \mathrm{MLP}(c), \\ c &= a + \mathrm{MHA}(a, b, b). \end{aligned} \quad (5)$$

It is noteworthy that $\{\mathbf{r}_i^l\}_{i=1}^4$ are parallelly computed by MRA^l via an attention mask. The final output of the region-aware image encoder is:

$$\mathbf{X} = [(\mathbf{y}_{cls})^\top, (\mathbf{r}_1^L)^\top, (\mathbf{r}_2^L)^\top, (\mathbf{r}_3^L)^\top, (\mathbf{r}_4^L)^\top]^\top. \quad (6)$$

$\mathbf{X} \in \mathbb{R}^{5 \times d}$ combines the frame-level representation and the region-level representations.

Semantic Query Generator We generate query feature $\mathbf{Q}$ for a frame with $\mathbf{S}$ and $\mathbf{X}$ as input. As illustrated in Fig. 3b, we feed the representations $\mathbf{X}$ to Semantic Attention (SA), a parameter-free block allows each visual representation in $\mathbf{X}$ to select the semantic units.

$$w_{i,j} = \frac{\langle \mathbf{x}_i, \mathbf{s}_j \rangle}{\| \mathbf{x}_i \| \cdot \| \mathbf{s}_j \|}, \quad a_{i,j} = \frac{\exp(w_{i,j}/\tau)}{\sum_{j=1}^K \exp(w_{i,j}/\tau)}, \quad (7)$$

where $\mathbf{x}_i \in \mathbb{R}^d$ and $\mathbf{s}_j \in \mathbb{R}^d$ are ith row and jth row of $\mathbf{X}$ and $\mathbf{S}$. $w_{i,j}$ is cosine similarity between $\mathbf{x}_i$ and $\mathbf{s}_j$. With the
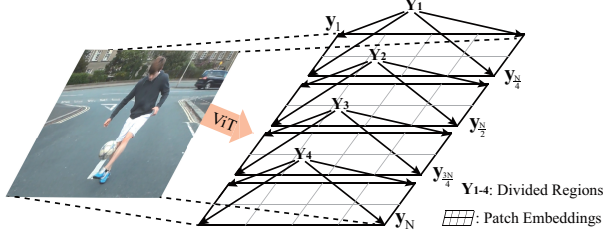
Figure 4. Visualization of dividing patch representations into regions. Note that when patches are not divisible by four along the vertical axis, we assign more lines of patches to middle regions.

attention weights $a_{i,j}$ calculated by the softmax activation of $w_{i,j}$ over the K-class semantic units $\mathbf{S}$ under temperature τ, we obtain the affinities matrix $\mathbf{A} \in \mathbb{R}^{5 \times K}$ between the representations $\mathbf{X}$ and $\mathbf{S}$. The query features $\mathbf{Q} \in \mathbb{R}^{5 \times d}$ for frame $\mathbf{I}$ is then calculated in a weight-sum manner:

$$\mathbf{Q} = \mathbf{A}\mathbf{S}. \quad (8)$$

Video Decoder Given a video $[\mathbf{I}_1, \mathbf{I}_2, ..., \mathbf{I}_T]$, we generate the feature maps and queries for each frame according to Eq. 6 and Eq. 8, obtaining $\widehat{\mathbf{X}} \in \mathbb{R}^{T \times 5 \times d}$ and $\widehat{\mathbf{Q}} \in \mathbb{R}^{T \times 5 \times d}$:

$$\widehat{\mathbf{X}} = [\mathbf{X}_1, ..., \mathbf{X}_i, ..., \mathbf{X}_T], \quad \widehat{\mathbf{Q}} = [\mathbf{Q}_1, ..., \mathbf{Q}_i, ..., \mathbf{Q}_T]. \quad (9)$$

The video decoder, as shown in Fig. 3c, consists of a sequence of a 1D-convolution layer and a cross-attention (CA) module, whose processing mechanism is similar to Eq. 5. The 1D-convolution layer processes the queries along the temporal dimension, while the CA module utilizes the representation of the queries to attend to the visual feature across the frames. It is noteworthy that the input queries to the video decoder are derived from the weighted sum of SUs' text embeddings, while the keys and values are from the region-aware image encoder. The video decoder promotes communication between the two modalities. More specifically, we perform the cross attention for each region and denote the visual representations and queries for the ith region as $\widehat{\mathbf{X}}_i \in \mathbb{R}^{T \times d}$ and $\widehat{\mathbf{Q}}_i \in \mathbb{R}^{T \times d}$. The procedure of the video decoder is as follows:

$$\begin{aligned} \widehat{\mathbf{H}}_i^m &= \widehat{\mathbf{Q}}_i^{m-1} + 1\text{Dconv}(\widehat{\mathbf{Q}}_i^{m-1}) \\ \widehat{\mathbf{Q}}_i^m &= \text{CA}^m(\widehat{\mathbf{H}}_i^m, \widehat{\mathbf{X}}_i), \quad m = 1, ..., M. \end{aligned} \quad (10)$$

where m indicates the layer index of the decoder and M = 4. The initialized query $\widehat{\mathbf{Q}}_i^0 = \widehat{\mathbf{Q}}_i$, and the output query after M-times attention is $\widehat{\mathbf{Q}}_i^M$. We concatenate the queries of different regions along the feature channel, obtaining $\widehat{\mathbf{Q}}^M \in \mathbb{R}^{T \times 5d}$. The video representation $\mathbf{z}$ is a vector calculated by applying MLP and MEAN over $\widehat{\mathbf{Q}}^M$ sequentially along the feature channel and the temporal dimension,

$$\mathbf{z} = \text{MEAN}(\text{MLP}(\widehat{\mathbf{Q}}^M)). \quad (11)$$

3.3. Training Details

Loss function We supervise the video representation with the text features of the action labels. Specifically, we feed action names with a prefixed template "a video of a person" into the frozen text encoder initialized by CLIP. The resulting text feature of the ith action is $\mathbf{c}_i$. The goal of the proposed network is to maximize the similarity between $\mathbf{z}_n$ and $\mathbf{c}_i$ if $\mathbf{z}$ belongs to the ith action. The loss function can be implemented by the cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{I} y^{i,n} \log \left(\frac{\exp(\mathbf{c}_i^\top \mathbf{z}_n)}{\sum_{j=1}^{I} \exp(\mathbf{c}_j^\top \mathbf{z}_n)} \right), \quad (12)$$

where $\mathbf{z}_n$ the representation of the nth video. The training set has N videos belonging to the I actions. If the nth video belongs to the ith action, $y^{i,n} = 1$, otherwise $y^{i,n} = 0$.

Network Training The ViT backbone in the region-aware image encoder is initialized by CLIP, while the other parameters in the region-aware image encoder and video decoder are randomly initialized. For network parameter training, we adopt an AdamW optimizer with an initial learning rate of 8×10^{-6} and 8×10^{-5} for the ViT backbone and the remaining parts, respectively. We train the network with 30 epochs and a weight decay of 0.001. The learning rate is warmed up for 5 epochs and decayed w.r.t. a cosine schedule. The input video follows the main sparse sampling method [54] and augmentation strategy [41] with a frame resolution 224×224. The hyperparameter temperature of softmax in the semantic query generator is set to 0.01. All the experiments are conducted with 8 32GB-V100-GPUs. More details are provided in *supplementary materials*.

4. Experiments

Datasets We benchmark our method on four popular video action recognition datasets. (1) Kinetics-400 [27] is a collection of around 240k training and 20k validation videos for 400 classes. Each clip covers around 10s. (2) Kinetics-600 [8] is an extension from Kinetics-400, which consists of around 410k training and 29k validation videos for 600 classes. (3) UCF-101 [51] contains 13,320 video clips with 101 classes. There are three splits of test data. (4) HMDB-51 [28] consists of 7,000 videos with 51 classes and has three test data splits. We conduct fully-supervised experiments on Kinetics-400 and Kinetics-600. With the pre-trained model on Kinetics-400, we conduct few-shot and zero-shot experiments on UCF-101 and HMDB-51.

4.1. Fully Supervised Comparison

Settings We conduct the fully-supervised experiments on Kinetics-400 and Kinetics-600. Each video is sampled with 8 or 16 frames. The multi-view inference is with 3 spatial crops and 4 temporal clips. Three variants of the proposed

Table 1. Comparison with state-of-the-art on Kinetics-400. FLOPs and throughput are reported per view with tools provided by X-CLIP [41]. For multi-modal approaches, parameters in the text branch are not counted. * indicates pretraining with a video-text collection.

Method	Pretrain	Frames	Top-1	Top-5	Views	#Param.(M)	FLOPs(G)	Throughput
Methods with ImageNet or web-scale image pretraining								
MViTv1-B, 64×3 [14]	-	64	81.2	95.1	3×3	36.6	455	7
Uniformer-B [30]	IN-1k	32	83.0	95.4	4×3	50.0	259	-
TimeSformer-L [4]	IN-21k	96	80.7	94.7	1×3	121.4	2380	3
Mformer-HR [44]	IN-21k	16	81.1	95.2	10×3	381.9	959	-
Swin-L [35]	IN-21k	32	83.1	95.9	4×3	200.0	604	6
Swin-L(384↑) [35]	IN-21k	32	84.9	96.7	10×5	200.0	2107	-
MViTv2-L(312↑) [31]	IN-21k	64	81.2	95.1	3×3	218.0	455	7
ViViT-H/16 [3]	JFT-300M	32	84.9	95.8	4×3	647.5	8316	-
MTV-H/16 [62]	JFT-300M	32	85.8	96.6	4×3	-	3705	-
Methods with web-scale language-image pretraining								
MTV-H/16 [62]	WTS-17B*	32	89.1	98.2	4×3	-	3705	-
PromptingCLIP-B/16 [26]		16	76.9	93.5	5×5	95.5	-	-
ActionCLIP-B/16 [55]		16	82.6	96.2	10×3	105.2	282	-
ActionCLIP-B/16 [55]		32	83.8	97.1	10×3	105.2	563	-
EVL-B/16 [34]		32	84.2	96.6	3×1	114.9	1777	-
EVL-L/14 [34]	CLIP-400M	32	87.3	97.6	3×1	357.9	2696	-
EVL-L/14(336↑) [34]		32	87.7	97.8	3×1	357.9	6065	-
X-CLIP-B/16 [41]		8	83.8	96.7	4×3	131.7	145	33
X-CLIP-B/16 [41]		16	84.7	96.8	4×3	131.7	287	17
X-CLIP-L/14 [41]		8	87.1	97.6	4×3	451.2	658	8
X-CLIP-L/14(336↑) [41]		16	87.7	97.4	4×3	451.2	3086	2
ASU-B/16 (ours)	CLIP-400M	8	84.1	96.3	4×3	140.2	146	30
ASU-B/16 (ours)	CLIP-400M	16	84.8	96.7	4×3	140.2	288	14
ASU-L/14 (ours)	CLIP-400M	8	87.8	97.8	4×3	425.3	660	8
ASU-L/14(336↑) (ours)	CLIP-400M	16	88.3	98.0	4×3	425.3	3084	2

Table 2. Comparison with state-of-the-art on Kinetics-600.

Method	Pretrain	Frames	Top-1	Views
MViT-B-24, 32×3 [14]	-	32	83.8	5×1
Swin-L(384↑) [35]	IN-21k	32	86.1	10×5
ViViT-H/16x2 320 [3]	JFT-300M	32	83.0	4×3
ViViT-H/16x2 [3]	JFT-300M	32	85.8	4×3
TokenLearner-L/10 [49]	JFT-300M	32	86.3	4×3
Florence(384↑) [63]	FLD-900M	32	87.8	4×3
CoVeR [65]	JFT-3B	96	87.9	1×3
MTV-H [62]	WTS-17B	32	89.6	4×3
X-CLIP-B/16 [41]		8	85.3	4×3
X-CLIP-B/16 [41]		16	85.8	4×3
X-CLIP-L/14 [41]	CLIP-400M	8	88.3	4×3
ASU-B/16 (ours)		8	85.7	4×3
ASU-B/16 (ours)		16	86.4	4×3
ASU-L/14 (ours)		8	88.6	4×3

network, *i.e.*, ASU-B/16, ASU-L/14, and ASU-L/14-336, adopt ViT-B/16, ViT-L/14, and ViT-L/14-336 for the region-aware image encoder, respectively.

Results In Tab. 1, we compare with the state-the-of-art methods on Kinetics-400. Under the standard 224×224 input resolution setting, our method achieves the second place

87.8%, which is 0.5% higher top-1 accuracy than the third-ranked method EVL [34]. It is noteworthy that MTV-H [62] performs the best by adopting much more pre-training data (70M video-text pairs with about 17B images [41]) and a larger pre-trained model(Vit-H). Our method outperforms MTV-H pre-trained on JFT-300M by 2.0% and achieves the best results among the CLIP-based methods. Under 336×336 input resolution setting, our method achieves the best Top-1 accuracy of 88.3%, which improves upon the previous highest result (X-CLIP [41] and EVL [34]) by 0.6%. Tab. 2 presents the results on Kinetics-600. Our ASU outperforms X-CLIP by 0.4% and 0.6% by using 8 and 16 frames per video, respectively. Moreover, our ASU achieves 88.6% with 4 × fewer input frames and 42.5 × fewer pre-training data compared to the current state-of-the-art method MTV-H. Our method achieves superior results among the methods that adopt similar-level pre-trained models. We attribute the effectiveness to the proposed visual-semantic attention mechanism, where the attentive semantic units can help exploit critical visual information for learning discriminative video representation.

Table 3. Few-shot comparisson on HMDB-51 and UCF-101.

Method	Frames	HMDB-51				UCF-101			
		K=2	K=4	K=8	K=16	K=2	K=4	K=8	K=16
TSM [33]	32	17.5	20.9	18.4	31.0	25.3	47.0	64.4	61.0
TimeSformer [4]	32	19.6	40.6	49.4	55.4	48.5	75.6	83.7	89.4
Swin-B [35]	32	20.9	41.3	47.9	56.1	53.3	74.1	85.8	88.7
ActionCLIP [55]	8	55.0	56.0	58.0	-	80.0	85.0	89.0	-
X-CLIP-B/16 [41]	32	53.0	57.3	62.8	64.0	76.4	83.4	88.3	91.4
X-Florence [41]	32	51.6	57.8	64.1	64.2	84.0	88.5	92.5	94.8
ASU-B/16 (ours)	8	**57.7**	**60.9**	**65.8**	**70.1**	**88.8**	**92.3**	**94.0**	**95.3**
ASU-B/16 (ours)	32	**60.1**	**63.8**	**67.2**	**70.8**	**91.4**	**94.6**	**96.0**	**97.2**

Table 4. Zero-shot comparison on HMDB-51 and UCF-101.

Method	HMDB-51	UCF-101
MTE [61]	19.7±1.6	15.8±1.3
ASR [56]	21.8±0.9	24.4±1.0
ZSECOC [45]	22.6±1.2	15.1±1.7
UR [68]	24.4±1.6	17.5±1.6
TS-GCN [19]	23.2±3.0	34.2±3.1
E2E [6]	32.7	48
ER-ZSAR [11]	35.3±4.6	51.8±2.9
ActionCLIP [55]	40.8±5.4	58.3±3.4
X-CLIP-B/16 [41]	44.6±5.2	72.0±2.3
ASU-B/16 (ours)	**48.1±2.8**	**75.0±3.7**

Table 5. Zero-shot comparison on Kinetics-600.

Method	Top-1	Top-5
DEVISE [18]	23.8±0.3	51.0±0.6
ALE [1]	23.4±0.8	50.3±1.4
SJE [2]	22.3±0.6	48.2±0.4
ESZSL [48]	22.9±1.2	48.3±0.8
DEM [66]	23.6±0.7	49.5±0.4
GCN [20]	22.3±0.6	49.7±0.6
ER-ZSAR [11]	42.1±1.4	73.1±0.3
X-CLIP-B/16 [41]	65.2±0.4	86.1±0.8
ASU-B/16 (ours)	**67.6±0.2**	**87.2±0.3**

4.2. Few-shot Comparisons

Settings The few-shot experiments are evaluated on HMDB-51 and UCF-101. The training set is constructed by randomly sampling 2,4,8, and 16 videos from each class and setting the frame number in each video to 8 or 32. Following [41], we adopt the first split of the test set as an evaluation set and report the result of a single-view inference.

Results In Tab. 3, we present the results of K-shot learning. Compared with methods employing single-modality pre-trained models, our method outperforms them by a large margin. *e.g.*, when K=2, it improves Swin-B by *39.2%* on HMDB-51 and *38.1%* on UCF-101, demonstrating the effectiveness of the semantic knowledge in the few-shot settings. Moreover, our method still achieves state-of-the-art performance compared with the methods that employed the cross-modal pre-trained model. On HMDB-51, our method with eight frames per video achieves *57.7%* top-1 accuracy under K=2-shot learning, which is higher than the previous best ActionCLIP with *2.7%*. On UCF-101, our method achieves *91.4%* top-1 accuracy under K=2-shot learning, which outperforms X-Florence by *7.4%*. Our method has achieved consistent superiority from K=2 to K=16, indicating the effectiveness of semantic units.

4.3. Zero-shot Comparisons

Settings We evaluate the zero-shot performance of ASU-B/16, which is pre-trained on Kinetics-400 with 8-frame videos. The protocol is the same as in [41]: For HMDB-51 and UCF-101, we average the results on the given three splits and report the top-1 accuracy and standard deviation. For Kinetics-600, we randomly selected 160 categories from 220 new categories, exclusive from Kinetics-600, three times. Then we report the averaged top-1 and top-5 accuracies and their standard deviations.

Results We report the zero-shot results in Tab. 4 and Tab. 5. With the model trained on Kinetics-400, our method outperforms X-CLIP-B/16 by *3.5%* and *3.0%* top-1 accuracy on HMDB-51 and UCF-101, respectively. When tested on a larger scale dataset Kinetics-600, our method outperforms X-CLIP-B/16 by *2.4%* top-1 accuracy. We attribute the success of our method on the few-shot/zero-shot setting to the utilization of semantic units, whose re-usability alleviates the difficulty of adapting our model to a new scenario.

4.4. Ablation Study

We use the ASU-B/16_{8f} with single-view inference for all the ablation experiments. By default, the fully-supervised experiments are evaluated on Kinetics-400, the few-shot experiments are conducted on the first split of HMDB-51, and the zero-shot evaluation is on the first split

Table 6. Ablation study on the effect of proposed components. *Semantics*: semantic units. *Region*: region-level representation. *Temporal*: 1D-convolution layer in video decoder.

Semantic.	Region.	Temporal.	Top-1 Acc.(%)	GFLOPs
-	-	-	81.1	141
✓	-	-	82.0	138
-	✓	-	81.8	145
✓	✓	-	82.4	146
✓	✓	✓	**82.6**	146

Table 7. Ablation study on different types of semantic units.

Semantic Units	Fully.	2-shot	0-shot
$\varnothing$	82.0	52.6	64.9
$\mathcal{S}^{body}$	82.2	56.5	72.6
$\mathcal{S}^{object}$	82.4	57.1	73.8
$\mathcal{S}^{scene}$	82.2	56.6	73.6
$\mathcal{S}^{motion}$	82.1	56.3	72.2
$\mathcal{S}$	**82.6**	**57.7**	**75.0**

Table 8. Ablation study on the effect of different region configurations. #Region is the number of regions. The configuration in Split denotes the horizontal number $\times$ the vertical number.

# Region	0	2	4		8	
Split	-	2×1	4×1	2×2	8×1	4×2
Top1-Acc(%).	82.0	82.3	**82.6**	82.2	82.4	81.8
GFLOPs	138	140	146	146	162	162

of the validation set of UCF-101.

Effects of Proposed Components We first investigate the effectiveness of the proposed components. The experiments are conducted under fully-supervised learning of Kinetics-400. (1) We treat the ActionCLIP-B/16-Transf, which introduces a 6-layer self-attention transformer followed by a mean pooling layer stacking on top of the CLIP image encoder as the baseline. Results are presented in Tab. 6, and the first row denotes the baseline. (2) Compared with the baseline, replacing the transformer of baseline with our video decoder and introducing semantic units to generate queries for cross attention bring a *0.9%* gain in term of the top-1 accuracy. The results demonstrate the effectiveness of enhancing visual representations with semantic embedding. (3) To validate the effectiveness of region-level representations, we disable the semantic-guided enhancement branch by feeding randomly initialized query embedding into the video decoder and retraining the model. The region-level representation can improve the top-1 accuracy of baseline by *0.7%*. (4) Combining the region-level representation and semantic unit further improve the value by *0.6%*. The results indicate using proper region-level representation as enhancement can help alignments about instances and texts. (5) Besides, equipping temporal relations in video decoder can further boost the top-1 accuracy by *0.2%*. Overall, our proposed components can raise the top-1 accuracy of the baseline from *81.1%* to *82.6%*.

Investigation of Semantic Units To validate the effect of semantic units, we first construct a baseline model, which replaces the queries to the video decoder with visual features from the region-aware image encoder. The results are reported in the first row in Tab. 7. Then we investigate the effectiveness of each sub-collection in semantic units as defined in Eq. 1. From Tab. 7, we can observe the models equipped with semantic units improves the baseline, especially for the 2-shot (*+5.1%*) and 0-shot (*+10.1%*) experiments. The results indicate that visual representations without semantic guidance are not robust when facing a severe lack of data. We also conduct experiments to evaluate the four types of semantic units. In particular, the set of "objects" is the most helpful to the results, indicating the objects are discriminative factors in the action recognition. The results in rows 2,3,4,5 show that each part of seman-

tic units can bring performance gain, and their combination (in row 6) can achieve the best results. The phenomenon indicates that different types of semantic units complement each other, and languages benefit action recognition performance, especially in the transferability of the model.

Effects of Different Region Configurations We conduct experiments with different spatial configurations to exploit spatial information for better action recognition. In particular, we split the whole image into regions according to two aspects: the number of split regions and the approach of spatial slicing. Results are presented in Tab. 8. There are two observations. (1) Increasing the number of partitions will not continuously improve performance. When the size of each region is too large, the visual representations cannot align with the semantic units of small targets. On the contrary, when the size of each region is too small, the receptive field of each region cannot cover the visual appearance of most semantic units. (2) Vertical divisions cannot bring gain to the performance. We propose an assumption for the phenomenon. In an action video, the variations of the body parts usually happen in horizontal regions so that the horizontal regions can keep the majority of part structures across the frames, which can help better describe an action. In comparison, the vertical divisions will break such structures, leading to a performance decrease.

5. Discussion

Conclusion In this paper, we propose a new video representation learning framework for action recognition based on pre-trained visual-language models. To leverage fine-grained visual-language correlation, we introduce "semantic units" that provide factorized and reusable textual knowledge hiding behind actions. A multi-region attention module that perceives region-aware information is proposed to better capture fine-grained alignments. Under the

guidance of selected semantic units, a cross-modal decoder is designed for decoding spatiotemporal video representations. Our approach produces discriminative video representations, outperforming the state-of-the-art approaches in fully supervised comparisons. The semantic units' reusability further alleviates the learning difficulty of adapting our model to new scenarios, achieving superior results in the zero-shot/few-shot settings.

Limitations and Future Work There are two potential improvements for ASU to explore in the future. Firstly, the generation of semantic units still requires a few manual operations. With the increasing maturity of large-scale language models such as the GPT series [7, 42], the current procedure can be fully automated with appropriate prompts and CoT [58] designs. Secondly, ASU and other recent VLM-based approaches consume significant computational resources on pre-trained backbones during the inference phase. From the perspective of practical application, incorporating more computational-efficient techniques is a valuable research field.

References

[1] Zeynep Akata, Florent Perronnin, Zaïd Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1425–1438, 2016. 7

[2] Zeynep Akata, Scott E. Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, pages 2927–2936, 2015. 7

[3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, pages 6816–6826, 2021. 2, 6

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, pages 813–824, 2021. 1, 2, 6, 7

[5] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly, 2009. 3

[6] Biagio Brattoli, Joseph Tighe, Fedor Zhdanov, Pietro Perona, and Krzysztof Chalupka. Rethinking zero-shot video classification: End-to-end training for realistic applications. In *CVPR*, pages 4612–4622, 2020. 7

[7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020. 9

[8] João Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *CoRR*, abs/1808.01340, 2018. 5

[9] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, pages 4724–4733, 2017. 1

[10] Dapeng Chen, Zejian Yuan, Badong Chen, and Nanning Zheng. Similarity learning with spatial constraints for person re-identification. In *CVPR*, June 2016. 4

[11] Shizhe Chen and Dong Huang. Elaborative rehearsal for zero-shot action recognition. In *ICCV*, pages 13618–13627, 2021. 7

[12] Leizhen Dong, Zhimin Li, Kunlun Xu, Zhijun Zhang, Luxin Yan, Sheng Zhong, and Xu Zou. Category-aware transformer network for better human-object interaction detection. In *CVPR*, pages 19516–19525, 2022. 3

[13] Chanho Eom and Bumsub Ham. Learning disentangled representation for robust person re-identification. In *NeurIPS*, pages 5298–5309, 2019. 4

[14] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, pages 6804–6815, 2021. 1, 6

[15] Haoshu Fang, Jinkun Cao, Yu-Wing Tai, and Cewu Lu. Pairwise body-part attention for recognizing human-object interactions. In *ECCV (10)*, pages 52–68, 2018. 3

[16] Christoph Feichtenhofer. X3D: expanding architectures for efficient video recognition. In *CVPR*, pages 200–210, 2020. 1, 2

[17] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6201–6210, 2019. 1, 2

[18] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomás Mikolov. Devise: A deep visual-semantic embedding model. In *NeurIPS*, pages 2121–2129, 2013. 7

[19] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. In *AAAI*, pages 8303–8311, 2019. 7

[20] Pallabi Ghosh, Nirat Saini, Larry S. Davis, and Abhinav Shrivastava. All about knowledge graphs for actions. *CoRR*, abs/2008.12432, 2020. 7

[21] Georgia Gkioxari, Ross B. Girshick, and Jitendra Malik. Actions and attributes from wholes and parts. In *ICCV*, pages 2470–2478, 2015. 3

[22] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022. 2

[23] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *ICCV Workshops*, pages 3154–3160, 2017. 1

[24] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatiotemporal scene graphs. In *CVPR*, pages 10233–10244, 2020. 3

[25] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation

learning with noisy text supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 2021. 1, 2

[26] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *ECCV (35)*, pages 105–124, 2022. 1, 2, 6

[27] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. 2, 5

[28] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso A. Poggio, and Thomas Serre. HMDB: A large video database for human motion recognition. In *ICCV*, pages 2556–2563, 2011. 5

[29] Christopher A. Kurby and Jeffrey M. Zacks. Segmentation in the perception and memory of events. *Trends in cognitive sciences*, pages 72–79, 2008. 3

[30] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *ICLR*, 2022. 6

[31] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *CVPR*, pages 4794–4804, 2022. 6

[32] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Shiyi Wang, Haoshu Fang, Ze Ma, Mingyang Chen, and Cewu Lu. Pastanet: Toward human activity knowledge engine. In *CVPR*, pages 379–388, 2020. 3

[33] Ji Lin, Chuang Gan, and Song Han. TSM: temporal shift module for efficient video understanding. In *ICCV*, pages 7082–7092, 2019. 7

[34] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen CLIP models are efficient video learners. In *ECCV (35)*, pages 388–404, 2022. 2, 6

[35] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, pages 3192–3201, 2022. 1, 2, 6, 7

[36] Yue Ma, Yali Wang, Yue Wu, Ziyu Lyu, Siran Chen, Xiu Li, and Yu Qiao. Visual knowledge graph for human action reasoning in videos. In *ACM Multimedia*, pages 4132–4141, 2022. 3

[37] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *CVPR*, pages 1046–1056, 2020. 3

[38] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, pages 39–41, 1995. 3

[39] Ron Mokady, Amir Hertz, and Amit H. Bermano. Clipcap: CLIP prefix for image captioning. *CoRR*, abs/2111.09734, 2021. 2

[40] Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. Clip-it! language-guided video summarization. In *NeurIPS*, pages 13988–14000, 2021. 2

[41] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *ECCV (4)*, pages 1–18, 2022. 1, 2, 5, 6, 7

[42] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022. 9

[43] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. St-adapter: Parameter-efficient image-to-video transfer learning for action recognition. In *NeurIPS*, 2022. 2

[44] Mandela Patrick, Dylan Campbell, Yuki M. Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and João F. Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. In *NeurIPS*, pages 12493–12506, 2021. 1, 6

[45] Jie Qin, Li Liu, Ling Shao, Fumin Shen, Bingbing Ni, Jiaxin Chen, and Yunhong Wang. Zero-shot action recognition with error-correcting output codes. In *CVPR*, pages 1042–1051, 2017. 7

[46] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *ICCV*, pages 5534–5542, 2017. 1, 2

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 1, 2

[48] Bernardino Romera-Paredes and Philip H. S. Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, pages 2152–2161, 2015. 7

[49] Michael S. Ryoo, A. J. Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. In *NeurIPS*, pages 12786–12797, 2021. 6

[50] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, pages 568–576, 2014. 1, 2

[51] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 5

[52] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015. 1, 2

[53] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, pages 6450–6459, 2018. 1, 2

[54] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment

networks: Towards good practices for deep action recognition. In *ECCV (8)*, pages 20–36, 2016. 1, 2, 5

[55] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *CoRR*, abs/2109.08472, 2021. 1, 2, 6, 7

[56] Qian Wang and Ke Chen. Alternative semantic representations for zero-shot human action recognition. In *ECML/PKDD (1)*, pages 87–102, 2017. 7

[57] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. CRIS: clip-driven referring image segmentation. In *CVPR*, pages 11676–11685, 2022. 2

[58] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022. 9

[59] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV (15)*, pages 318–335, 2018. 1, 2

[60] 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. In *EMNLP (1)*, pages 6787–6800, 2021. 2

[61] Xun Xu, Timothy M. Hospedales, and Shaogang Gong. Multi-task zero-shot action recognition with prioritised data augmentation. In *ECCV (2)*, pages 343–359, 2016. 7

[62] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *CVPR*, pages 3323–3333, 2022. 1, 2, 6

[63] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision. *CoRR*, abs/2111.11432, 2021. 6

[64] Jeffrey M. Zacks, Barbara Tversky, and Gowri Iyer. Perceiving, remembering, and communicating structure in events. *Journal of experimental psychology: General*, page 29, 2001. 3

[65] Bowen Zhang, Jiahui Yu, Christopher Fifty, Wei Han, Andrew M. Dai, Ruoming Pang, and Fei Sha. Co-training transformer with videos and images improves action recognition. *CoRR*, abs/2112.07175, 2021. 6

[66] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *CVPR*, pages 3010–3019, 2017. 7

[67] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *ECCV (1)*, pages 831–846, 2018. 1, 2

[68] Yi Zhu, Yang Long, Yu Guan, Shawn D. Newsam, and Ling Shao. Towards universal representation for unseen action recognition. In *CVPR*, pages 9436–9445, 2018. 7