Align before Adapt: Leveraging Entity-to-Region Alignments for Generalizable Video Action Recognition

Yifei Chen **Dapeng Chen** **Ruijin Liu** **Sai Zhou** **Wenyuan Xue** **Wei Peng**
IT Innovation and Research Center, Huawei Technologies
{chenyifei14, chendapeng8, liuruijin1}@huawei.com

Abstract

Large-scale visual-language pre-trained models have achieved significant success in various video tasks. However, most existing methods follow an "adapt then align" paradigm, which adapts pre-trained image encoders to model video-level representations and utilizes one-hot or text embedding of the action labels for supervision. This paradigm overlooks the challenge of mapping from static images to complicated activity concepts. In this paper, we propose a novel "Align before Adapt" (ALT) paradigm. Prior to adapting to video representation learning, we exploit the entity-to-region alignments for each frame. The alignments are fulfilled by matching the region-aware image embeddings to an offline-constructed text corpus. With the aligned entities, we feed their text embeddings to a transformer-based video adapter as the queries, which can help extract the semantics of the most important entities from a video to a vector. This paradigm reuses the visual-language alignment of VLP during adaptation and tries to explain an action by the underlying entities. This helps understand actions by bridging the gap with complex activity semantics, particularly when facing unfamiliar or unseen categories. ALT demonstrates competitive performance while maintaining remarkably low computational costs. In fully supervised experiments, it achieves 88.1% top-1 accuracy on Kinetics-400 with only 4947 GFLOPs. Moreover, ALT outperforms the previous state-of-the-art methods in both zero-shot and few-shot experiments, emphasizing its superior generalizability across various learning scenarios.

1. Introduction

Video action recognition is a fundamental task in the pursuit of intelligent video understanding. The recent trend of utilizing the visual-language pre-trained (VLP) models [26, 31, 49, 73] have significantly advanced the research of action recognition [27, 37, 43, 45, 64, 71]. By lightly fine-tuning the model, VLP-based methods outperform the previous end-to-end network architectures, including two-stream networks [54, 63, 78], 3D convolutional neural networks [6, 14, 15, 22, 48, 58, 59, 67], and vision-transformer-based (ViT) networks [3, 12, 38, 46, 70]. Employing a pre-trained VLP model for action recognition can better encode the semantic meaning of items in images, even if they have very different visual appearances. This is very helpful in understanding human action and also explains why VLP models have achieved superior performance. As shown in Fig. 1, the current VLP-based action recognition methods follow an "adapt then align" paradigm. They either introduce temporal interaction upon image representations or insert temporal modules into pre-trained image encoders. However, the "adapt then align" paradigm will merely fit the video representation to the action name, which potentially destroys the other visual-semantic correspondences provided by VLP models. As far as we are concerned, actions are complex concepts that involve multiple fine-grained entities, such as body parts, scenes, and objects. With VLP model, the text embedding of the relevant entities should also be grounded in some image region. "adapt then align" paradigm does not emphasize the underlying entities-to-regions correspondence behind the action concept. Furthermore, human-centric activities often share common entities, implying that visual-language correspondences can be reused across different actions, even for those that were not included in the training set. The re-usability of entities-to-regions correspondences allow the model to quickly recognize new action categories.

In this paper, we propose the "ALign before adapt" (ALT) paradigm. Unlike the "adapt then align" approaches that align the image-level visual embedding with the text embedding of the action name, ALT aims to establish an entity-level correspondence to support action recognition. The relevant entities should have evidence in specific regions of the image. To achieve entities-to-regions alignment, the VLP model is leveraged in two aspects: (1) Aggregating adjacent and similar image token embeddings from the VLP model. The resulting embedding typically represents a region containing the same entity. (2) Selecting the most relevant entities for each region by matching their image embeddings to the text

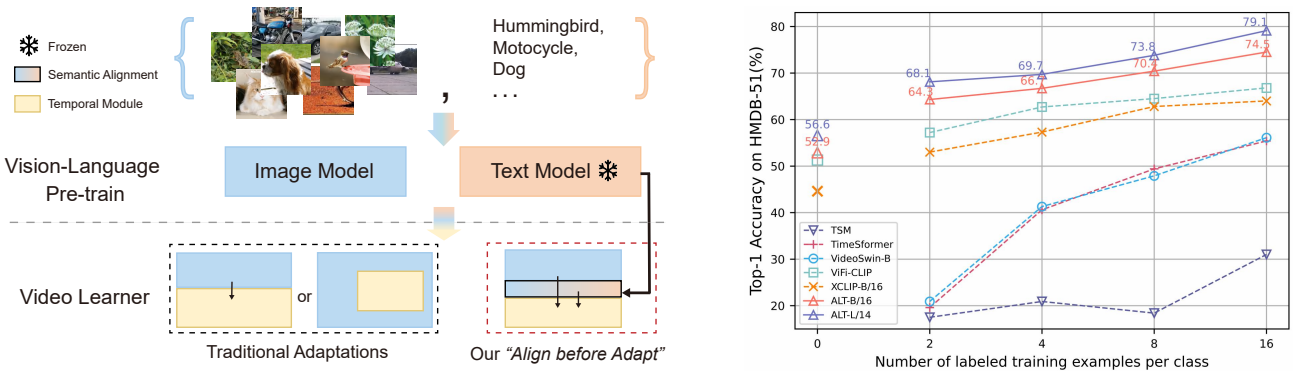

Figure 1. **Left:** Paradigm comparison between traditional adaptation approaches and our "Align before Adapt" method. **Right:** Zero-shot and few-shot performance comparison on HMDB-51 dataset. Pretrained on Kinetics-400, our method surpasses the previous state of the arts.

embeddings of a text corpus.

Using the established alignments, we utilize the text embedding of the entities as queries in a transformer-based decoder for action recognition. This adaptation step helps bridge the gap between general image representations and video action representation while preserving the visual-language correspondences from VLP models. ALT exhibits impressive generalization ability with low computational complexity. It enhances our framework by 6.8% in top-1 accuracy on the HMDB-51 dataset under the 2-shot configuration, while reducing computational cost by 23% with the ViT-base backbone. In summary, our contributions are as follows:

- We propose an "align before adapt" paradigm that leverages entities-to-region correspondences to guide the adaptation from VLP to action recognition. The paradigm preserves the visual-language alignment of VLP during the video representation adaption, achieving better interpretability and generalization ability.
- We propose a new transformer-based video adapter to extract the semantics of the most important entities from a video to a vector. The adapter adopts a transformer-based architecture, which utilizes the text embedding of the selected entities as the query and the multi-frame visual embeddings as the key and values.
- Extensive experiments under various learning configurations are conducted. Besides demonstrating competitive performance with low computational complexity (surpasses the current leading approach with the same VLP backbone by *0.4%* top-1 accuracy while requiring *55%* fewer GFLOPs). Our method reveals superior generalizability due to the reusable text entities (surpasses the previous state-of-the-art by more than *10%*).

2. Related Work

Large Scale Visual-Language pretraining. In the past few years, the surge of large-scale visual-language pre-trained (VLP) models [26, 31, 34, 49, 56, 76] have revolutionized multiple fields of computer vision, including image classification, captioning [73], grounding [33], image-text retrieval, and so on. With the availability of massive amounts of web-scale visual-text paired data, these models learn cross-modal representations through masked language modeling and/or contrastive learning. Specifically, one of the most representative works, CLIP [49] , is trained on 400M data following a contrastive manner, and shows remarkable performance on zero-shot image classification. The success of VLP models inspires the "fine-tuning" trend on multiple downstream tasks, such as open-vocabulary detection [21], segmentation [18, 66, 69], caption [41], summarization [42], generation [50], etc. Our method adopts CLIP as the backbone for video action recognition tasks under fully-supervised, few-shot, and zero-shot scenarios.

Video Action Recognition. The prosperity of deep learning has sparked various works for effective video action recognition. Initially, there were two directions of methods: two-stream 2D CNNs [54, 63, 78] that process and spatial and temporal context parallelly, and 3D CNNs [14, 15, 48, 58, 59, 67] that factorize the convolution across spatial and temporal dimensions simultaneously. Later transformer-based approaches, including ViViT [1], Timesformer [3], and VideoSwin [38] outperformed the convolutional methods, by better capturing long-term dependencies through scalable self-attention mechanisms. More Recently, leveraging available VLP models such as CLIP [49] and Florence [74], has become a data-friendly trend. EVL [37], ST-Adapter [45], and AIM [71] add lightweight modules to the fixed CLIP backbone for close-set recognition tasks, while ActionCLIP [64] , X-CLIP [43], and ViFi-CLIP [52] propose frameworks that enable adaptation to new scenarios. While the above methods focus on adapting the visual branch of models to video input directly, our approach explores 'entity-to-region' visual-semantic alignments before the adaptation. This bridges the gap of mapping with complicated activity semantics during video representation learning.

Region-Aware Perception for Vision Transformer. In re-

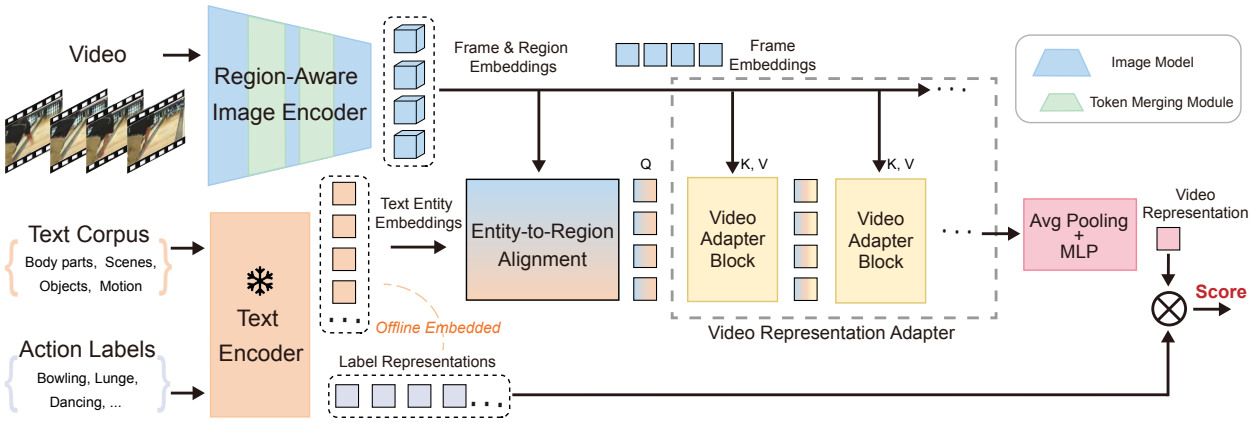

Figure 2. An overview of our framework: we utilize a video clip and an offline text corpus as inputs to learn a video representation, which is supervised with the objective of maximizing the similarity score with the text representation of the corresponding action label.

cent research on ViT architectures, it has been well-studied that capturing fine-grained patterns in visual signals leads to improved representation learning. Various approaches, such as Swin Transformer [38], Region ViT [8], and GCViT [23], propose incorporating multi-scale attention into the ViT to achieve better performance in various downstream tasks including recognition, detection, and segmentation. With the rise of visual-language pretraining, GLIP [33] suggests learning better instance-level language-aware representations through grounded image-text data, while FILIP [72] and Dense CLIP [51] focus on introducing patch-level contrastive losses. These works have shown impressive progress in various scenarios. In contrast to methods that require additional structures, data, or supervision, we adopt ToMe [5], which initially aims at boosting ViTs with minor performance drops, in our image encoder. Finding its soft bipartite matching strategy tends to cluster patch tokens of frames with similar region/instance representativeness, we utilize the merged tokens to achieve "entity-to-region" visual-semantic alignments for video action recognition.

3. Methodology

Our method proposes an "align before adapt" framework to learn discriminative and transferable video representation. The overview of our framework is illustrated in Fig. 2. The framework has two main steps. First, we explore the entities-to-regions alignment from a constructed text-corpus and the region-aware image embeddings (Sec. 3.1). Then we leverage the text embeddings of aligned entities to guide the adaption to the video representation (Sec. 3.2).

3.1. Entitiy-to-Region Alignments

Text corpus of action-related entities. Drawing inspiration from cognitive science and recent research [13, 30, 35, 75], we believe that perceiving and leveraging intermediate spatiotemporal-variated patterns, such as bodies, objects, and scenes, can greatly mitigate the difficulty of understanding activity concepts. Thanks to the VLP models, these patterns can be linguistically expressed and perceived based on their similarities with visual representations in the embedding space. We construct a knowledge base for these patterns, referred to as "text corpus".

To generate a reusable text corpus, we first collect a set of action labels from several recognition datasets including [7, 29, 55]. The action-related text corpus is then generated with our proposed automatic pipeline: (1) Extracting and expanding entities from original action names by utilizing POS (Nouns) tagging tools [4, 24] and prompting LLM [44]. (2) Generating a set of descriptions for each entity by utilizing WordNet [40] and LLM. (3) Employing word sense disambiguation techniques, including the Lesk algorithm [2] and T5 language model [62], to filter out unrelated descriptions of an entity to action. In addition, body parts, such as `head` and `feet` are added to the corpus by default. They are involved in most human activities. The details and utilized prompts are provided in supplementary materials.

The text corpus, denoted as $\mathcal{S} = \{s_i\}_{i=1}^{K}$, consists of K entities, where each entity s_i is represented by its entity name and description: $s_i = \{\texttt{unit}_i, \texttt{description}_i\}$. The text embeddings for all the text entities can be denoted as $\mathbf{S} \in \mathbb{R}^{K \times d}$, where d represents the dimension of embeddings.

Region-aware image embeddings. To fully explore the entity-to-region alignments, we need to perceive the region-level image embeddings. With a CLIP-ViT, the frame input is initially divided into N non-overlapping patches and projected to a sequence of d-dimension embeddings $\mathbf{E}_0$:

$$\mathbf{E}_0 = [\mathbf{e}_{cls,0}, \mathbf{e}_{1,0}, \mathbf{e}_{2,0}, ..., \mathbf{e}_{N,0}] + \mathbf{E}_{pos}, \quad (1)$$

where each embedding $\mathbf{e}_{n,0}$ corresponds to a patch. $\mathbf{e}_{cls,0}$ and $\mathbf{E}_{pos}$ denote prepended $[class]$ embedding and position embeddings, respectively. However, the image patches are too redundant to represent a region. We adopt token-

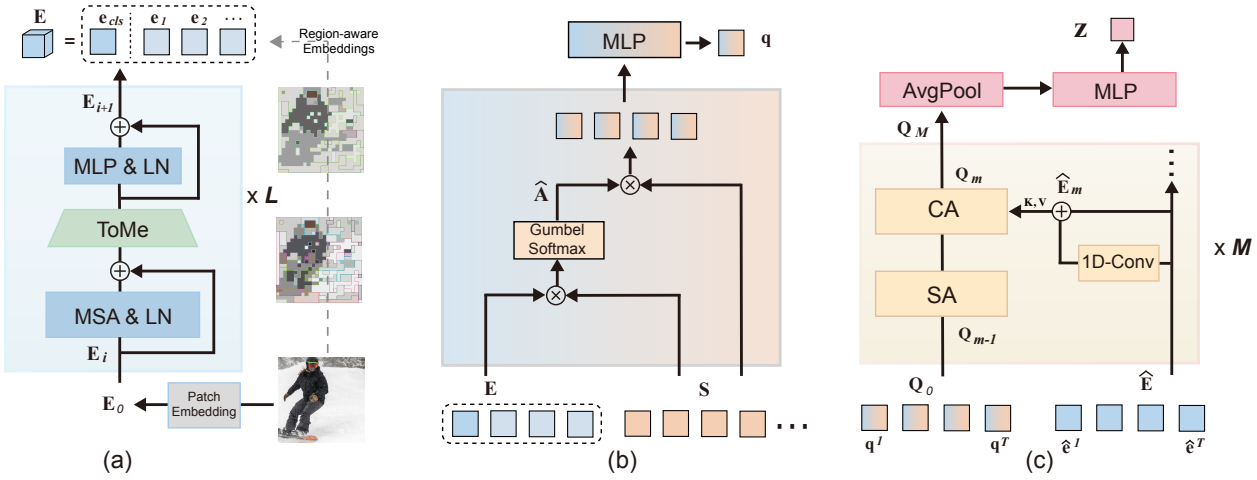

Figure 3. Detailed network components: (a) The region-aware image encoder includes a ViT with plug-in token merging modules. *MSA*, *MLP*, and *LN* indicate multi-head self-attention, multilayer perception, and layernorm, respectively. (b) The entity-to-region alignment module obtains the aligned query in a softmax-weight-sum manner. and (c) shows the multi-modal video adapter, with each block containing a stack of hybrid modules composed of attention layers and 1D temporal convolution.

merging [5] to aggregate patches. The token merging technique originally aimed to accelerate the ViT architecture. It employs a soft bipartite matching algorithm to find the most similar r pairs of embeddings (excluding $[cls]$). Each pair of embeddings are merged into one, reducing the total number by r. The token merging module is inserted into each transformer block, denoted by $\texttt{BLOCK}_{ToMe}$. The forward pass of the modified transformer block can be formulated as:

$$\mathbf{E}_l = \texttt{BLOCK}_{ToMe}(\mathbf{E}_{l-1}), \tag{2}$$

where l denotes the index of transformer blocks, and $\mathbf{E}_l \in \mathbb{R}^{(N+1-l\times r)\times d}$ reduces r patch embeddings compared with $\mathbf{E}_{l-1}$. After passing through L blocks (the second last layer of ViT), the final image embeddings are represented by:

$$\mathbf{E} = [\mathbf{e}_{cls}, \mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_{N-L\times r}], \tag{3}$$

where $\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_{N-L\times r}$ are region-aware embeddings that account for corresponding merged patches which have similar visual representativeness. We visualize the merging procedure along the transformer layers in Fig. 3a, where the patches with the same color and border are merged into one and form region-aware embeddings.

Entity-to-regions alignments. ALT aims to build the entities-to-regions correspondences by matching the text embeddings of entities and the region embeddings of each frame. As depicted in Fig. 3b, the process starts by calculating the similarity matrix $\mathbf{A}$ between the visual embeddings $\mathbf{E}$ and text entity embeddings $\mathbf{S}$ through Gumbel-Softmax [25, 39] over $\mathbf{S}$ with $Gumbel(0, 1)$ distributed samples γ, where:

$$\mathbf{A}_{i,j} = \frac{\exp(\langle \mathbf{e}_i, \mathbf{s}_j \rangle / (\|\mathbf{e}_i\| \cdot \|\mathbf{s}_j\|) + \gamma_j)}{\sum_{k=1}^{K} \exp(\langle \mathbf{e}_i, \mathbf{s}_k \rangle / (\|\mathbf{e}_i\| \cdot \|\mathbf{s}_k\|) + \gamma_k)}. \tag{4}$$

Note that $\mathbf{e}_i \in \mathbb{R}^d$ and $\mathbf{s}_j \in \mathbb{R}^d$ are ith and jth embeddings of $\mathbf{E}$ and $\mathbf{S}$, respectively.

To address the ambiguities of aligned semantics, we highlight the most similar entity for each region by enforcing a differentiable one-hot trick [60, 68]:

$$\widehat{\mathbf{A}} = one\text{-}hot(\mathbf{A}_{argmax}) + \mathbf{A} - detach(\mathbf{A}), \tag{5}$$

where $detach$ stops the gradient. $\widehat{\mathbf{A}} \in \mathbb{R}^{(N+1-L\times r)\times K}$ only aligns the most correlated text entity for each frame-level or region-aware embedding with dominating weights while keeping differentiable. It is noteworthy that, to validate the precision and interpretability of visual-semantic alignments, in Fig. 4 left, we visualize the correspondence between region-aware embeddings and text entities.

3.2. Video Representation Adapter

The video adapter aims to extract the most discriminative information from multi-frame visual embeddings to a single vector for action recognition. Previous methods utilize the labels or the text embeddings of the action names to supervise the adapter, while our method can leverage the relevant entities obtained from the entity-to-region alignments.

More specifically, we utilize the transformer-based architecture for the adapter. The queries, keys, and values are reduced embeddings of matched entities and video frames. Take one frame as an example: query $\mathbf{q}$ is calculated by first summing the text embedding $\mathbf{S}$ weighted by the similarity matrix $\widehat{\mathbf{A}}$ then reducing to a vector by MLP. Meanwhile, key and value are the same vector $\widehat{\mathbf{e}}$, obtained by processing the image embedding $\mathbf{E}$ with another MLP:

$$\mathbf{q} = \text{MLP}(\widehat{\mathbf{A}}\mathbf{S}), \quad \widehat{\mathbf{e}} = \text{MLP}(\mathbf{E}). \tag{6}$$

Given a video with T frames $[\mathbf{I}^1, \mathbf{I}^2, ..., \mathbf{I}^T]$, the initial queries $\mathbf{Q}_0$ to the adapter are query vectors of different

frames, while the keys and values are $\widehat{\mathbf{E}}$, which is the processed image embeddings of input frames (Eq. 6):

$$\mathbf{Q}_0 = [\mathbf{q}^1, ..., \mathbf{q}^T], \quad \widehat{\mathbf{E}} = [\widehat{\mathbf{e}}^1, ..., \widehat{\mathbf{e}}^T], \tag{7}$$

The structure of our adapter is illustrated in Fig. 3c. This adapter includes a sequence decoding block that consists of a 1D-convolution module, a self-attention (SA) module, and a cross-attention (CA) module. The attention modules function in the same way as the ones in the transformer [61]. The procedure of the video adapter can be formulated as:

$$\begin{aligned} \mathbf{Q}'_m &= \mathrm{SA}_m(\mathbf{Q}_{m-1}), \quad \widehat{\mathbf{E}}_m = \widehat{\mathbf{E}} + \text{1D-Conv}_m(\widehat{\mathbf{E}}), \\ \mathbf{Q}_m &= \mathrm{CA}_m(\mathbf{Q}'_m, \widehat{\mathbf{E}}_m), \quad m = 1, ..., M, \end{aligned} \tag{8}$$

where m indicates the block index of the video adapter. The SA module and the 1D-convolution module serve for temporal interactions for $\mathbf{Q}_{m-1}$ and $\widehat{\mathbf{E}}$, respectively. The CA module utilizes evolved text embeddings as queries to aggregate to the visual embeddings across the frames, preserving the entity-level visual information during the adaption. The output query after M blocks is $\mathbf{Q}_M$. We obtain the final video representation $\mathbf{z}$ by applying Average Pooling and MLP layer over the $\mathbf{Q}_M$ sequentially along the temporal dimension and the feature channel:

$$\mathbf{z} = \mathrm{MLP}(\mathrm{AvgPool}(\mathbf{Q}_M)). \tag{9}$$

3.3. Training Details

Loss function. Our proposed network aims to maximize the similarity between video representations and textual representations of corresponding action labels. Specifically, we utilize the frozen text encoder of CLIP to perform prompt ensembling for action labels with a bunch of handcraft templates [43]. Given the representations of the ith action $\mathbf{c}_i$ and nth video $\mathbf{z}_n$ (as described in Eq. (9)), the loss function can be implemented by the cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{I} y^{i,n} \log\left(\frac{\exp(\mathbf{c}_i^\top \mathbf{z}_n)}{\sum_{j=1}^{I}\exp(\mathbf{c}_j^\top \mathbf{z}_n)}\right). \tag{10}$$

The training set has N videos belonging to the I actions. If the nth video belongs to the ith action, $y^{i,n}$ equals 1; otherwise, $y^{i,n}$ equals 0.

Network training. The ViT backbone in the region-aware image encoder is initialized by CLIP, while the token merging module is parameter-free with the reduction number r to be 8. The number of blocks in the multi-modal video decoder is set to 4 and 6 for ViT-B and ViT-L backbones, respectively. We adopt an AdamW optimizer for network parameter training with initial learning rates of 8×10^{-6} for the ViT backbone and 8×10^{-5} for the remaining parts. The networks are trained with 30 epochs (including a five-epoch warmup) and a weight decay of 0.001 w.r.t. a cosine schedule. The input video follows the main sparse sampling method [63] and augmentation strategy [43] with a frame resolution 224×224. Experiments are conducted with 8 32GB-V100-GPUs.

4. Experiments

Datasets. Our proposed method is evaluated on four widely used video action recognition datasets: Kinetics-400 [28], Kinetics-600 [7], UCF-101 [55], HMDB-51 [29], and Something Something V2 [20](see in supplementary materials). Kinetics-400 consists of approximately 240k training and 20k validation videos, covering 400 classes, with each clip spanning around 10 seconds. Kinetics-600 is an extension of Kinetics-400, including around 410k training and 29k validation videos for 600 classes. UCF-101 contains 13,320 video clips with 101 classes, and HMDB-51 consists of 7,000 videos with 51 classes. We conduct fully-supervised experiments on Kinetics-400 and Kinetics-600. Additionally, for Kinetics-600, UCF-101, and HMDB-51, we perform few-shot and zero-shot experiments with models pre-trained on Kinetics-400.

4.1. Fully Supervised Comparison

Settings. We conduct fully-supervised experiments on Kinetics-400. Each video clip is sampled with 16 or 32 frames. Two variants of the network, namely ALT-B/16 and ALT-L/14, employ ViT-B/16 and ViT-L/14, respectively. The results on Kinetics-600 and Something-Something-v2 [20] are exhibited in the supplementary materials.

Results. In Tab. 1, we compare with the state-the-of-art methods on Kinetics-400 with the input resolution 224×224. Taking 16 sampled frames of each video as input, ALT-B/16 achieves *84.8%* top-1 accuracy with only 657 GFLOPs. When the input frames increase to 32, ALT-B/16 surpasses the performance of ViViT-H/16 [1], which takes more than 30× computation cost (1308 vs. 48916 GFLOPs). By employing the larger backbone, ALT-L/14 achieves superior performance with *88.1%* top-1 accuracy among CLIP-400M pretraining works and significant computational advantage (*0.4%* higher than AIM [71] but *55%* fewer GFLOPs). It is noteworthy that the leading method MTV-H [70] adopts larger-scale pretraining data (70M video-text pairs with about 17B images) and consumes 9× GFLOPs. We attribute the superiority and efficiency of ALT to the seamless coupling of entity-to-region alignments token merging. As shown in Fig. 4 right, we visualize the '*performance v.s. GFLOPs*' relationships of some representative works, where ALT sets new Pareto frontiers.

4.2. Few-shot Comparisons

Settings. We evaluate our few-shot experiments on the HMDB-51 and UCF-101 datasets, utilizing ALTs trained on Kinetics-400 data and a text corpus. For the training set

Method	Pretrain	Top-1	Top-5	GFLOPs	Frames	Views	#Param.(M)
Methods with ImageNet or web-scale image pretraining							
MViTv1-B [12]	-	81.2	95.1	4095	64	3×3	36.6
Uniformer-B [32]	IN-1k	83.0	95.4	3108	32	4×3	50.0
TimeSformer-L [3]	IN-21k	80.7	94.7	7140	64	1×3	121.4
VideoSwin-L [38]	IN-21k	83.1	95.9	7248	32	4×3	200.0
ViViT-H/16 [1]	JFT-300M	84.9	95.8	48916	32	4×3	647.5
Methods with web-scale language-image pretraining							
MTV-H/16 [70]	WTS-17B*	89.1	98.2	45537	32	4×3	-
PromptingCLIP-B/16 [27]	CLIP-400M	76.9	93.5	-	16	5×1	95.5
ActionCLIP-B/16 [64]	CLIP-400M	83.8	97.1	16890	32	10×3	105.2
ViFi-CLIP [52]	CLIP-400M	83.9	96.3	3372	16	4×3	124.7
ASU-B/16 [11]	CLIP-400M	84.7	96.8	3444	16	4×3	132
ST-Adapter-L/14 [45]	CLIP-400M	87.2	97.6	8248	32	3×1	-
EVL-L/14 [37]	CLIP-400M	87.3	97.6	8088	32	3×1	357.9
AIM-L/14 [71]	CLIP-400M	87.5	97.7	11208	32	3×1	341
X-CLIP-L/14 [43]	CLIP-400M	87.1	97.6	7896	8	4×3	451.2
X-CLIP-L/14(336↑)† [43]	CLIP-400M	87.7	97.4	37032	16	4×3	451.2
Our method							
ALT-B/16	CLIP-400M	84.8	96.4	657	16	3×1	134.4
ALT-B/16	CLIP-400M	85.5	96.7	1308	32	3×1	134.4
ALT-L/14	CLIP-400M	87.8	97.6	2478	16	3×1	437.1
ALT-L/14	CLIP-400M	88.1	97.7	4947	32	3×1	437.1

Table 1. Comparison to state-of-the-art on Kinetics-400. Views are denoted with *"temporal clips × spatial crops."* Parameters in text encoders are not counted. * indicates pretraining with a video-text collection. † indicates the input frame size is 336×336.

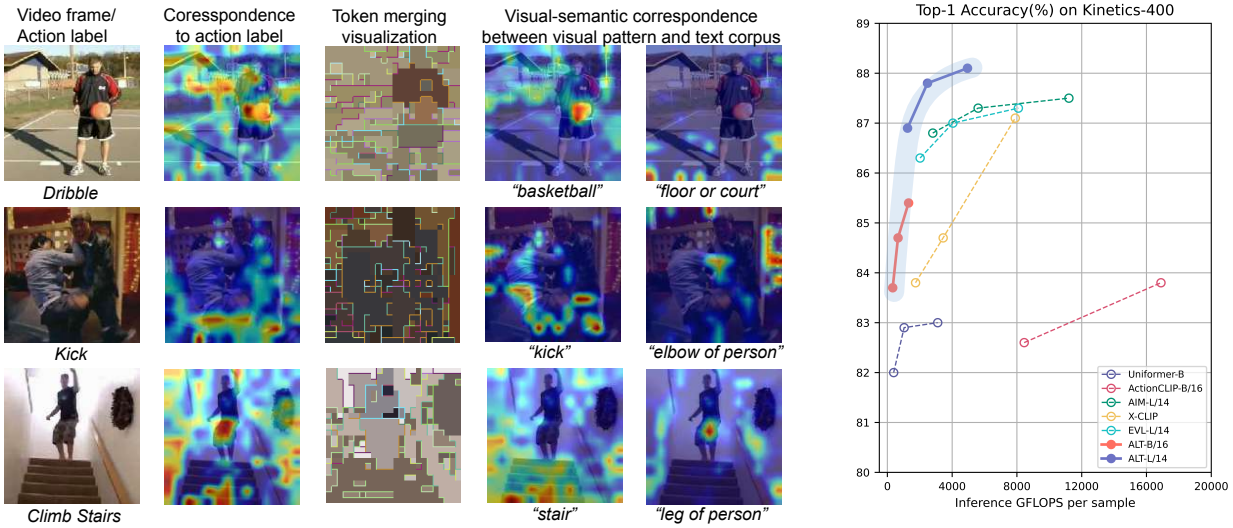

Figure 4. **Left:** Visualization of visual-semantic correspondences with the tool [9]. For each row: Column (2) visualizes the visual correspondence to text entities generated by the action label; Column (3) visualizes region-aware embeddings under ToMe; Column (4) and (5) show the two of the fine-grained corresponding visual patterns to specific text entities, which are geometrically consistent with Column (3). **Right:** Visualization of Accuracy v.s. FLOPs performance.

construction, we randomly sample 2, 4, 8, and 16 videos from each class, and we set the frame number in each video to either 8 or 32. Following the protocols of X-CLIP [43], we use the first split of the test set for evaluation.

Results. Sec. 3.3 shows the performance comparison on K-shot learning. Our method significantly outperforms image-pretrained methods. For instance, when K=2, ALT-B/16 surpasses VideoSwin-B [38] by *43.4%* on HMDB-51 and *39.9%* on UCF-101. Among the approaches that leverage image-language pretraining: In two-shot scenarios, ALT-B/16 surpasses the previous state of arts by *7.1%* and *9.2%* on HMDB-51 and UCF-101, respectively. The lead remains consistent across 16-shot scenarios and continues to expand when switching to ALT-L/14. This showcases the superior generalization capabilities of our paradigm, which establishes reusable entity-to-region alignments.

Method	Frames	GFLOPs	HMDB-51				UCF-101			
			K=2	K=4	K=8	K=16	K=2	K=4	K=8	K=16
TSM [36]	32	-	17.5	20.9	18.4	31.0	25.3	47.0	64.4	61.0
TimeSformer [3]	32	238	19.6	40.6	49.4	55.4	48.5	75.6	83.7	89.4
VideoSwin-B [38]	32	321	20.9	41.3	47.9	56.1	53.3	74.1	85.8	88.7
ActionCLIP [64]	8	141	55.0	56.0	58.0	-	80.0	85.0	89.0	-
X-CLIP-B/16 [43]	32	658	53.0	57.3	62.8	64.0	76.4	83.4	88.3	91.4
X-Florence [43]	32	2822	51.6	57.8	64.1	64.2	84.0	88.5	92.5	94.8
ViFi-CLIP [52]	32	562	57.2	62.7	64.5	66.8	80.7	85.1	90.0	92.7
ALT-B/16	32	436	64.3	66.7	70.4	74.5	93.2	95.3	96.4	97.3
ALT-L/14	32	1649	**68.1**	**69.7**	**73.8**	**79.1**	**96.0**	**97.4**	**98.0**	**98.1**

Table 2. Few-shot comparison: we compare ALTs with previous SOTAs on HMDB-51 and UCF-101. All the models are trained on Kinetics-400, with top-1 accuracies(%) reported under a single-view inference.

Method	HMDB-51	UCF-101
ASR [65]	21.8±0.9	24.4±1.0
ZSECOC [47]	22.6±1.2	15.1±1.7
UR [79]	24.4±1.6	17.5±1.6
TS-GCN [17]	23.2±3.0	34.2±3.1
ER-ZSAR [10]	35.3±4.6	51.8±2.9
ActionCLIP [64]	40.8±5.4	58.3±3.4
X-CLIP-B/16 [43]	44.6±5.2	72.0±2.3
ASU-B/16 [11]	48.1±2.8	75.0±3.7
ViFi-CLIP [52]	51.3±0.6	76.8±0.7
ALT-B/16	52.9±1.0	79.4±0.9
ALT-L/14	**56.6±0.8**	**83.9±1.1**

Method	Top-1	Top-5
DEVISE [16]	23.8±0.3	51.0±0.6
ESZSL [53]	22.9±1.2	48.3±0.8
DEM [77]	23.6±0.7	49.5±0.4
GCN [19]	22.3±0.6	49.7±0.6
ER-ZSAR [10]	42.1±1.4	73.1±0.3
ActionCLIP [64]	66.7±1.1	91.6±0.3
X-CLIP-B/16 [43]	65.2±0.4	86.1±0.8
ASU-B/16 [11]	67.6±0.2	87.2±0.3
ViFi-CLIP [52]	71.2±1.0	92.2±0.3
ALT-B/16	72.7±0.6	91.7±0.4
ALT-L/14	**74.9±0.4**	**92.2±0.3**

Table 3. Zero-shot comparison between ALTs with representative image & image-language pretraining methods on HMDB-51 & UCF-101 (left) and Kinetics-600 (right). Pretrained on Kinetics-400, the accuracies(%) are reported under a single-view inference.

4.3. Zero-shot Comparisons

Settings. For zero-shot evaluation, we train ALTs on Kinetics-400 data and corpus, following the protocol outlined in the [43]: For HMDB-51 and UCF-101, we conducted experiments using the three provided splits. Regarding Kinetics-600, the test set is constructed by randomly selecting 160 categories, three times, from 220 categories that are distinct from those in Kinetics-400. We report results in the format of *"average accuracy ± standard deviations."*
Results. We present the zero-shot results in Tab. 3. ALT-B/16 outperforms ViFi-CLIP by *1.6%*, *2.6%*, and *1.5%* in terms of top-1 accuracy on HMDB-51, UCF-101, and Kinetics-600, respectively. It is noteworthy that our approach requires *22%* fewer GFLOPs and ViFi-CLIP unfreezes the CLIP text encoder in the pretraining stage. We attribute the superiority to the utilization of the text corpus, whose factorized and reusable semantics mitigate the difficulty of adapting our model to a new scenario.

4.4. Ablation Study

We employ ALT-B/16 to conduct detailed ablation experiments. By default, the fully-supervised experiments are evaluated on Kinetics-400 with 8 frames per video clip. Taking 32 frames per sample as input, The few-shot and zero-shot experiments are conducted on the first split of HMDB-51 and UCF-101, respectively. With 32 frames per video clip, we report results under a single-view inference.

Component analysis. We investigate the effectiveness of the proposed components and report the results in Tab. 4: (a) We consider X-CLIP-B/16 [43] (without text prompts) as the baseline. It incorporates a cross-frame module within the CLIP image encoder and follows the "align then adapt" approach. The top-1 accuracy on Kinetics-400 is *81.7%*. (b) Then we enhance the baseline with our framework but only using the text embedding of the action name rather than the corpus of entities in Eq.4. The improvement reveals the effectiveness of the proposed video adapter, which utilizes text embedding to guide the adaption from image embeddings to the final video embedding. (c) When we replace the action name with the corpus of entities, the results further increase especially in the few/zero-shot scenarios. The phenomenon indicates using the relevant entities can achieve a better generalization ability in action recognition. (d) We further change the global visual image to region-aware embeddings in the alignment(Eq.4), fulfilling the 'entities-to-regions' alignment. Our approach finally improves the baselines by *0.9%*, *11.3%*, and *7.4%*, respectively.

Video adapter component analysis. One important function of our proposed video adapter is to enable modality interactions of established alignments through the cross-attention (CA) modules. To examine its effects, in Tab. 5: (a) we initially set up a baseline by substituting the video adapter with four self-attention layers. (b) Then we evaluate the performance with a CA-only video adapter, which im-

	Align.	Corpus.	Region.	Fully.	2-shot	0-shot
a.	×	×	×	81.7	53.0	72.0
b.	✓	×	×	82.0	55.2	74.4
c.	✓	✓	×	82.2	58.1	76.9
d.	✓	✓	✓	**82.8**	**64.3**	**79.4**

Table 4. Effect of proposed components. ***Align***: 'Alignment' before adaptation step; ***Corpus***: Utilize action-related text corpus; ***Region***: Empowered by region-aware visual embeddings. 2-shot and 0-shot are evaluated on HMDB-51 and UCF-101, respectively.

	Method	Top-1 Acc.(%)	GFLOPs	Param. (M)	Tunable Param.(M)
a.	Frozen	81.4	110	134	**38**
b.	S-adapter [71]	81.7	116	138	42
c.	SM-adapter [71]	81.8	123	141	45
d.	STM-adapter [71]	82.5	163	145	49
e.	Finetune	**82.8**	**110**	**134**	134

Table 6. Effect of different training strategies on the image encoder. ***S***: spatial; ***M***: MLP; ***T***: temporal; ***Param.***: # parameters. We report Top-1 Acc. on Kinetics-400 under fully-supervised settings.

proves the baseline significantly, validating the effectiveness of 'entity-to-region' alignments. (c)-(e) Additionally, we investigated the effects of 1D-convolution and self-attention (SA) modules which enable spatiotemporal signal communication. The results reveal that both modules are beneficial and compatible with each other.

Training strategy and efficiency of image encoder. Our method finetunes the pre-trained image encoder, which influences the final performance as shown in Tab. 6: (a) when freezing the image encoder during training, fewer tunable parameters are required while the accuracy decreases to *81.4%*. (b)-(d) Based on the 'frozen' setting, we leverage approaches proposed by AIM [71], which introduces adapters into the image encoder. The performances are improved by stacking spatial, temporal, and MLP adapters in the transformer blocks. It is noteworthy that the STM method (d) computes the attention layers twice, therefore significantly increasing the computational complexity of our framework (e) by *48%*.

Finetune strategy in few-shot scenarios. We investigate different finetuning strategies in few-shot learning, and the results are shown in Tab. 7: (a) STM-adapter [71] can significantly improve the frozen model baseline (*52.5%*) with a small number of training parameters, (b) Finetuning the whole visual branch further improves performance without overfitting. (c) The recent prompt tuning method, VL-prompt* [52], demonstrates impressive performance in few-shot learning. (d) We adopt the technique in ALT by freezing the backbone and adding ten learnable tokens to each layer of the image & text encoders. This modification leads to improved accuracies when dealing with lower shots.

Linear evaluation on learned representations. To assess

	CA	SA	1D-Conv	Fully.	2-shot	0-shot
a.	×	×	×	81.1	55.0	68.3
b.	✓	×	×	81.8	61.3	74.6
c.	✓	✓	×	82.5	62.7	76.3
d.	✓	×	✓	82.3	62.1	76.1
e.	✓	✓	✓	**82.8**	**64.3**	**79.4**

Table 5. Effect of components in the video adapter: ***CA***: cross attention module; ***SA***: Self attention module; ***1D-Conv***: 1D convolution module. Top-1 Acc. are reported in a single view. 2-shot and 0-shot are evaluated on HMDB-51 and UCF-101, respectively.

	Method	Tunable Param.(M)	2-shot	8-shot	16-shot
a.	STM-adapter [71]	49	58.1	66.9	71.0
b.	V-Finetune	125	64.3	70.4	**74.5**
c.	VL-prompt* [52]	0.15(+63)	63.0	69.6	72.0
d.	ALT+VL-prompt	0.15+38.4	**65.3**	**71.1**	74.2

Table 7. Analysis on different finetune strategies in HMDB-51 few-shot learning. The base model is ALT-B except * employs ViFi-CLIP, which additionally pretrains the text encoder(63M paramters).

Dataset	Split 1	Split 2	Split 3	Average	Benchmark [57]
UCF-101	95.6	95.8	96.1	95.83	**96.1**
HMDB-51	73.8	73.5	74.0	**73.77**	73.3

Table 8. Linear evaluation on ALT-B/16 (Kinetics-400 pretrained) with Top-1 Acc. reported under a single view. *Benchmark* denotes VideoMAE that fully trained on Kinetics-400 and target datasets.

the quality of the video representations obtained, we conduct linear probe experiments on UCF101 and HMDB-51 datasets with a frozen ALT-B/16. Specifically, the representations of ALT are fixed and fed into a tunable linear classification layer with one-hot label supervision. The top-1 accuracies are presented in Tab. 8 and compared with VideoMAE [57], which adopts video reconstruction supervision to learn powerful representations and achieve impressive recognition performance. Notably, our linear-probed results are comparable to a fully-trained VideoMAE, underscoring the discrimination and generalization capabilities of our approach.

5. Conclusion

In this paper, we propose a novel method with the "align before adapt" paradigm for video action recognition. By leveraging the alignments between local visual appearance and action-related entity semantics, we conduct a better video representation adaption with improved interpretability and generalizability. Our method demonstrates superior performance especially in zero-shot and few-shot scenarios while maintaining low computational costs.

References

[1] 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, 5, 6

[2] Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. An enhanced lesk word sense disambiguation algorithm through a distributional semantic model. In *COLING*, pages 1591–1600. ACL, 2014. 3

[3] 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

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

[5] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *CoRR*, abs/2210.09461, 2022. 3, 4

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

[7] João Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *CoRR*, abs/1808.01340, 2018. 3, 5

[8] Chun-Fu Chen, Rameswar Panda, and Quanfu Fan. Regionvit: Regional-to-local attention for vision transformers. In *ICLR*. OpenReview.net, 2022. 3

[9] Peijie Chen, Qi Li, Saad Biaz, Trung Bui, and Anh Nguyen. gscorecam: What objects is CLIP looking at? In *ACCV (4)*, pages 588–604. Springer, 2022. 6

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

[11] Yifei Chen, Dapeng Chen, Ruijin Liu, Hao Li, and Wei Peng. Video action recognition with attentive semantic units. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10170–10180, 2023. 6, 7

[12] 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

[13] 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

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

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

[16] 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

[17] 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

[18] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV (36)*, pages 540–557. Springer, 2022. 2

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

[20] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *ICCV*, pages 5843–5851. IEEE Computer Society, 2017. 5

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

[22] 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

[23] Ali Hatamizadeh, Hongxu Yin, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Global context vision transformers. In *ICML*, pages 12633–12646. PMLR, 2023. 3

[24] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017. 3

[25] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *ICLR (Poster)*. OpenReview.net, 2017. 4

[26] 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*, pages 4904–4916. PMLR, 2021. 1, 2

[27] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding, 2022. 1, 6

[28] 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. 5

[29] 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. 3, 5

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

[31] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900. PMLR, 2022. 1, 2

[32] 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

[33] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, pages 10955–10965. IEEE, 2022. 2, 3

[34] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV (30)*, pages 121–137, 2020. 2

[35] 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

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

[37] 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, 2022. 1, 2, 6

[38] 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, 3, 6, 7

[39] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *ICLR (Poster)*. OpenReview.net, 2017. 4

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

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

[42] Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. Clip-it! language-guided video summarization, 2021. 2

[43] 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

[44] OpenAI. Chatgpt: Optimizing language models for dialogue. https://chat.openai.com/, 2022. 3

[45] 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. 1, 2, 6

[46] 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

[47] 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

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

[49] 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

[50] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125, 2022. 2

[51] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *CVPR*, pages 18061–18070. IEEE, 2022. 3

[52] Hanoona Abdul Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman H. Khan, and Fahad Shahbaz Khan. Fine-tuned CLIP models are efficient video learners. In *CVPR*, pages 6545–6554. IEEE, 2023. 2, 6, 7, 8

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

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

[55] 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. 3, 5

[56] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: pre-training of generic visual-linguistic representations. In *ICLR*. OpenReview.net, 2020. 2

[57] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022. 8

[58] 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

[59] 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

[60] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NIPS*, pages 6306–6315, 2017. 4

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 5

[62] Jan Philip Wahle, Terry Ruas, Norman Meuschke, and Bela Gipp. Incorporating word sense disambiguation in neural language models. *arXiv preprint arXiv:2106.07967*, 2021. 3

[63] 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

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

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

[66] 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

[67] 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

[68] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas M. Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *CVPR*, pages 18113–18123. IEEE, 2022. 4

[69] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. *CVPR*, abs/2303.04803, 2023. 2

[70] 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, 5, 6

[71] Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. AIM: adapting image models for efficient video action recognition. *CoRR*, abs/2302.03024, 2023. 1, 2, 5, 6, 8

[72] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: fine-grained interactive language-image pre-training. In *ICLR*. OpenReview.net, 2022. 3

[73] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *CoRR*, abs/2205.01917, 2022. 1, 2

[74] 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. 2

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

[76] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *ICML*, pages 25994–26009. PMLR, 2022. 2

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

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

[79] 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