https://doi.org/10.2352/EI.2025.37.8.IMAGE-263

ZAR: Zero-Shot Action Recognition with Dynamic Prompt Tuning

***Qiyue Liang*[1], *Cheng Lu*[2], *Chun Tao*[1], *and Jan P. Allebach*[1]**
[1] ***Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907***
[2] ***Xmotors.ai, Santa Clara, CA 95054***

Abstract

Pre-trained vision-language models, exemplified by CLIP, have exhibited promising zero-shot capabilities across various downstream tasks. Trained on image-text pairs, CLIP is naturally extendable to video-based action recognition, due to the similarity between processing images and video frames. To leverage this inherent synergy, numerous efforts have been directed towards adapting CLIP for action recognition tasks in videos. However, the specific methodologies for this adaptation remain an open question. Common approaches include prompt tuning and fine-tuning with or without extra model components on video-based action recognition tasks. Nonetheless, such adaptations may compromise the generalizability of the original CLIP framework and also necessitate the acquisition of new training data, thereby undermining its inherent zero-shot capabilities. In this study, we propose zero-shot action recognition (ZAR) by adapting the CLIP pre-trained model without the need for additional training datasets. Our approach leverages the entropy minimization technique, utilizing the current video test sample and augmenting it with varying frame rates. We encourage the model to make consistent decisions, and use this consistency to dynamically update a prompt learner during inference. Experimental results demonstrate that our ZAR method achieves state-of-the-art zero-shot performance on the Kinetics-600, HMDB51, and UCF101 datasets.

Overall Document Guidelines: Head

Advancements in natural language processing, epitomized by large-scale language models like BERT[1], GPT[2], ERNIE[3], and T5[4], showcase an impressive capacity to grasp contextual intricacies and exhibit proficiency in few-shot and zero-shot learning settings. This remarkable strength has ignited a surge of interest in the computer vision domain. For example, vision-language models (VLMs), such as CLIP [5] and ALIGN [6], leverage the contextual information encoded within text data, paving the way for zero-shot transfer to a myriad of downstream tasks. This paradigm has prompted exploration into leveraging the knowledge encapsulated within these pre-trained VLMs for zero-shot video recognition tasks.

However, bridging the gap between image and video recognition poses a notable challenge. Unlike static images, video data inherently encapsulate temporal information. While VLMs effectively harness vast image-text datasets for contrastive learning, they lack the inherent temporal cues encoded within video data. To address the modality gap associated with leveraging image-text pre-trained models for video recognition tasks, current explorations can be broadly categorized into two approaches. One approach involves augmenting the original vision-language model with additional components, such as a decoder[7] or spatial-temporal extension[8], to accommodate temporal dynamics. While these methods have demonstrated efficacy in video action recognition tasks, full fine-tuning of the model imposes significant demands on GPU resources, rendering it time-consuming and inaccessible for some applications. Moreover, considering that the original CLIP model was trained on an extensive dataset comprising 400 million image-text pairs, fine-tuning requires a comparably large dataset to avoid overfitting issues. Even when appropriately fine-tuned, models often excel on the tuned dataset but struggle with generalization to unseen datasets, thereby impeding the original zero-shot learning capabilities of CLIP. An alternative approach, spurred by recent research efforts in the natural language processing domain, entails prompt tuning[7]—a method that adds a few additional learnable prompts while keeping the original model backbone frozen. Prompting is considered efficient and demands less computational resources and training time compared to full fine-tuning techniques. Under a fully supervised setting, prompt tuning can facilitate effective adaptation from the image to the vision domain. However, it is essential to note that prompt tuning still necessitates a substantial amount of labeled data.

Given the aforementioned obstacles, we propose ZAR (zero-shot action recognition), an innovative solution that circumvents the requirements of additional labeled training data. ZAR operates by dynamically tuning the prompt exclusively with the provided test video through entropy minimization. Entropy minimization allows the model to learn from the inherent distribution of the data itself, and operates independently of a specific training regimen, making it particularly well-suited for zero-shot settings where additional labels are not provided. Additionally, we propose the utilization of multiple frame rates from the same test video, a simple yet powerful approach to effectively capture diverse temporal dynamics. The on-the-fly prompt adaptation ensures that ZAR is finely tailored to each specific video, and facilitates zero-shot action recognition generalization without the need for additional training data or annotations. This not only streamlines the process but also enhances the model's versatility and applicability across a broad spectrum of video recognition tasks. This approach enhances the model's ability to make informed decisions, ultimately leading to improved performance.

Our main contributions are as follows:

- We introduce ZAR (zero-shot action recognition), that targets video action recognition tasks by obviating the need for additional training data. ZAR uniquely leverages prompt tuning on the fly using only the current test sample. To the best of our knowl-

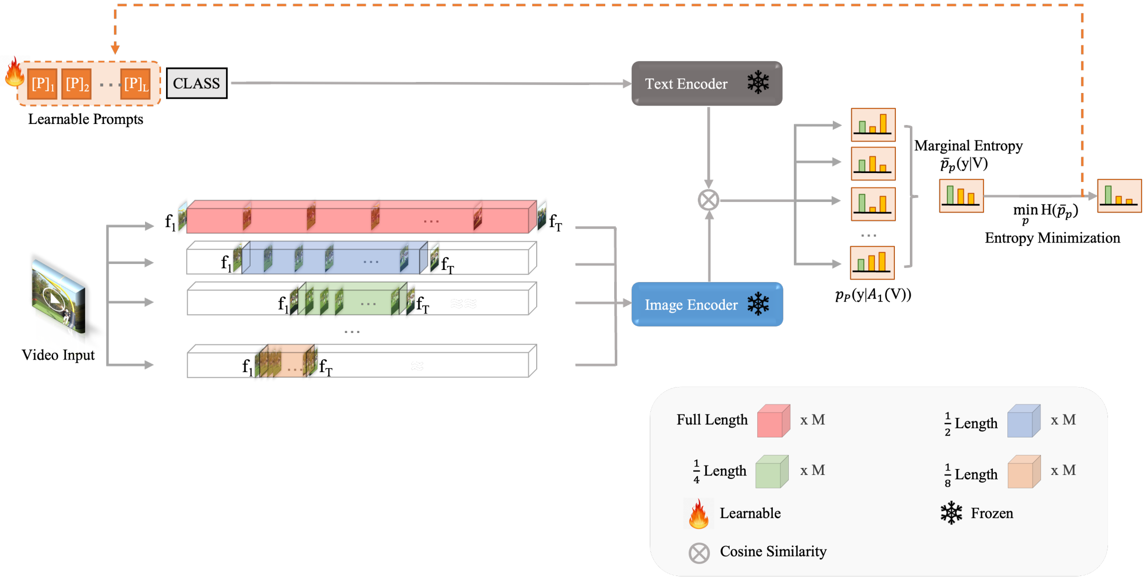


Figure 1. *An overview of **ZAR** for video action recognition.*

edge, we are the first to use prompt-tuning in a zero-shot manner in the video action recognition domain.

- We propose the incorporation of multiple frame rates to augment decision-making within the model, fostering consensus and enhancing accuracy. This innovative technique provides a robust framework for optimizing performance in video action recognition.

- We conduct experiments to comprehensively investigate the impact of various frame rate selections and the percentage of frames utilized for entropy minimization. Furthermore, we compare our results across diverse datasets against established baselines, providing valuable insights into the efficacy and versatility of our proposed approach.

Related Works

Vision-Language Models (VLMs). The integration of multi-modal representations through large-scale image-text pretraining [7, 9, 10, 11] has become a cornerstone in the development of many applications. The key concept underlying the pretraining of foundational vision language models (VLMs) such as CLIP [5] and ALIGN [6] is to immerse them in a plethora of image-text pairs, thereby fostering an understanding of the semantic connections between images and their accompanying textual descriptions within a shared embedding space by contrasting coherent pairs with incoherent ones. Acquiring robust image-textual representations, CLIP can operate in an open vocabulary manner, liberating it from constraints associated with predefined sets of words or classes during training. This characteristic empowers it to undertake an extensive array of downstream tasks, encompassing classification [12, 13, 14], object detection [15, 16, 17], image or video text retrieval [18, 19, 20], among others, with notable accuracy and efficiency. While extending pretrained image-text models to video processing seems intuitive, a domain gap exists between image and video processing due to the temporal clues present in videos, which are absent in static image data. Common approaches to bridge this gap and adapt CLIP to the video domain include integrating separate decoder components [21, 7], employing auxiliary branches [22], and incorporating spatial-temporal components [8]. While these approaches may achieve favorable outcomes on video datasets, they typically entail fine-tuning the CLIP model, thereby potentially diminishing its inherent open vocabulary and zero-shot capabilities.

Prompt Tuning. Prompting [23] is a prevalent technique originated from natural language processing, utilized to tailor models for specific tasks or datasets in a heuristic way. This involves crafting specific prompts or input templates that guide the model to generate desired outputs, aligning with the task requirements or dataset characteristics. Advancements have facilitated the automation of this process [24, 25] through the incorporation of a few trainable tokens into the model while preserving the frozen backbone, thus ensuring the model's generalizability. CoOp[26] incorporates additional tokens into the input prompt of the CLIP model, enabling the conditioning of its predictions on specific attributes or characteristics of the input data. CoCoOp [27] extends the capabilities of CoOp by generating an input-conditional token vector for each image to address a notable limitation in CoOp: the learned context in CoOp may not generalize effectively to unseen classes within the same dataset. Vifi-CLIP [28] is a video action recognition model that introduces a two-step approach: firstly, it undergoes fine-tuning on a video dataset to address the disparity between image and video modalities; subsequently, prompts are added to both the image and text branches while freezing the backbone. Although these prompt tuning methods offer promising approaches for adapting pre-trained models such as CLIP to specific tasks with reduced parameter overhead compared to fine-tuning techniques, their efficacy in automatically generating effective prompts hinges on the availability of labeled data to accurately capture the task requirements, which poses a challenge to the zero-shot capability of the models.

Test-time Learning. Applying machine learning models in a zero-shot manner [29, 30, 31] presents a primary challenge that require effective solutions: it is critical to adapt the model to make predictions for unseen categories or tasks without necessitating extra labeled data. Test-time training (TTT) and its variants [32, 33] introduce an approach to model adaptation during inference by integrating a self-supervised branch that operates by computing an optimization objective tailored to the test sample. In TTT, the selection of test-time optimization objectives relies on a proxy task. However, as emphasized by its authors, the choice of proxy task is critical and must be " both well-defined and non-trivial " in the new domain. Another option for the test-time optimization objective is entropy minimization [34, 35]. This involves regularizing entropy, as seen in methods like TENT (Test-time Entropy Minimization) [36], which imposes a penalty on decisions made by the model at high-density regions within the data distribution [37]. However, TENT requires multiple test samples to operate effectively. MEMO [38] bypasses the multi-sample requirements using data augmentations and proves that updating the entire model at test time on the augmentations of a single test sample yields robust test-time adaption results. TPT [39] expands on this research by introducing an additional confidence selection mechanism for data augmentation, and refrains from utilizing all the augmented samples for backpropagation. It also opts to choose the parameter to be the text prompt learner instead of the entire model. Our approach operates by selecting a different augmentation method, leveraging the temporal information inherent in videos.

Methodology

Preliminaries

Using CLIP for zero-shot classification We commence by delineating the conventional procedure for applying CLIP to downstream classification tasks in a zero-shot fashion. In the context of video action recognition, given a single test video, we sample T frames from the video $\{f_1, f_2, ..., f_T\}$ to form the input $V \in \mathbb{R}^{T\times C\times H\times W}$ belonging to class $y_i \in Y$. Here, T represent the number of sampled frames from the video, C denotes the color channel, and H and W represents the spatial dimensions. $Y = \{y_1, y_2, ...y_K\}$ denotes the set of all classes, where y_i signifies one class out of a total of K classes. The frames of the video are processed by the image encoder of the CLIP model to generate image embeddings, which are subsequently aggregated to obtain a video feature representation v. Concurrently, the class names are passed through the text encoder of the CLIP model to obtain text embeddings $\{t_1, t_2, ..., t_k\}$ where t_i represents the text embedding of the i^{th} class. The similarity score $S_i(v, t_i)$ is obtained by calculating the cosine similarity between the video embedding v and each text feature t_i:

$$S_i(v, t_i) = \frac{<v, t_i>}{\|v\|\|t_i\|}. \tag{1}$$

The probability of the given video V belonging to class y_i can be expressed using conditional probability as:

$$p(y_i \mid V) = \frac{\exp(S_i(v, t_i)/\tau)}{\Sigma_{i=1}^{K} \exp(S_i(v, t_i)/\tau)}, \tag{2}$$

where τ refers to the learned temperature of the softmax function. This temperature helps control the sharpness or smoothness of the resulting probability distribution. During zero-shot inference, the model selects class y_i that maximizes the prediction probability $p(y_i \mid V)$, indicating that it is the most likely class according to the model's predictions for the input video.

Prompt-Tuning for Model Adaptation Notwithstanding its commendable performance, the direct application of CLIP to video-based tasks in a zero-shot manner encounters challenges arising from the intrinsic temporal dynamics inherent in videos. These dynamics, absent in the static image data utilized for pre-training, necessitate tailored adaptation strategies to effectively leverage the model's capabilities in video analysis tasks. Prompt tuning is a prevalent approach for adapting models to downstream tasks. Formally, for the context prompt setting (where the same prompts are prepended to each class label, the input provided to the text encoder is represented as:

$$[P]_1[P]_2...[P]_L[y_i], \tag{3}$$

where $[P]_j$ for $(j = 1, 2, ..., L)$ are learnable prompt vectors, L denotes the number of prompt vectors, and each prompt vector $[P]_j$ has the same dimension D as the word embeddings ($D = 512$ for CLIP). $[y_i]$ represents the class label token. The goal of prompt tuning is to learn an optimal prompt $P = [P]_1[P]_2...[P]_L \in \mathbb{R}^{L\times D}$. This optimal prompt, when combined with the class labels and passed through the text encoder, provides the most helpful context information about the downstream task, leading to improved performance in terms of generating more accurate prediction. As this is a classification task, the objective of improving prediction accuracy can be formulated by minimizing the cross-entropy loss $\mathscr{L}$:

$$\mathscr{L} = -\mathbb{E}_{\{V,y\}}\left[\frac{1}{K}\sum_{i=1}^{K}\log p(y_i \mid V)\right], \tag{4}$$

where $\mathbb{E}_{\{V,y\}}$ represents the expected input video provided by a training set with labels, K is the number of classes, and $p(y_i \mid V)$ denotes the conditional probability calculated based on the similarity score between the input video features and the class embeddings prepended with prompts. Although the CLIP backbone remains frozen during prompt tuning, with the only learnable parameters being the few added tokens, this approach still relies on a labeled training set and trains for a few shots (16 shots for CoOp [26]) to achieve optimal results.

Evaluation settings. We define two evaluation settings to assess the performance of our method. In the first setting, we conduct a **zero-shot** evaluation, wherein the model is trained on a pretraining dataset D_p with a set of classes Y_p and subsequently evaluated on a target dataset D_t with a distinct set of classes Y_t, ensuring no class overlap: $Y_p \cap Y_t = \emptyset$. In the second setting, known as **base-to-novel** setting, we follow the dataset split delineated in [28], dividing the dataset into base and novel sets. The model is trained on the base set, comprising classes Y_b, and evaluated on both the novel set, encompassing classes Y_n, and the base validation set. Notably, there are no overlapping classes between the base and novel splits, ensuring $Y_b \cap Y_n = \emptyset$. This approach effectively reflects the method's performance on previously unseen classes.

Experiments

Datasets: We conduct our experiments on four widely used action recognition datasets:

Comparison of accuracy (%) of zero-shot action recognition methods with state-of-the-art performance on HMDB-51[40], UCF101[41], and Kinetics-600[42]. We applied two pretraining approaches: vanilla CLIP pretraining (WIT-400M), and further pretraining on Kinetics-400 (WIT-400M + K400). For the second pretraining, our approach outperforms the previous best zero-shot performance on HMDB-51 and UCF101, and achieves competitive results on Kinetics-600. We indicate the gain/decline relative to the previous best performance in blue.

Method	Publication	Input	Pretrain	Views	HMDB-51	UCF101	Kinetics-600	GFLOPS
				Methods with vision pretraining				
ASR[43]	ECML'17	16×112^2	Sports-1M [44]	1 × 1	21.8 ± 0.9	24.4 ± 1.0	-	38.5
ZSECOC[45]	CVPR'17	-	-	1 × 1	22.6 ± 1.2	15.1 ± 1.7	-	-
UR [46]	CVPR'18	16×224^2	ImageNet + ActivityNet	-	24.4 ± 1.6	17.51.6	-	-
TS-GCN [47]	AAAI'19	$16 \times -$	YFCC100M	1 × 1	23.2 ± 3.0	34.2 ± 3.1	-	-
E2E [48]	CVPR'20	16×112^2	Kinetics-700	1 × 1	32.7	48	-	-
ER-ZSAR[49]	ICCV'21	16×224^2	ImageNet21k	1 × 1	35.3 ± 4.6	51.8 ± 2.9	42.1 ± 1.4	65
SJE[50]	ICCV'15	$- \times 224^2$	-	-	-	-	22.3 ± 0.6	-
ESZSL[51]	ICML'15	-	-	-	-	-	22.9 ± 1.2	-
DEM[52]	CVPR'17	$- \times 224^2$	ImageNet 1K	-	-	-	23.6 ± 0.7	-
				Methods with vision-language multimodal pretraining				
Vanilla CLIP B/16 [5]	ICML'21	32×224^2	WIT-400M	1 × 1	40.8 ± 0.3	63.2 ± 0.2	59.8 ± 0.3	563
ActionCLIP B/16 [53]	arXiv'21	32×224^2	WIT-400M	10 × 3	40.8 ± 5.4	58.3 ± 3.4	66.7 ± 1.1	563 × 30
XCLIP B/16 [7]	ECCV'22	16×224^2	WIT-400M	4 × 3	44.6 ± 5.2	72.0 ± 2.3	65.2 ± 0.3	281 × 12
A5 [54]	ECCV'22	16×224^2	WIT-400M	5 × 1	44.3 ± 2.2	69.3 ± 4.2	55.8 ± 0.7	281 × 5
Vita-CLIP B/16 [55]	CVPR'23	32×224^2	WIT-400M + K400	1 × 1	48.6 ± 0.6	75.0 ± 0.6	67.4 ± 0.5	563
ViFi-CLIP B/16 [28]	CVPR'23	32×224^2	WIT-400M + K400	1 × 1	51.3 ± 0.6	76.8 ± 0.7	71.2 ± 1.0	563
ZAR	-	16×224^2	WIT-400M	1 × 1	43.6 ± 0.7	64.1 ± 0.9	60.2 ± 0.4	281
ZAR + K400 pretrain	-	16×224^2	WIT-400M + K400	1 × 1	54.2 ± 0.8	77.4 ± 0.8	70.5 ± 0.4	281
					+2.9	+0.6	-0.7	

Base-to-novel generalization results. HM stands for harmonic mean of the base and novel performance.

	Kinetics400			HMDB-51			UCF101		
Method	**Base**	**Novel**	**HM**	**Base**	**Novel**	**HM**	**Base**	**Novel**	**HM**
Vanilla CLIP B/16 [5]	53.3	46.8	49.8	53.3	46.8	49.8	78.5	63.6	70.3
ActionCLIP B/16 [53]	69.0	57.2	62.6	69.1	37.3	48.5	90.1	58.1	70.7
XCLIP B/16 [7]	74.1	56.4	64.0	69.4	45.5	55.0	89.9	58.9	71.2
A5 [54]	74.1	56.4	64.0	46.2	16.0	23.8	90.5	40.4	55.8
ViFi-CLIP B/16 [28]	76.4	61.1	67.9	73.8	53.3	61.9	92.9	67.7	78.3
ZAR	75.2	60.7	67.2	75.2	55.2	63.7	91.3	67.2	77.4

Kinetics-400 [56]: Kinetics-400 is a human action recognition dataset sourced from YouTube videos, comprising approximately 300,000 short clips. The dataset consists of 400 different action classes, with each clip lasting around 10 seconds. Each video is assigned a single label, indicating it belongs to only one class.

Kinetics-600 [42]: Kinetics-600 is an extension of the Kinetics-400 dataset, featuring 600 different action classes and a total of around 480,000 short clips.

HMDB-51 [40]: HMDB-51 is a collection of videos sourced from various resources including movies and web-based platforms. It includes over 6,000 videos spanning 51 action classes.

UCF101 [41]: UCF101 is an action recognition dataset containing 13,320 videos across 101 action categories. The dataset encompasses variations in camera view, object viewpoint, and backgrounds.

Experimental Setup. We use the VIT-B/16 backbone of the CLIP image encoder for our experiments. We update our prompt with four learnable tokens, and we use the default initialization setting 'a video of a'. We use 16 frames for our evaluation, with single view, the pictures are cropped into 224x224 size to fit into the backbone of VIT-B/16, and we compare our results to other methods that adapt CLIP or directly using CLIP for zero-shot action recognition on the datasets that we mentioned above. We optimize the prompt based on entropy minization and we update the paramter of the prompt learning for 1 step, we use AdamW with a learning rate of 0.005. We present our results of directly prompt-tuning CLIP following our method for zero-shot action recognition on UCF101, HMDB-51, and Kinetics-600 in Tab. 1. We observed from previous works that models built upon CLIP and subsequently trained with Kinetics-400 generally exhibit improved performance on downstream video-based action recognition tasks, with an accuracy improvement of over 8%. Therefore, we also report the results of further pretraining CLIP on Kinetics-400 and cross-evaluating on other action recognition datasets using the same prompt-tuning and zero-shot setting.

Confidence selection ablation study. We found that selecting the top 80% highest confidence augmentations to update the prompt learner yields the best results on UCF101 and HMDB-51.

	HMDB-51						HMDB-51			
UCF101										
10%	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**	**10%**
20%	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**		
53.94	53.88	54.07	53.94	53.81	54.01	54.14	54.17	54.14	54.07	77.10
76.84	77.26	77.34	77.34	77.37	77.37	77.42	77.37	77.29		

To illustrate the generalizability of our model to distribution shifts, we also include its performance in base-to-novel settings, following the setup and division outlined by Rasheed et al. [28]. In this setup, the data is divided into base and novel sets. The model is trained on the base set while being evaluated on both the base validation and novel sets, which effectively reflects the model's ability to generalize to unseen data points.

Zero-shot setting. In the zero-shot setting, we evaluate the performance of our ZAR method by applying prompt tuning to CLIP and assessing its effectiveness on unseen datasets without utilizing any additional labels from those test datasets. We assess the zero-shot performance of our ZAR method under two different pretraining scenarios. In the first setting, we directly prompt tune CLIP, indicating that the model is solely pretrained on the original CLIP model's training dataset WIT-400M [5]. In the second setting, we further pretraining CLIP on Kinetics-400 [56], imbues the model with knowledge from both WIT-400M and Kinetics-400. In both pretraining scenarios, we evaluate the performance of the model using our ZAR method in a zero-shot manner on three testing datasets: UCF101 [41], Kinetics-600[42] and HMDB-51[40]. Since Kinetics-600 is an extension of Kinetics-400 with substantial overlap in classes, we exclusively test on the 220 classes that did not appear in Kinetics-400 to ensure a fair evaluation of zero-shot performance. We report the accuracy based on three data splits randomly selected from the pool of 220 categories, with each split containing 160 categories, consistent with the splitting reported in [49]. The results of our ZAR-method's zero-shot performance, compared with other SOTA methods, are presented in Tab. 1. Notably, ZAR with Kinetics-400 pretraining surpasses the previous best performance on HMDB-51 by 3%, while performing comparably to previous results on UCF101 and Kinetics-600. It's worth highlighting that our method achieves these results with a significantly fewer number of trainable parameters to be considered, with only the prompt learner requiring updates.

Base-to-novel setting. Base-to-novel setting serves as an effective way to evaluate the model's performance on unseen data points, akin to assessing its performance in a zero-shot learning scenario. In this setting, we divide the dataset into two parts: the base and the novel sets. The model is trained on the base set and its performance is evaluated on both the base validation sets and the novel set, following the same split as reported in [28]. We trained the model on the base set for 16 shots and evaluated the results on the novel set as well as the base validation set. Notably, the classes present in the base validation set are included in the training data but are not utilized during the training process. The reported results are averaged across three random seeds to ensure robustness. The results of base-to-novel generalization are shown in Tab. 2.

Pretraining CLIP. As outlined in the zero-shot setting section, we conducted evaluations of our model's zero-shot performance on both vanilla CLIP and CLIP further pretrained with Kinetics-400. The pretraining on Kinetics-400 dataset involves training for 10 epochs, utilizing a batch size of 256, and employing a learning rate of 8e-6. During the pretraining phase, both the image encoder and the text encoder parameters are updated, while the prompt learner is not considered. Although this update is computationally expensive, for the sake of fair comparison, we also present the results of this pretraining. Subsequently, during zero-shot evaluation on the UCF101, Kinetics-600, and HMDB-51 datasets, only the parameters of the light-weighted prompt learner are updated, while the image encoder and the text encoder remain frozen. To facilitate the exchange of temporal information between frames and obtain a comprehensive representation of the entire video, we employed embedding level fusion. In this fusion technique, the image inputs undergo separate processing by the CLIP image encoder, and the resulting frame-level embeddings are average-pooled to be fused together, forming the video-level embedding representation. Alternative fusion methods include image-level fusion, where losses are computed for each image independently, and decision-level fusion, where the final logits are averaged after calculating the probability or decision of each individual image frame with the text prompt. In image-level fusion, each image contributes to parameter updates and data flow independently, with limited knowledge transfer between images. In decision-level fusion, each image contributes to the decision-making process, and their votes are averaged after calculating the logits for each individual image frame with the text prompt. In our adoption of the embedding-level fusion method, knowledge exchange occurs at the embedding level, fostering mutual contribution to knowledge exchange. In this research, we adopt embedding-level fusion due to its superior performance, as empirically demonstrated in [28]. Further exploration of each fusion method is beyond the scope of this study. The objective function during pretraining simply involves cross entropy, as depicted in Eq. 2. In this equation, the similarity between each video-level embedding and its corresponding text embedding is maximized.

Parameter selections. Through empirical investigation, we determined the optimal values for N (number of crops), Q (number of different sampling rates), and M (number of video samples per sampling rate) to be 3, 4, and 4, respectively. For the prompt updating stage, involving both directly adapting CLIP and pretraining CLIP on Kinetics-400 and further evaluating on other datasets in a zero-shot manner, we crop 16 frames from the video. This results in a total of $3 \times 4 \times 4 = 48$ augmentations and a combined total of $48 \times 16 = 768$ image frames. The input tensor to the model for prompt updating has a shape of $768 \times 3 \times 224^2$, where 3×224^2 represents the shape of a single image. In the base-to-novel setup for prompt updating, we crop 8 frames from the video, resulting in a tensor with a shape of $384 \times 3 \times 224^2$. During inference, all operations are performed on a single $16 \times 3 \times 224^2$ video clip, which is center cropped from the original video.

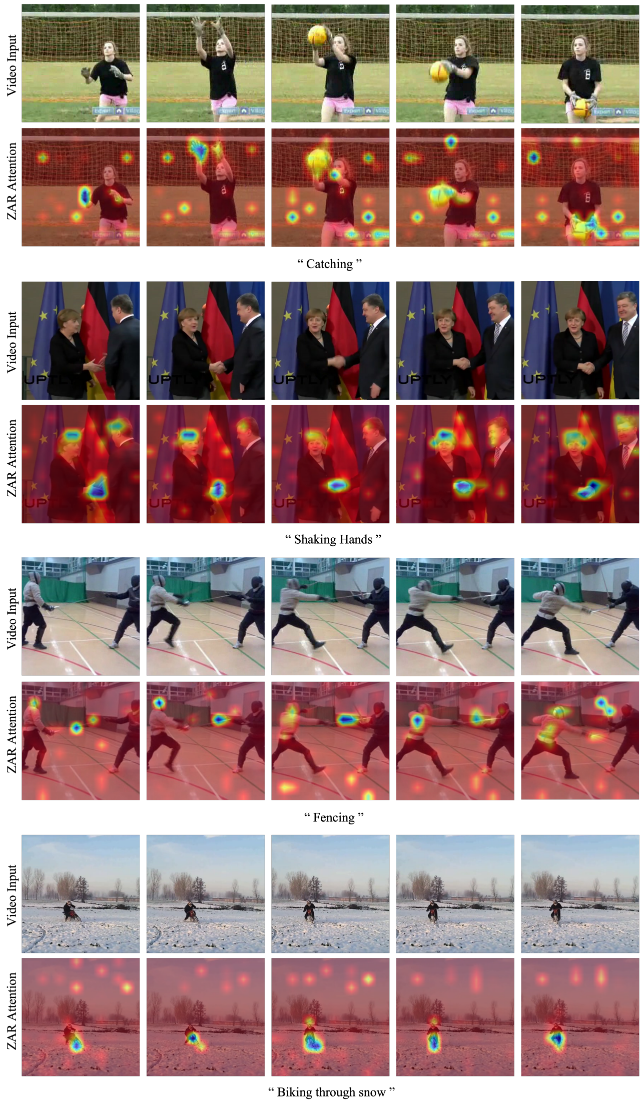
Methodology to implement different frame rates. To implement different frame rate sampling, the video is initially processed into images. Subsequently, a random starting point of the video is selected, with the sampling rate simulated by uniformly sampling from a truncated range of video frames. For example, if a 10-second video consists of a total of 300 frames with a physical frame rate of 30 fps, half the length of the video is chosen as the starting point, with the starting frame randomly selected at 40 and the corresponding ending frame at 190. Uniform sampling is then applied, which effectively doubles the sampling rate compared to working on the full-length video and performing uniform sampling. For example, in a scenario where a 10-second video comprises 300 frames with a physical frame rate of 30 fps, working on a truncated video—where the starting frame is randomly chosen at 40 and the corresponding ending frame at 190, or equivalently, a half-length video—results in a sampling rate that is effectively doubled compared to working on the full-length video and performing uniform sampling.

Results visualizations. To provide a more intuitive understanding of our results, we present visualizations of attention maps generated by our approach, depicted Fig. 2. These attention maps reveal the model's ability to selectively focus on crucial features even amidst motion, capturing abstractly related aspects of the depicted actions. We also visualze the t-sne of the dataset as presented in Fig. 6. The t-sne visualization technique essentially transforms high-dimensional output into a lower dimension to facilitate the visualization of cluster separations. In our case, the output of our backbone consists of 512-dimensional features, which are then converted into 2D features for t-sne visualization. We observe that Kinetics-600 and UCF101 exhibit clearer cluster separations in the t-sne visualization, indicating better discriminative features among classes. Conversely, HMDB-51 displays overlapping classes at the center, suggesting a more challenging scenario for classification. Despite this inherent difficulty, our method achieves improved performance on HMDB-51 compared to previous approaches, showcasing its effectiveness in handling the intricacies of this dataset.

Ablation Study

We conduct an ablation study on each parameter of our model, empirically validating the efficacy of our approach and identifying the optimal configuration.

Percentage of selected video augmentations utilized for entropy minimization. We recognize the significance of the percentage of video augmentations utilized for entropy minimization, which contributes to the gradient backpropagation of the prompt learner. Prior methodologies such as MEMO [52] employ all augmentations of a single image to update the entire model during test time, whereas TPT proposes to leverage the top 10% augmentations with the highest confidence for entropy minimization, and updating only the prompt learner. Given the divergence between their training on images and our focus on video-based tasks, as well as our distinct augmentation methodologies using different frame rates compared to traditional rotate and flip augmentation approaches, we conduct empirical investigations to determine the optimal percentage of video augmentations necessary for updating the prompt learner. We conduct our experiments on UCF101 and HMDB-51 in a zero-shot manner, selecting augmentations with the highest confidence and varying the selection percentage from 10% to 100% in increments of 10%. Our findings reveal that selecting 80% highest confidence augmentations yields the best results on both datasets.The accuracy corresponding to each confidence selection percentage is presented in Tab. 3 Therefore, the top 80% video augmentations with the highest confidence are selected for the updating of prompt learner parameters, unless otherwise stated.



" Catching "

" Shaking Hands "

" Fencing "

" Biking through snow "

Figure 2. *Attention map visualization of ZAR focus.*

Number of different sampling rates, Q. To find out the influence of using different frame rates, we also conduct an ablation study regarding the number of different sampling rates. We use a total of four different sampling rates, achieved through uniform sampling of segments from the video, with the slowest sampling

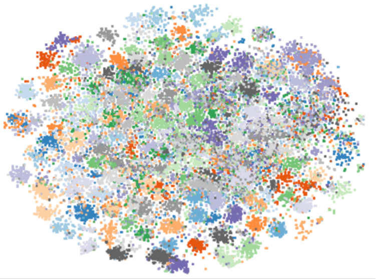


Figure 3. *

Kinetics-600
Accuracy 70.5

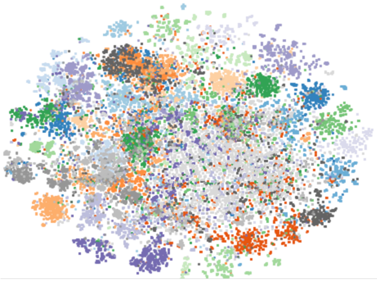


Figure 4. *

HMDB-51
Accuracy 54.2

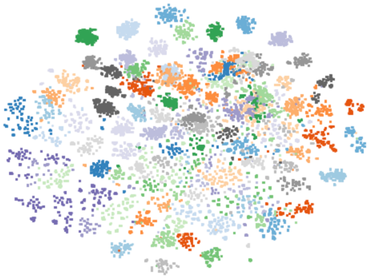


Figure 5. *

UCF101
Accuracy 77.4

Figure 6. *t-sne visualization of zero-shot performance on Kinetics-600, HMDB-51 and UCF101.*

rate corresponding to the entire video length. We begin our ablation study by removing the fastest sampling rate (obtained by uniform sampling of 1/8 of the entire video) and successively ablate the video samples derived from using 1/4 and 1/2 of the entire video. The sampling rate ablation study result is shown in Tab. 4. As we systematically remove augmentations with fast sampling rates and transition to slower ones, we observe a corresponding decline in accuracy, eventually converging to accuracy equivalent to models employing fine-tuning without prompt tuning. This observation underscores the importance of using multiple sampling rates, as it enables the model to capture diverse temporal dynamics and glean essential motion cues from the video data.

Number of crops, N We investigate the impact of different numbers of crops to augment the video. Specifically, we evaluate the performance when using a single crop, 3 crops, 5 crops, and 9 crops. Here, ”1 crop” refers to the center crop, ”3 crops” encompass left, middle, and right crops, ”5 crops” include top and bottom crops in addition to the previous three, and ”9 crops” further expand to include top-left, top-right, bottom-left, and bottom-right crops. Our ablation study results, as presented in Table 5, indicate that employing 3 crops yields the highest accuracy. However, accuracy drops significantly when using 5 or 9 crops. We attribute this decline to the increased diversity of viewpoints introduced by the additional crops, leading to difficulties in effectively minimizing entropy and capturing mutual features across the different views.

Sampling rates ablation study.

HMDB-51			
Use all four sampling rates	**Drop 1/8**	**Drop 1/8 and 1/4**	**Drop 1/8, 1/4, and 1/2**
54.2	53.6	52.1	51.5

Number of crops ablation study.

HMDB-51			
1 crop	**3 crops**	**5 crops**	**9 crops**
53.6	54.2	52.0	51.8

Conclusion

We propose a zero-shot prompt learning framework for video based action recognition, using different sampling rates augmentations to update the prompt learner and enhance accuracy across various action recognition datasets. Employing the CLIP backbone, our method operates in a true zero-shot fashion, devoid of additional training data, and relies solely on individual test videos. We empirically demonstrate the efficacy of our approach via comprehensive experimentation, affirming its viability and effectiveness.

Limitations. While our approach consistently achieves superior zero-shot results on HMDB-51, its improvement over previous best on other datasets remains marginal. We posit that this limitation may stem from the nature of entropy minimization, which relies on the fidelity of the label-to-data point linkage. In scenarios where labels lack details, pertinent information may be lost during video encoding, leading to suboptimal entropy reduction and prompt undates.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[3] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, “Ernie: Enhanced language representation with informative entities,” *arXiv preprint arXiv:1905.07129*, 2019.

[4] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning

transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[6] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International conference on machine learning*. PMLR, 2021, pp. 4904–4916.

[7] B. Ni, H. Peng, M. Chen, S. Zhang, G. Meng, J. Fu, S. Xiang, and H. Ling, "Expanding language-image pretrained models for general video recognition," in *European Conference on Computer Vision*. Springer, 2022, pp. 1–18.

[8] R. Liu, J. Huang, G. Li, J. Feng, X. Wu, and T. H. Li, "Revisiting temporal modeling for clip-based image-to-video knowledge transferring," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6555–6564.

[9] M. Maaz, H. Rasheed, S. Khan, F. S. Khan, R. M. Anwer, and M.-H. Yang, "Class-agnostic object detection with multi-modal transformer," in *European Conference on Computer Vision*. Springer, 2022, pp. 512–531.

[10] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019.

[11] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion, "Mdetr-modulated detection for end-to-end multi-modal understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1780–1790.

[12] R. Abdelfattah, Q. Guo, X. Li, X. Wang, and S. Wang, "Cdul: Clip-driven unsupervised learning for multi-label image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1348–1357.

[13] R. Zhang, W. Zhang, R. Fang, P. Gao, K. Li, J. Dai, Y. Qiao, and H. Li, "Tip-adapter: Training-free adaption of clip for few-shot classification," in *European Conference on Computer Vision*. Springer, 2022, pp. 493–510.

[14] T. Huang, B. Dong, Y. Yang, X. Huang, R. W. Lau, W. Ouyang, and W. Zuo, "Clip2point: Transfer clip to point cloud classification with image-depth pre-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 157–22 167.

[15] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," *arXiv preprint arXiv:2104.13921*, 2021.

[16] H. Shi, M. Hayat, Y. Wu, and J. Cai, "Proposalclip: Unsupervised open-category object proposal generation via exploiting clip cues," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9611–9620.

[17] Y. Zhong, J. Yang, P. Zhang, C. Li, N. Codella, L. H. Li, L. Zhou, X. Dai, L. Yuan, Y. Li *et al.*, "Regionclip: Region-based language-image pretraining," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 793–16 803.

[18] H. Fang, P. Xiong, L. Xu, and Y. Chen, "Clip2video: Mastering video-text retrieval via image clip," *arXiv preprint arXiv:2106.11097*, 2021.

[19] Y. Ma, G. Xu, X. Sun, M. Yan, J. Zhang, and R. Ji, "X-clip: End-to-end multi-grained contrastive learning for video-text retrieval," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 638–647.

[20] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li, "Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning," *Neurocomputing*, vol. 508, pp. 293–304, 2022.

[21] Z. Lin, S. Geng, R. Zhang, P. Gao, G. de Melo, X. Wang, J. Dai, Y. Qiao, and H. Li, "Frozen clip models are efficient video learners," in *European Conference on Computer Vision*. Springer, 2022, pp. 388–404.

[22] W. Wu, X. Wang, H. Luo, J. Wang, Y. Yang, and W. Ouyang, "Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6620–6630.

[23] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[24] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *European Conference on Computer Vision*. Springer, 2022, pp. 709–727.

[25] X. Liu, K. Ji, Y. Fu, W. Tam, Z. Du, Z. Yang, and J. Tang, "P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2022, pp. 61–68.

[26] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[27] ——, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 816–16 825.

[28] H. Rasheed, M. U. khattak, M. Maaz, S. Khan, and F. S. Khan, "Finetuned clip models are efficient video learners," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[29] X. Cheng, Z. Fu, and J. Yang, "Zero-shot image super-resolution with depth guided internal degradation learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*. Springer, 2020, pp. 265–280.

[30] J. W. Soh, S. Cho, and N. I. Cho, "Meta-transfer learning for zero-shot super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3516–3525.

[31] D. Bau, H. Strobelt, W. Peebles, J. Wulff, B. Zhou, J.-Y. Zhu, and A. Torralba, "Semantic photo manipulation with a generative image prior," *arXiv preprint arXiv:2005.07727*, 2020.

[32] Y. Sun, X. Wang, Z. Liu, J. Miller, A. Efros, and M. Hardt, "Test-time training with self-supervision for generalization under distribution shifts," in *International conference on machine learning*. PMLR, 2020, pp. 9229–9248.

[33] Y. Gandelsman, Y. Sun, X. Chen, and A. Efros, "Test-time training with masked autoencoders," *Advances in Neural Information Processing Systems*, vol. 35, pp. 29 374–29 385, 2022.

[34] G. S. Dhillon, P. Chaudhari, A. Ravichandran, and S. Soatto, "A baseline for few-shot image classification," *arXiv preprint arXiv:1909.02729*, 2019.

[35] S. Roy, A. Siarohin, E. Sangineto, S. R. Bulo, N. Sebe, and E. Ricci, "Unsupervised domain adaptation using feature-whitening and consensus loss," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9471–9480.

[36] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," *arXiv preprint*

arXiv:2006.10726, 2020.

[37] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," *Advances in neural information processing systems*, vol. 17, 2004.

[38] M. Zhang, S. Levine, and C. Finn, "Memo: Test time robustness via adaptation and augmentation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 38 629–38 642, 2022.

[39] M. Shu, W. Nie, D.-A. Huang, Z. Yu, T. Goldstein, A. Anandkumar, and C. Xiao, "Test-time prompt tuning for zero-shot generalization in vision-language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 14 274–14 289, 2022.

[40] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *2011 International conference on computer vision*. IEEE, 2011, pp. 2556–2563.

[41] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[42] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, "A short note about kinetics-600," *arXiv preprint arXiv:1808.01340*, 2018.

[43] Q. Wang and K. Chen, "Alternative semantic representations for zero-shot human action recognition," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part I 10*. Springer, 2017, pp. 87–102.

[44] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[45] J. Qin, L. Liu, L. Shao, F. Shen, B. Ni, J. Chen, and Y. Wang, "Zero-shot action recognition with error-correcting output codes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2833–2842.

[46] Y. Zhu, Y. Long, Y. Guan, S. Newsam, and L. Shao, "Towards universal representation for unseen action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9436–9445.

[47] J. Gao, T. Zhang, and C. Xu, "I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 8303–8311.

[48] B. Brattoli, J. Tighe, F. Zhdanov, P. Perona, and K. Chalupka, "Rethinking zero-shot video classification: End-to-end training for realistic applications," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4613–4623.

[49] S. Chen and D. Huang, "Elaborative rehearsal for zero-shot action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 638–13 647.

[50] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele, "Evaluation of output embeddings for fine-grained image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2927–2936.

[51] B. Romera-Paredes and P. Torr, "An embarrassingly simple approach to zero-shot learning," in *International conference on machine learning*. PMLR, 2015, pp. 2152–2161.

[52] L. Zhang, T. Xiang, and S. Gong, "Learning a deep embedding model for zero-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2021–2030.

[53] M. Wang, J. Xing, and Y. Liu, "Actionclip: A new paradigm for video action recognition," *arXiv preprint arXiv:2109.08472*, 2021.

[54] C. Ju, T. Han, K. Zheng, Y. Zhang, and W. Xie, "Prompting visual-language models for efficient video understanding," in *European Conference on Computer Vision*. Springer, 2022, pp. 105–124.

[55] S. T. Wasim, M. Naseer, S. Khan, F. S. Khan, and M. Shah, "Vita-clip: Video and text adaptive clip via multimodal prompting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 034–23 044.

[56] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

Author Biography

Qiyue Liang received her BS (2018) and Ph.D. (2024) in Electrical and Computer Engineering from Purdue University. Her current work focuses on the optimization of large language models, and her research interests include computer vision, language models, and deep learning.

JOIN US AT THE NEXT EI!

electronic IMAGING

Imaging across applications . . . Where industry and academia meet!

• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org