

Learning to Generalize without Bias for Open-Vocabulary Action Recognition

Yating Yu^{1*} Congqi Cao^{1*†} Yifan Zhang² Yanning Zhang¹

¹Northwestern Polytechnical University ²Institute of Automation, Chinese Academy of Sciences

yatingyu@mail.nwpu.edu.cn, congqi.cao@nwpu.edu.cn,
 yfzhang@nlpr.ia.ac.cn, yanningzhang@nwpu.edu.cn

Abstract

Leveraging the effective visual-text alignment and static generalizability from CLIP, recent video learners adopt CLIP initialization with further regularization or recombination for generalization in open-vocabulary action recognition in-context. However, due to the static bias of CLIP, such video learners tend to overfit on shortcut static features, thereby compromising their generalizability, especially to novel out-of-context actions. To address this issue, we introduce **Open-MeDe**, a novel *Meta*-optimization framework with static *De*biasing for *Open*-vocabulary action recognition. From a fresh perspective of generalization, Open-MeDe adopts a meta-learning approach to improve “*known-to-open generalizing*” and “*image-to-video debiasing*” in a cost-effective manner. Specifically, Open-MeDe introduces a cross-batch meta-optimization scheme that explicitly encourages video learners to quickly generalize to arbitrary subsequent data via virtual evaluation, steering a smoother optimization landscape. In effect, the free of CLIP regularization during optimization implicitly mitigates the inherent static bias of the video meta-learner. We further apply self-ensemble over the optimization trajectory to obtain generic optimal parameters that can achieve robust generalization to both in-context and out-of-context novel data. Extensive evaluations show that Open-MeDe not only surpasses state-of-the-art regularization methods tailored for in-context open-vocabulary action recognition but also substantially excels in out-of-context scenarios. Code is released at <https://github.com/Mia-YatingYu/Open-MeDe>.

1. Introduction

Open-vocabulary action recognition (OVAR) aims to identify test videos whose classes are not previously encountered during the training phase, which challenges the generalization and zero-shot capabilities of the video learn-

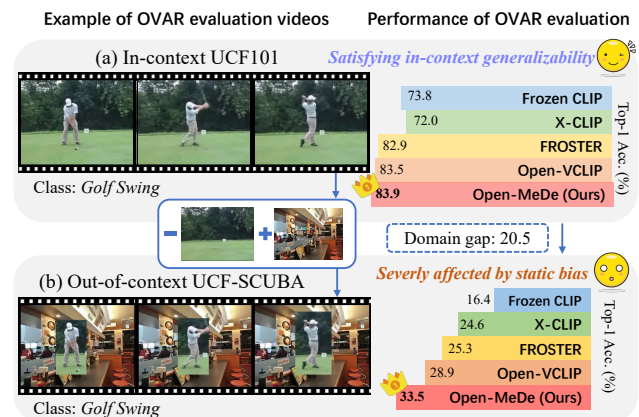


Figure 1. Performance comparison (Top-1 Acc (%)) under various open-vocabulary evaluation settings where the video learners except for CLIP are tuned on Kinetics-400 [26] with frozen text encoders. The satisfying in-context generalizability on UCF101 [43] (a) can be severely affected by static bias when evaluating on out-of-context UCF-SCUBA [30] (b) by replacing the video background with other images.

ers [3, 48, 52, 57]. Recently, the emergence of image-based visual-language (I-VL) pre-training, such as CLIP [39] and ALIGN [23], has shown promising zero-shot inference in image-based tasks. Inspired by this success, recent attempts [3, 4, 8, 34, 36, 49] have been made to adapt CLIP for general action recognition via additional temporal modeling following the “*pre-train, prompt and fine-tune*” paradigm [47]. Broadly, these video learners optimize the learnable parameters from the start point of CLIP, pursuing decent performance on the training videos, known as standard fine-tuning objectives. However, adapting CLIP to the video domain, especially for OVAR, is extremely challenging, as the video learners with standard fine-tuning objectives often lead to overfitting, which achieves improved specialization at the cost of generalization degradation.

To build an improved zero-shot video learner, Open-VCLIP [53] and FROSTER [21] propose to regularize the fine-tuning process curbing deviation from CLIP’s gener-

*Equal Contribution

†Corresponding Author

alization from the perspective of model patching [22] and knowledge distillation [7, 13, 38], respectively. In Fig. 1, these methods have achieved satisfying performance compared to frozen CLIP and X-CLIP [34] on UCF101 [43] dataset under in-context open-vocabulary evaluation, where the action categories have strong correlations with the context in videos. However, when it comes to the out-of-context evaluation in SCUBA [30], where the video background is replaced by other images, the performance degrades severely. As these video learners are intimately tied to the learning of shortcut static features, which manifest as static bias, they interfere with the learning of motion cues, resulting in poor out-of-context generalization [16]. Based on these observations, we argue that the static generalization of CLIP can (1) effectively adapt to in-context scenarios for OVAR by regularizing video learners; yet (2) it undesirably hinders the sensitivity of such video learners to motion cues, exerting a notable detrimental impact on generalization under out-of-context, open-vocabulary setting.

How can we encourage the emergence of such robust open-vocabulary generalization for both in-context and out-of-context scenarios? We explore an explicit approach to this problem: as the video learner is trained with a sampled batch of videos at each gradient step, our objective is to optimize the learner from a meta-learning standpoint so that it can quickly adapt to arbitrary subsequent data, thereby minimizing inherent biases toward known data and static cues.

Based on this insight, we propose **Open-MeDe**, the first Meta-learning based framework with static Debiasing for in-context and out-of-context Open-vocabulary action recognition. Meta-learning, also known as “*learning to learn*”, incorporates virtual evaluation during the training process for better generalization [1, 17, 35]. In our meta-learning scheme, the “*learning to generalize*” process is enhanced by naturally treating sequences of adjacent batches sampled from the training set as a distribution of tasks. More concretely, our procedure optimizes the video learner to obtain fast weights by gradient descent updates on the current batch (*i.e.*, *meta training*), while evaluating the subsequent batch (*i.e.*, *meta testing*) based on fast weights of the learner, which mimics a known-to-open task. Based on the evaluation performance in *meta testing*, our procedure can further optimize the learner to obtain more generalizable video-specific knowledge against inherent known and static biases. In effect, this cross-batch meta-optimization formulates a meta-learner free of CLIP regularization, thereby facilitating smoother optimization and robust video representation learning for fast known-to-open generalizing, thus enhancing image-to-video debiasing. Tailored to the optimization trajectory of the video learner, we further employ self-ensemble stabilization, *i.e.*, Gaussian Weight Average (GWA), to derive generic optima for robust generalization at open-vocabulary test time. Overall, while inte-

grating the same video learner, our model-agnostic Open-MeDe outperforms existing regularization-based methods, which strikes a promising balance on in-context and out-of-context generalization settings (Fig. 1).

The contribution of our work can be summarized as:

- We introduce a novel meta-learning based framework, Open-MeDe, which provides new insights for more generalized open-vocabulary action recognition.
- We propose cross-batch meta-optimization and self-ensemble stabilization, which effectively power known-to-open generalizing and image-to-video debiasing of the video learner for robust generalizability.
- We conduct extensive evaluations on various scenarios including base-to-novel, cross-dataset, and out-of-context open-vocabulary action recognition. Experimental results show that Open-MeDe consistently improves performance across all the benchmarks.

2. Related Work

2.1. Adapting CLIP to Action Recognition

A seminal work of I-VL, CLIP [39] has demonstrated remarkable static generalization, achieving promising performance in image-based zero-shot inference. Despite extensive works [40, 47, 52] fully fine-tuning the video learner, a collection of studies focuses on adopting lightweight adapters [5, 36, 55] or incorporating learnable prompts [24, 49] for easy video adaptation. However, these video learners adhere to the standard fine-tuning paradigm, which tends to overfit in the closed-set setting, thereby limiting expertise in open-vocabulary settings. To this end, Open-VCLIP [50] regularizes the fine-tuning process of the video learner, preventing deviation from CLIP’s generalization, by interpolating frozen CLIP weights with the current learner on the fly. FROSTER [21] and STDD [56] enforce the regularization from the perspective of knowledge distillation [7, 10, 14, 41], aligning features of the video learner and frozen CLIP via a tailored residual module. Despite demonstrating superiority in open-vocabulary evaluations, the increased computational overhead and excessive reliance on static cues introduced by CLIP regularization hinder efficient adaptation and robust generalization. In contrast, we approach the problem of adapting CLIP-based video learners to OVAR from a fresh view of “learning to generalize without bias”. During training, the learner is explicitly forced to quickly generalize to forthcoming data by sorely resorting to the knowledge learned by itself rather than by the virtue of CLIP’s static generalization.

2.2. Meta-learning

Rather than directly learning from experiences, with the goal of learning to learn, meta-learning can quickly generalize to new tasks by leveraging prior learning abilities [19].

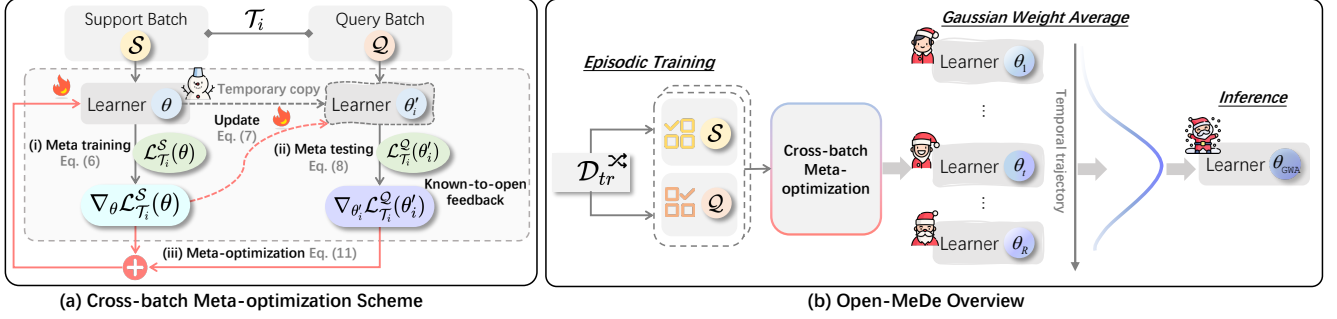


Figure 2. Illustration of our framework. (a) The cross-batch meta-optimization scheme aims to mimic the known-to-open generalization task \mathcal{T}_i by performing the gradient descent update (*i.e.*, *meta training*) on the support batch \mathcal{S} and virtual evaluation (*i.e.*, *meta testing*) on the query batch \mathcal{Q} . Then, the video learner is optimized by both class-specific losses from \mathcal{S} and task feedback from \mathcal{Q} for more generalizable knowledge against inherent known and static biases. (b) Overview of the Open-MeDe framework with self-ensemble stabilization. During the episodic training process, we exploit the optimization trajectory of the video learner to perform Gaussian Weight Average (GWA) to derive generic optima for robust generalization.

As the representative works in meta-learning, MAML [17] boasts simplicity and has actively driven the development of the gradient-based methods in few-shot learning. Recently, meta-learning techniques have also been explored in zero-shot learning [20, 32, 37, 45] and domain adaptation [29], which typically perform episode-wise training by dividing the training set into support and query sets with different classes distributions. Targeting long-tailed issues within closed-set video scene generation, MVSGG [54] employs meta-learning across several manually predefined task types, which are partitioned based on specific conditional biases in the training data. However, these approaches are often prone to meta-overfitting due to insufficient meta tasks and limited application scopes of generalization. Differently, our work tackles ubiquitous challenges in video understanding beyond closed-set and in-context settings, *i.e.*, mitigating static bias of video learners for open-vocabulary generalization. To the best of our knowledge, we are the first to directly integrate the mini-batch training mechanism with meta-learning to naturally mimic diverse known-to-open tasks utilizing cross-batch data without additional computational overhead.

3. Method

3.1. Preliminaries

Action recognition with CLIP-based video learner. Consider a CLIP-based video learner with a ViT architecture [15], that incorporates temporal modeling for video understanding [47, 49, 50, 52, 55, 57]. Next, we present the standard vision-only fine-tuning paradigm that applies such a video learner f_{θ_v} with a frozen text encoder f_{θ_t} to action recognition. Specifically, given a video clip V_i , and a candidate action label $T_j \in \mathcal{Z}_{tr}$ described in predefined textual templates (*e.g.*, “a video of {action}”) from the training set

\mathcal{D}_{tr} , the similarity is calculated as:

$$s_{i,j} = \frac{\langle v_i, t_j \rangle}{\|v_i\| \|t_j\|}, v_i = f_{\theta_v}(V_i), t_j = f_{\theta_t}(T_j), \quad (1)$$

where the training objective is to maximize it of the matched V_i and T_j , or to minimize it otherwise. The loss function is implemented by the cross-entropy loss in [9, 39, 52] as:

$$\mathcal{L}_{CE} = -\frac{1}{B} \sum_i \sum_k y_{i,k} \log \left(\frac{\exp(s_{i,k})}{\sum_j \exp(s_{i,j})} \right), \quad (2)$$

where B and K denote the minibatch size and the number of all known classes, respectively. If the i -th video belongs to the k -th class, $y_{i,k}$ equals 1; otherwise, $y_{i,k}$ equals 0. In OVAR, the trained video learner should achieve good generalization on test data with the class label $T_i \in \mathcal{Z}_{te}$, where $\mathcal{Z}_{te} \cap \mathcal{Z}_{tr} = \emptyset$.

Model-agnostic meta-learning (MAML). MAML [17] is a gradient-based meta-optimization framework designed for few-shot learning, which aims to learn good initialization such that a few gradient steps will lead to fast learning on new tasks. Formally, consider a model f_θ with parameters θ , MAML learns a set of initial weight values, which will serve as a good starting point for fast adaptation to a new task \mathcal{T}_i , sampled from a task distribution $p(\mathcal{T})$. When adapting to the task \mathcal{T}_i , the fast weights θ'_i are computed *w.r.t.* examples from \mathcal{T}_i though single inner-loop update as:

$$\theta'_i = \theta - \alpha \nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}(f_\theta), \quad (3)$$

where α denotes the step size for inner loops. Then, the model with fast weights $f_{\theta'_i}$ is evaluated on new samples from the same task \mathcal{T}_i , to act as the feedback (*i.e.*, loss gradients) to adapt to current task \mathcal{T}_i to optimize the initialization θ for generalization as:

$$\theta \leftarrow \theta - \beta \nabla_{\theta} \sum_{\mathcal{T}_i} \mathcal{L}_{\mathcal{T}_i}(f_{\theta'_i}). \quad (4)$$

where β is the step size for outer loops. Computationally, due to the additional backward propagation burden of the gradient by gradient update, MAML presents a first-order approximation, FOMAML, by dropping the backward pass.

3.2. Open-MeDe

As discussed above, the standard fine-tuning paradigm can cause the video learner to overfit to the known classes during training, leading to poor zero-shot capabilities. Also, CLIP regularization-based approaches face challenges in achieving robust generalization due to the excessive reliance on superficial static cues in videos. To tackle these issues, we draw upon the philosophy and methodology from meta-learning, and propose Open-MeDe framework, which is illustrated in Fig. 2, to enhance both know-to-open generalizing and image-to-video debiasing simultaneously.

3.2.1. Cross-batch meta-optimization

Our Open-MeDe framework primarily adopts a cross-batch meta-optimization scheme (in Fig. 2(a)) to enhance the video learner via *meta training and testing*, enabling it to acquire generalizable, video-specific knowledge instead of overly exploiting static biases. Note that we neither sample from a distribution of N -way K -shot tasks as done in few-shot MAML nor deliberately split the training set into support and query sets as Meta-ZSL [32, 45] suggested. Instead, our support and query examples are constructed effortlessly and arbitrarily by the default training data sampler. In effect, we consider this arbitrariness a blessing for building the natural “*known-to-open generalization task*”, since the known biases in *meta training* data do not hold in *meta testing* data due to different inherent label distributions across batches. A known-to-open task can be created by extending the original gradient step into two consecutive mini-batches in one pass, with the current batch acting as support data and the subsequent batch as query data. Specifically, in line with the episode-wise training akin to MAML, we first train the learner within an inner loop (*i.e.*, *meta training*), where the fast weights are obtained through a single gradient step for each support batch. Following this adaptation, in the outer loop, query videos are sampled to evaluate the generalization performance of the adapted learner with fast weights (*i.e.*, *meta testing*). In this work, our framework further updates the fast weights of the learner based on the evaluation performance during *meta testing*, which then provides feedback for the task to derive more generalizable optimization for the learner.

Meta training. At each training iteration, we first utilize each support batch $\mathcal{S} = \{V_i, T_i\}^B$ from the task \mathcal{T}_i to train the video learner f_θ (with parameters θ), via one standard gradient step. The inner loop update is governed by the loss on the support batch as:

$$\mathcal{L}_{\mathcal{T}_i}^{\mathcal{S}}(\theta) = \mathcal{L}(f_\theta(\mathcal{S})), \quad (5)$$

where $\mathcal{L}(\cdot)$ refers to the loss function (*e.g.*, the cross-entropy loss \mathcal{L}_{CE} *w.r.t.* Eq. (2)). Then, we make a temporary copy for the original parameters θ and update the intermediate parameters for fast weights as follows:

$$\theta'_i = \theta - \alpha \nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}^{\mathcal{S}}(\theta), \quad (6)$$

where α denotes the learning rate for *meta training*. Intuitively, this step simulates a direct update to train the learner to obtain class-specific knowledge of the support data.

Meta testing. After meta training on the support batch, we then scheme a virtual testing process, leveraging the query batch $\mathcal{Q} = \{V_i, T_i\}^B$, where $\mathcal{S} \cap \mathcal{Q} = \emptyset$, to evaluate the generalization performance of the base learner $f_{\theta'_i}$. Formally, we measure the known-to-open performance on \mathcal{T}_i by calculating the class-specific loss *w.r.t.* the query data as:

$$\mathcal{L}_{\mathcal{T}_i}^{\mathcal{Q}}(\theta'_i) = \mathcal{L}(f_{\theta'_i}(\mathcal{Q})). \quad (7)$$

Here, the formulation closely relates to the standard fine-tuning process, which aims to obtain decent class-specific performance for all training batches. Differently, this step merely evaluates the intermediary base learner for its known-to-open generalizability on each task, due to the original parameters θ remaining immune to the task-specific updates. Hence, it can be used to provide feedback for the learner on *what video-specific knowledge should be learned to derive the robust generalization* against inherent known and static biases in the following meta-optimization.

Meta-optimization. As mentioned above, the intuition behind our approach is that the virtual evaluation during meta testing can provide useful feedback to encourage the learning of more robust representations for fast known-to-open generalization after *meta training* on the support data (*i.e.*, $\theta'_i \leftarrow \theta$). Note that original MAML approaches focus on optimizing parameters for a strong initialization, enabling quick adaptation to new tasks with minimal gradient updates. Conversely, open-vocabulary recognition requires zero-shot capabilities, where no further adaptation can be applied for new tasks. Therefore, class-specific knowledge should be strengthened in terms of global optimization. To this end, within the outer loop, the parameters of the learner are optimized to minimize the class-specific errors for the support data and the adaptation cost for the query data simultaneously. The combination of both Eq. (5) and Eq. (7) is used to carry out the outer loop update, thus the objective for meta-optimization can be defined as:

$$\begin{aligned} \min_{\theta} \mathcal{L}_{\mathcal{T}_i}(\theta) &= \min_{\theta} (\mathcal{L}_{\mathcal{T}_i}^{\mathcal{S}}(\theta) + \mathcal{L}_{\mathcal{T}_i}^{\mathcal{Q}}(\theta'_i)) \\ &= \min_{\theta} (\mathcal{L}_{\mathcal{T}_i}^{\mathcal{S}}(\theta) + \mathcal{L}_{\mathcal{T}_i}^{\mathcal{Q}}(\theta - \alpha \nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}^{\mathcal{S}}(\theta))). \end{aligned} \quad (8)$$

Here, the first term refers to the class-specific knowledge learned on the support batch, while the second term provides the known-to-open generalization feedback based on

θ'_i towards robust representation learning *w.r.t.* the task \mathcal{T}_i . The optimizing process of the parameter θ can be given by:

$$\theta \leftarrow \theta - \beta \nabla_{\theta} \sum_{i=1}^N (\mathcal{L}_{\mathcal{T}_i}^{\mathcal{S}}(\theta) + \mathcal{L}_{\mathcal{T}_i}^{\mathcal{Q}}(\theta - \alpha \nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}^{\mathcal{S}}(\theta))), \quad (9)$$

where N is the batch size of the task for meta-optimization. Since the MAML meta-gradient update needs to differentiate through the optimization process (*i.e.*, a gradient by a gradient), it's not an ideal solution where we need to optimize a large number of tasks during the training phase. Therefore, we opt for the one-step update approximation by dropping the backward pass of $\theta \leftarrow \theta'_i$ as:

$$\theta \leftarrow \theta - \beta \sum_{i=1}^N (\nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}^{\mathcal{S}}(\theta) + \delta \nabla_{\theta'_i} \mathcal{L}_{\mathcal{T}_i}^{\mathcal{Q}}(\theta'_i)), \quad (10)$$

where β and δ are the learning rates for meta-optimization. With the genuine update of the learner in Eq. (10) without CLIP regularization, we can optimize a parallel or batch version that evaluates on N known-to-open tasks of different class distributions (*i.e.*, class-specific knowledge), which encourages to learn more generalizable features against known and static biases.

3.2.2. Gaussian self-ensemble stabilization

Typically, training the video learner for longer iterations to gain specialization on the supervised tasks comes with the risk of diminished plasticity and generalizability. Model patching [22, 42, 50, 51] of weight ensembling has been shown to improve both the performance and generalization. Given that the fine-tuning videos are limited in class-specific knowledge, while the open-vocabulary tasks are unconstrained, the static generalizable flexibility derived from large-scale I-VL pre-training should be scrupulously exploited to enhance the adaptation of the video learner while minimizing the impact of static bias. Therefore, we further incorporate self-ensemble stabilization tailored to the video learner over its optimization trajectory, which utilizes the knowledge from previous training iterations for a generalizable solution. In a fine-tuning procedure of R epochs with l step length for each, the learner's optimization trajectory is represented by $\{\theta_t\}_{t=1}^R$, and θ_0 is the pre-trained weights. The self-ensemble averages the weights of the learner as:

$$\theta_{\text{WA}} = (1 - \sum_{t=1}^R \alpha_t) \cdot \theta_0 + \sum_{t=1}^R \alpha_t \cdot \theta_t, \quad (11)$$

where $\alpha_t \in [0, 1]$ specifies the weights contributed by the parameters at t -th epoch. Intuitively, during the early fine-tuning epochs (*i.e.*, at a smaller epoch t), the video learner lacks the maturity to effectively capture video-specific knowledge while still retaining substantial static-related orientation from large-scale pre-training, which introduces vulnerable information for temporal understanding. Conversely, the parameters at the last few epochs (*i.e.*,

Algorithm 1: Training Procedure

Input: Training set $D_{tr} = \{V_i, T_i\}^M$, Video learner f_{θ} .

Require: GWA Params θ_{GWA} update at each epoch with l step length. CLIP Params θ_{CLIP} . Batch size of training samples B . Learning rate α, β, δ .

Output: The final GWA learner $f_{\theta_{\text{GWA}}}$.

```

1 Initialize  $\theta, \theta_{\text{GWA}} \leftarrow \theta_{\text{CLIP}}$ ; Step = 0;  $t = 0$ 
2 while not converged do
3   Step  $\leftarrow$  Step + 1
4   Construct batch of tasks  $\mathcal{T}_i = \{\mathcal{S}, \mathcal{Q}\}$  by sampling
      $\mathcal{S}, \mathcal{Q} \leftarrow \{V_a, T_a\}^B, \{V_b, T_b\}^B \subseteq D_{tr}$ 
5   forall  $\mathcal{T}_i$  do
6     // meta training
7     Evaluate  $\nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}^{\mathcal{S}}(\theta)$  w.r.t. Eq. (5)
8     Compute adapted parameters with gradient
       decent:  $\theta'_i = \theta - \alpha \nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}^{\mathcal{S}}(\theta)$  w.r.t. Eq. (6)
9   end
10  // meta testing
11  Evaluate  $\nabla_{\theta'_i} \mathcal{L}_{\mathcal{T}_i}^{\mathcal{Q}}(\theta'_i)$  w.r.t. Eq. (7)
12  // meta-optimization
13  Update  $\theta$  w.r.t. Eq. (10)
14  // Gaussian Weight Average
15  if mod(Step,  $l$ ) == 0 then
16     $t \leftarrow t + 1$ ;  $\theta_t \leftarrow \theta$ 
17    Update  $\theta_{\text{GWA}}$  w.r.t. Eq. (13)
18  end
19 end

```

at a larger epoch t) have integrated more video-specific knowledge, highly featuring the supervised downstream task distribution, whereas the plasticity of the unconstrained zero-shot capability is not guaranteed. As both sides degrade the final open-vocabulary generalizability, we aim to weaken the contribution of the parameters near the initial and terminal epochs by employing a distribution prior, resulting in a generic optima for robust generalization.

Driven by [27] in prompt learning, we perform Gaussian Weight Average (GWA) based on model patching, as shown in Fig. 2(b), which assigns the parameters with lower weights at initial epochs, higher weights at middle epochs, and relatively lower weights at final epochs. Given a Gaussian distribution $w_t \sim \mathcal{N}(\mu, \sigma^2)$ defined over the epochs, we sample the weight values for the parameters θ_t as its corresponding probability in the distribution as:

$$w_t = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}}, t = 1, \dots, R. \quad (12)$$

Here, we exclude the integration of CLIP weights θ_0 for the purpose of static debiasing. μ and σ^2 are hyper-parameters for the distribution, and in practice, we determine the value of μ according to the epoch number. Then, we perform normalization towards the weights of total epochs *i.e.*, $\alpha_t =$

Table 1. Performance comparison (Top1-Acc (%)) with the CLIP-adapted methods using ViT-B/16 under the in-context base-to-novel setting. We also report the harmonic mean (HM) of base and novel recognition accuracy. The **best** and the second-best results are highlighted. * and † denote the results reproduced with our implementation using frozen text learners.

Method	Venue	K400			HMDB			UCF			SSv2		
		Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
Frozen CLIP [39]	ICML'21	62.3	53.4	57.5	53.3	46.8	49.8	78.5	63.6	70.3	4.9	5.3	5.1
ActionCLIP [47]	arXiv'21	61.0	46.2	52.6	69.1	37.3	48.5	90.1	58.1	70.7	13.3	10.1	11.5
X-CLIP [34]	ECCV'22	74.1	56.4	64.0	69.4	45.5	55.0	89.9	58.9	71.2	8.5	6.6	7.4
VPT [24]	ECCV'22	69.7	37.6	48.8	46.2	16.0	23.8	90.5	40.4	55.8	8.3	5.3	6.4
ST-Adapter [36]	NeurIPS'22	74.6	62.0	67.3	65.3	48.9	55.9	85.5	76.8	80.9	9.3	8.4	8.8
ViFi-CLIP [40]	CVPR'23	<u>76.4</u>	61.1	67.9	73.8	<u>53.3</u>	<u>61.9</u>	92.9	67.7	78.3	<u>16.2</u>	<u>12.1</u>	<u>13.9</u>
Open-VCLIP * [50]	ICML'23	76.3	<u>62.3</u>	<u>68.6</u>	70.2	50.2	58.5	<u>94.6</u>	<u>77.2</u>	<u>85.0</u>	15.9	10.8	12.9
FROSTER † [21]	ICLR'24	76.0	61.9	68.3	70.0	49.9	58.3	94.3	76.9	84.7	15.5	10.3	12.4
Open-MeDe		77.2	63.8	69.9	<u>73.6</u>	56.4	63.9	94.9	78.5	85.9	17.1	12.3	14.3

$\frac{w_t}{\sum_{i=1}^t w_i}$. We also formulate GWA as a moving average to avoid increasing the storage cost of saving multiple snapshots of the parameters by updating the average of current learner θ_t on the fly (*i.e.*, at epoch t) as:

$$\theta_{\text{GWA}} \leftarrow \frac{\sum_{i=1}^{t-1} w_i}{\sum_{i=1}^t w_i} \cdot \theta_{\text{GWA}} + \frac{w_t}{\sum_{i=1}^t w_i} \cdot \theta_t. \quad (13)$$

3.3. Algorithm overview

We present the overall training procedure of the proposed model-agnostic Open-MeDe in Algorithm 1. The video learner is fine-tuned based on training videos based on our cross-batch meta-optimization scheme cost-effectively. And the Gaussian self-ensemble stabilization is performed on the video learner via our GWA for robust generalization under open-vocabulary settings.

4. Experiments

4.1. Experimental Setup

Datasets. We explore two distinct types of open-vocabulary action recognition evaluation in this work: *in-context* and *out-of-context* settings. For in-context scenarios, we conduct experiments following the common practice in the literature [21, 40, 40, 50] on the Kinetics-400 (K400) [26], UCF-101 (UCF) [43], HMDB-51 (HMDB) [28], Something-Something V2 (SSv2) [18] and Kinetics-600 (K600) [6] datasets under widely-used evaluation protocols: *cross-dataset* and *base-to-novel* evaluation. For more challenging out-of-context scenarios, we newly conduct general cross-dataset evaluations using K400 dataset as the training set and testing on the synthetic UCF-SCUBA [30] and UCF-HAT [2, 12] benchmarks.

Implementation details. Generally, we use the official CLIP ViT-B/16 backbone for all experiments, and our video learner is the adaptation of the CLIP model follows [50], unless stated otherwise. During our meta-optimization process, we construct a batch of 4 tasks, each task contains 8

support and query samples from the training set. The learning rates of inner and outer loops for support batches *i.e.*, α , and β , are synchronized with the initial value of 3.33×10^{-6} and decay to 3.33×10^{-8} utilizing the AdamW [33] optimizer following a cosine decay scheduler, while the hyperparameter δ for query batches is set to 1.67×10^{-3} . For cross-dataset evaluation, we warm up the training on the K400 dataset for the first 2 epochs and further fine-tune the video learner for 20 epochs. For base-to-novel evaluation, we train the learner for 12 epochs with the first two warm-up epochs on training data. During inference, we use 3 temporal and 1 spatial views per video and linearly aggregate the recognition results. See Appendix for experimental details.

4.2. Comparison with state-of-the-art methods

We compare our framework with the state-of-the-art open-vocabulary action recognition methods on the following commonly used *in-context* and newly proposed *out-of-context* evaluation protocols.

In-context base-to-novel generalization. In Tab. 1, we compare the proposed framework with other CLIP-based methods under the popular in-context base-to-novel setting. All methods are initially learned on the frequently occurring base classes and evaluated on both base and novel classes, where the novel classes represent a realm of previously uncounted scenarios. From the results, we can summarize the observations: (1) Most of the methods show reasonable improvements from the frozen CLIP [39], except for ActionCLIP [47], X-CLIP [34] and VPT [24] suffering inferior performances especially on the novel sets of K400, HMDB and UCF, indicating the strong generalization of CLIP and the potential overfitting of these adapted video learners toward the training samples. (2) Our framework experiences noticeable gains in novel class performance and consistent achievements on all four datasets, spanning spatially dense and temporally focused scenarios, which validates the effectiveness of enhancing generalization and static debiasing for both known and open classes.

Table 2. Comparison with the previous methods under the in-context cross-dataset setting. The results are top-1 accuracies (%) with mean and standard deviation on the evaluation across three validation splits within each dataset. * and † denote our re-implementation with frozen text learners.

Method	Venue	UCF	HMDB	K600
Frozen CLIP [39]	ICML'21	73.8±0.6	47.9±0.5	68.1±1.1
ActionCLIP [47]	arXiv'21	77.5±0.8	48.2±1.5	62.5±1.2
X-CLIP [34]	ECCV'22	72.0±2.3	44.6±5.2	65.2±0.4
VPT [24]	ECCV'22	69.3±4.2	44.3±2.2	55.8±0.7
ST-Adapter [36]	NeurIPS'22	77.6±0.7	51.1±0.6	60.2±1.8
Vita-CLIP [49]	CVPR'23	75.0±0.6	48.6±0.6	67.4±0.5
MAXI [31]	ICCV'23	78.2±0.7	52.3±0.6	71.5±0.8
Open-VCLIP * [50]	ICML'23	<u>83.3±1.4</u>	<u>53.8±1.5</u>	<u>73.0±0.8</u>
ViLT-CLIP [46]	AAAI'24	73.6±1.1	45.3±0.9	-
FROSTER † [21]	ICLR'24	82.9±0.6	53.4±1.2	71.1±0.8
VicTR [25]	CVPR'24	72.4±0.3	51.0±1.3	-
ALT [11]	CVPR'24	79.4±0.9	52.9±1.0	72.7±0.6
Open-MeDe		83.7±1.3	54.6±1.1	73.7±0.9

Table 3. Performance comparison (Top-1 / Top-5 Acc. (%)) on UCF dataset. We evaluate both in-context and out-of-context recognition (marked with *) performances. We also report the harmonic mean (HM) of the results. * and † indicate our implementation with frozen text learners.

Method	UCF	UCF-SCUBA *	UCF-HAT *	HM
X-CLIP	74.5 / 95.4	24.6 / 43.3	56.8 / 78.1	20.3 / 64.7
Open-VCLIP *	<u>83.5 / 96.9</u>	<u>28.9 / 48.0</u>	<u>59.6 / 79.5</u>	<u>47.4 / 68.6</u>
FROSTER †	82.9 / 96.4	25.2 / 43.2	58.6 / 78.9	43.6 / 64.9
Ours	83.9 / 96.9	33.5 / 52.7	64.5 / 82.3	52.4 / 72.4

In-context cross-dataset generalization. In Tab. 2, we present the compared results under in-context cross-dataset zero-shot evaluations, where all learners undergo further fine-tuning on K400 training set and are tested directly on downstream cross-datasets *i.e.*, UCF, HMDB and K600. Similar findings can be noticed from the results as base-to-novel evaluations that frozen CLIP outperforms several adapted learners, especially on the most generalizability demanding benchmark, *i.e.*, K600, further demonstrating the generalization degradation of overfitting within these methods. Remarkably, our framework based on meta-learning consistently surpasses state-of-the-art approaches on all three benchmarks, demonstrating its superior effectiveness and enhanced generalizability.

Out-of-context cross-dataset generalization. In Tab. 3, we further compare our method with the previous state of the arts under more challenging out-of-context cross-dataset evaluations on SCUBA and HAT benchmarks of the UCF dataset. It can be noticed that: (1) Integrating with CLIP regularization, both Open-VCLIP [50] and FROSTER [21] achieve promising improvements compared with X-CLIP under original UCF in-context scenarios. (2) However, the compared methods suffer from severely limited generalization when encountering out-of-context scenarios due to the

Table 4. In-context cross-dataset comparison (Top-1 Acc. (%)) when integrating our Open-MeDe with different video learners.

Adaptation	Method	UCF	HMDB	K600
Adapter-based	ST-Adapter [36]	77.6±0.7	51.1±0.6	60.2±1.8
	+ Ours	78.9±1.1	52.0±1.1	72.7±0.8
	Δ Gains	+ 1.3	+ 0.9	+ 12.5
Prompt-based	Vita-CLIP [49]	75.0±0.6	48.6±0.6	67.4±0.5
	+ Ours	77.9±0.8	50.7±1.3	71.5±0.9
	Δ Gains	+ 2.9	+ 2.1	+ 4.1
Partially-tuned	X-CLIP [34]	72.0±2.3	44.6±5.2	65.2±0.4
	+ Ours	79.3±1.3	52.3±1.5	72.9±1.1
	Δ Gains	+ 7.3	+ 7.7	+ 7.7
Fully-tuned	VCLIP [50]	78.5±1.0	50.3±0.8	65.9±1.0
	+ Ours	83.7±1.3	54.6±1.1	73.7±0.9
	Δ Gains	+ 5.2	+ 4.3	+ 7.8

static bias within these video learners. (3) Our method significantly outperforms partially fine-tuned X-CLIP and CLIP regularization methods on various out-of-context scenarios. We outperform the second-best competitor by 4.6% on UCF-SCUBA and 4.9% on UCF-HAT, with the highest HM striking an impressive balancing on cross-dataset generalization for in-context and out-of-context scenarios. We attribute the superiority of our video learner to the natural know-to-open generalizing and image-to-video debiasing via the newly proposed meta-optimization and self-ensemble independent from CLIP’s persistent interference of static biases for robust and generic generalizability.

4.3. Ablation Studies

Applicability with different video learners. In Tab. 4, we adopt other video learners (with the frozen text encoder) from adapter-based ST-Adapter [36], prompt-based Vita-CLIP [49], partially fine-tuned X-CLIP [34] and fully fine-tuned VCLIP [50] to validate the effectiveness of our model-agnostic framework. We find that: (1) All CLIP-adapted video learners integrating with our method achieve consistent improvements on in-context cross-dataset evaluations, highlighting its broad and flexible applicability. (2) Our approach generally exhibits more improvements for partially and fully fine-tuned methods than PEFT learners, suggesting the importance of sufficient fitting capacity (*i.e.*, learnable parameters) for video learners to attain video-specific generalizability.

Effect of cross-batch meta-optimization. In Tab. 5, we conduct experiments to verify the effect of our cross-batch meta-optimization scheme. The compared strategies and analyses are as follows: (a) Consider VCLIP with standard fine-tuning objectives as a baseline of the plain learner. (b) When adopting RFD to VCLIP, the K400 closed-set performance experiences a slight decline for both IC and OC scenarios, while cross-dataset in-context generalization improves, with gains of +4.5% on UCF-IC, whereas it

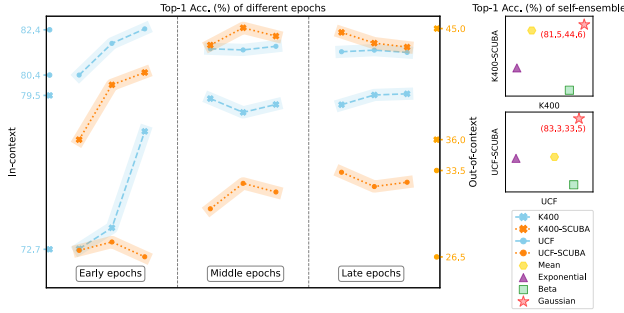


Figure 3. Performance comparison at different epochs vs. various weight self-ensemble strategies. We train the video learner on K400 and test on the in-context UCF, K400, and out-of-context K400-SCUBA and UCF-SCUBA benchmarks. Points on the curves represent epochs of [2, 4, 6], [10, 12, 14] and [18, 20, 22] from left to right, respectively.

Table 5. We compare the performances of different optimization schemes under various settings. IC: in-context evaluations, OC: SCUBA [30] out-of-context evaluations, HM: harmonic mean. RFD: Residual Feature Distillation, IWR: Interpolated Weight Regularization, Meta Unseen: MAML for meta seen to unseen, Meta Cross-batch: our cross-batch meta-optimization.

Optimization	Method	K400 (closed-set)			UCF (zero-shot)		
		IC	OC	HM	IC	OC	HM
Plain	(a) VCLIP [50]	80.1	42.4	55.4	78.5	28.3	41.6
	(b) + RFD [21]	79.9	41.5	54.6	82.5	25.2	38.9
	(c) + IWR [50]	80.5	40.3	53.7	82.9	28.9	42.9
Meta learning	(d) + Meta Unseen [45]	79.5	41.7	54.7	83.2	31.8	46.0
	(e) + Meta Cross-batch	81.5	46.6	59.3	83.9	33.5	47.9

severely impairs generalization for UCF-OC (-3.1%). (c) Similar results are observed when integrating IWR regularization with VCLIP. (d) For the previous meta unseen optimization method for zero-shot learning, all three accuracies under UCF cross-dataset evaluation increase, where K400 evaluations challenge its closed-set generalizations, indicating the potential overfitting to meta unseen tasks. (e) Notably, our cross-batch meta-optimization scheme ((a) \rightarrow (e)) enhances all closed-set and zero-shot performance on harmonic mean with gains of $+3.9\%$ and $+6.3\%$, respectively. This showcases the superiority of our scheme for enhancing know-to-open generalizing and image-to-video debiasing, which establishes a promising balance for robust generalization capabilities.

Effect of weight self-ensemble. In Fig. 3, we investigate the trend of generalization performance during K400 training and the efficacy of weight self-ensemble stabilization using various strategies. In particular, the curves illustrate the performance within the video learner’s optimization trajectory at different epochs, where the x -axis and y -axis display the different stages of training epochs and various generalization evaluation protocols, respectively. It is notice-

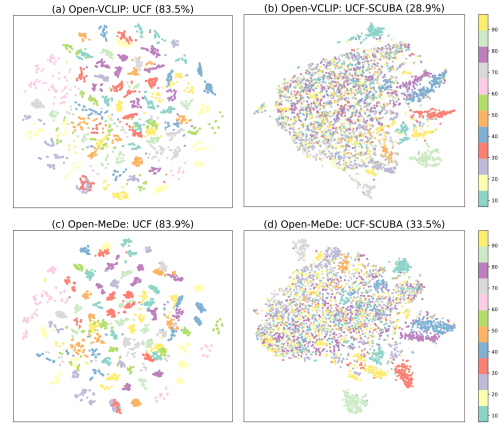


Figure 4. t-SNE [44] visualization of the predictions from Open-VCLIP and our Open-MeDe on UCF and UCF-SCUBA.

able that the overall performance has experienced trends of significant enhancement on both closed-set and zero-shot generalization while quickly leading to drops in zero-shot performance at the tail of the fine-tuning phase, suggesting the plasticity degradation that highly features supervised task-specific distributions on the downstream dataset. The results show that weight ensembling methods improve both specialty and generalizability, with our Gaussian self-ensemble excelling significantly, strongly suggesting it as a better choice for robust generalization.

4.4. Visualizations

Fig. 4 compares the t-SNE visualizations of Open-VCLIP and our framework for in-context and out-of-context UCF predictions. Note that our predictions for videos within the same category are more concentrated, with reduced confusion between different categories, compared to Open-VCLIP. This suggests that the proposed framework effectively learns temporal information, mitigating known and static biases while demonstrating robust generalizability. However, there remains considerable room for improvement in out-of-context scenarios for video-adapted learners.

5. Conclusion

We introduce Open-MeDe, a novel meta-learning framework for open-vocabulary action recognition. It adopts a cross-batch meta-optimization, which encourages the video learner to attain generalizable knowledge counteracting inherent known and static biases for effective known-to-open generalizing and image-to-video debiasing. It also incorporates Gaussian Weight Average to achieve generic optima for robust generalization. Extensive evaluations in both in-context and out-of-context open-vocabulary scenarios validate the applicability and superiority of our framework.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (No. 62376217, 62273347, and 62301434), and the Young Elite Scientists Sponsorship Program by CAST (No. 2023QNRC001).

References

- [1] Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your maml. In *International conference on learning representations*, 2018. 2
- [2] Kyungho Bae, Geo Ahn, Youngrae Kim, and Jinwoo Choi. Devias: Learning disentangled video representations of action and scene for holistic video understanding. *arXiv preprint arXiv:2312.00826*, 2023. 6
- [3] Biagio Brattoli, Joseph Tighe, Fedor Zhdanov, Pietro Perona, and Krzysztof 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*, pages 4613–4623, 2020. 1
- [4] Congqi Cao, Hanwen Zhang, Yue Lu, Peng Wang, and Yanning Zhang. Scene-dependent prediction in latent space for video anomaly detection and anticipation. *IEEE transactions on pattern analysis and machine intelligence*, 2024. 1
- [5] Congqi Cao, Yueran Zhang, Yating Yu, Qinyi Lv, Lingtong Min, and Yanning Zhang. Task-adapter: Task-specific adaptation of image models for few-shot action recognition. In *ACM Multimedia 2024*, 2024. 2
- [6] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 6
- [7] Santiago Castro and Fabian Caba Heilbron. Fitclip: Refining large-scale pretrained image-text models for zero-shot video understanding tasks. *arXiv preprint arXiv:2203.13371*, 2022. 2
- [8] Tongjia Chen, Hongshan Yu, Zhengeng Yang, Zechuan Li, Wei Sun, and Chen Chen. Ost: Refining text knowledge with optimal spatio-temporal descriptor for general video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18888–18898, 2024. 1
- [9] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021. 3
- [10] Yudong Chen, Sen Wang, Jiajun Liu, Xuwei Xu, Frank de Hoog, and Zi Huang. Improved feature distillation via projector ensemble. *Advances in Neural Information Processing Systems*, 35:12084–12095, 2022. 2
- [11] Yifei Chen, Dapeng Chen, Ruijin Liu, Sai Zhou, Wenyuan Xue, and Wei Peng. Align before adapt: Leveraging entity-to-region alignments for generalizable video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18688–18698, 2024. 7
- [12] Jihoon Chung, Yu Wu, and Olga Russakovsky. Enabling detailed action recognition evaluation through video dataset augmentation. *Advances in Neural Information Processing Systems*, 35:39020–39033, 2022. 6
- [13] Wenliang Dai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. Enabling multimodal generation on clip via vision-language knowledge distillation. *arXiv preprint arXiv:2203.06386*, 2022. 2
- [14] Xiang Deng and Zhongfei Zhang. Comprehensive knowledge distillation with causal intervention. *Advances in Neural Information Processing Systems*, 34:22158–22170, 2021. 2
- [15] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3
- [16] Haodong Duan, Yue Zhao, Kai Chen, Yuanjun Xiong, and Dahua Lin. Mitigating representation bias in action recognition: Algorithms and benchmarks. In *European Conference on Computer Vision*, pages 557–575. Springer, 2022. 2
- [17] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017. 2, 3
- [18] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The” something something” video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 6
- [19] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5149–5169, 2021. 2
- [20] He Huang, Changhu Wang, Philip S Yu, and Chang-Dong Wang. Generative dual adversarial network for generalized zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 801–810, 2019. 3
- [21] Xiaohu Huang, Hao Zhou, Kun Yao, and Kai Han. Froster: Frozen clip is a strong teacher for open-vocabulary action recognition. In *International Conference on Learning Representations*, 2024. 1, 2, 6, 7, 8
- [22] Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. Patching open-vocabulary models by interpolating weights. *Advances in Neural Information Processing Systems*, 35:29262–29277, 2022. 2, 5
- [23] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1
- [24] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *European Conference on Computer Vision*, pages 105–124. Springer, 2022. 2, 6, 7

- [25] Kumara Kahatapitiya, Anurag Arnab, Arsha Nagrani, and Michael S Ryoo. Victr: Video-conditioned text representations for activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18547–18558, 2024. 7
- [26] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1, 6
- [27] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15190–15200, 2023. 5
- [28] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 6
- [29] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 3
- [30] Haoxin Li, Yuan Liu, Hanwang Zhang, and Boyang Li. Mitigating and evaluating static bias of action representations in the background and the foreground. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19911–19923, 2023. 1, 2, 6, 8
- [31] Wei Lin, Leonid Karlinsky, Nina Shvetsova, Horst Possegger, Mateusz Kozinski, Rameswar Panda, Rogerio Feris, Hilde Kuehne, and Horst Bischof. Match, expand and improve: Unsupervised finetuning for zero-shot action recognition with language knowledge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2851–2862, 2023. 7
- [32] Zhe Liu, Yun Li, Lina Yao, Xianzhi Wang, and Guodong Long. Task aligned generative meta-learning for zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8723–8731, 2021. 3, 4
- [33] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6
- [34] 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 *European Conference on Computer Vision*, pages 1–18. Springer, 2022. 1, 2, 6, 7
- [35] A Nichol. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. 2
- [36] Junting Pan, Ziyi Lin, Xiayan Zhu, Jing Shao, and Hongsheng Li. St-adapter: Parameter-efficient image-to-video transfer learning. *Advances in Neural Information Processing Systems*, 35:26462–26477, 2022. 1, 2, 6, 7
- [37] Jinyoung Park, Juyeon Ko, and Hyunwoo J Kim. Prompt learning via meta-regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26940–26950, 2024. 3
- [38] Renjing Pei, Jianzhuang Liu, Weimian Li, Bin Shao, Songcen Xu, Peng Dai, Juwei Lu, and Youliang Yan. Clipping: Distilling clip-based models with a student base for video-language retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18983–18992, 2023. 2
- [39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 6, 7
- [40] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6545–6554, 2023. 2, 6
- [41] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 2
- [42] Yang Shu, Xingzhuo Guo, Jialong Wu, Ximei Wang, Jianmin Wang, and Mingsheng Long. Clipood: Generalizing clip to out-of-distributions. In *International Conference on Machine Learning*, pages 31716–31731. PMLR, 2023. 5
- [43] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1, 2, 6
- [44] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 8
- [45] Vinay Kumar Verma, Dhanajit Brahma, and Piyush Rai. Meta-learning for generalized zero-shot learning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6062–6069, 2020. 3, 4, 8
- [46] Hao Wang, Fang Liu, Licheng Jiao, Jiahao Wang, Zehua Hao, Shuo Li, Lingling Li, Puhua Chen, and Xu Liu. Vilt-clip: Video and language tuning clip with multimodal prompt learning and scenario-guided optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5390–5400, 2024. 7
- [47] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 1, 2, 3, 6, 7
- [48] Yizhe Wang, Congqi Cao, and Yanning Zhang. Visual-semantic network: a visual and semantic enhanced model for gesture recognition. *Visual Intelligence*, 1(1):25, 2023. 1
- [49] Syed Talal Wasim, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Vita-clip: Video and text adaptive clip via multimodal prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23034–23044, 2023. 1, 2, 3, 7
- [50] Zejia Weng, Xitong Yang, Ang Li, Zuxuan Wu, and Yu-Gang Jiang. Open-vclip: Transforming clip to an open-vocabulary video model via interpolated weight optimization. In *International Conference on Machine Learning*, pages 36978–36989. PMLR, 2023. 2, 3, 5, 6, 7, 8

- [51] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7959–7971, 2022. [5](#)
- [52] Wenhao Wu, Zhun Sun, and Wanli Ouyang. Revisiting classifier: Transferring vision-language models for video recognition. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2847–2855, 2023. [1](#), [2](#), [3](#)
- [53] Zuxuan Wu, Zejia Weng, Wujian Peng, Xitong Yang, Ang Li, Larry S Davis, and Yu-Gang Jiang. Building an open-vocabulary video clip model with better architectures, optimization and data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. [1](#)
- [54] Li Xu, Haoxuan Qu, Jason Kuen, Jiuxiang Gu, and Jun Liu. Meta spatio-temporal debiasing for video scene graph generation. In *European Conference on Computer Vision*, pages 374–390. Springer, 2022. [3](#)
- [55] Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. Aim: Adapting image models for efficient video action recognition. *arXiv preprint arXiv:2302.03024*, 2023. [2](#), [3](#)
- [56] Yating Yu, Congqi Cao, Yueran Zhang, Qinyi Lv, Lingtong Min, and Yanning Zhang. Building a multi-modal spatiotemporal expert for zero-shot action recognition with clip. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9689–9697, 2025. [2](#)
- [57] Yan Zhu, Junbao Zhuo, Bin Ma, Jiajia Geng, Xiaoming Wei, Xiaolin Wei, and Shuhui Wang. Orthogonal temporal interpolation for zero-shot video recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7491–7501, 2023. [1](#), [3](#)