Learning to Generalize without Bias for Open-Vocabulary Action Recognition

Supplementary Material

This supplementary material provides additional details and further experiments to complement the main paper. The content is organized as follows:

A. Additional Experimental Details (Appendix § A)
B. Additional Experimental Results (Appendix § B)
C. Discussions (Appendix § C)
D. Broader Impacts and Limitations (Appendix § D)

A. Additional Experimental Details

A.1. Datasets

In this work, we categorize the datasets into *in-context* and *out-of-context* datasets. The videos from *in-context* datasets consist of actions with frequent static context, *e.g*. swimming in the swimming pool, while the videos from *out-of-context* datasets contain actions occurring with an unusual static context, *e.g*. dancing in the mall [3]. We conduct the experiments on five *in-context* benchmarks: Kinectics-400 [8] (K400), Kinectis-600 [2] (K600), UCF101 [14] (UCF), HMDB51 [10] (HMDB), and Something-Something V2 [7] (SSv2). Additionally, we evaluate our approach on two *out-of-context* benchmarks: SCUBA [11] and HAT [4].

K400 and K600 are both comprehensive video datasets for human action recognition. K400 contains 400 action categories of approximately 240k training and 20k validation videos collected from YouTube, which covers a wide range of human actions, including sports activities, daily life actions, and various interactions, serving as a widely-used action recognition dataset for pre-training. The duration of video clips in K400 varies, with most clips being around 10 seconds long. This diversity in video duration helps models learn temporal dynamics and context for action recognition. K600 extends K400 by incorporating 220 additional new categories, thus enabling the evaluation of zero-shot learning capabilities on these novel categories.

UCF is a human action recognition dataset collected from YouTube, and consists of 13,320 video clips, which are classified into 101 categories. These 101 categories encompass a wide range of realistic actions including body motion, human-human interactions, human-object interactions, playing musical instruments and sports. Officially, there are three splits allocating 9,537 videos for training and 3,783 videos for testing.

HMDB is a relatively small video dataset comprising a diverse range of sources, including movies, public databases, and YouTube videos, and is composed of 6,766 videos across 51 action categories (such as "jump", "kiss" and "laugh"), ensuring at least 101 clips within each category. The original evaluation scheme employs three distinct training/testing splits, allocating 70 clips for training and 30 clips for testing of each category in each split.

SSv2 is a temporally focused video dataset across 174 fine-grained action categories, consisting of 168,913 training videos and 24,777 testing videos showing the objects and the actions performed on them. These action categories are presented using object-agnostic templates, such as "Dropping [something] into [something]" containing slots ("[something]") that serve as placeholders for objects. This dataset focuses on basic, physical concepts rather than higher-level human activities, which challenges the temporal modeling capabilities.

SCUBA is an out-of-distribution (OOD) video benchmark designed to quantitatively evaluate static bias in the background. It comprises synthetic out-of-context videos derived from the first test split of HMDB and UCF, as well as the validation set of K400. These videos are created by superimposing action regions from one video onto diverse scenes, including those from Place365 [18] and VQGAN-CLIP [5] generated scenes. Due to the differences in test sets and background sources, the domain gaps of SCUBA benchmarks vary. A domain gap is defined as the ratio of accuracies between the original test sets and synthetic datasets obtained by a 2D reference network, where a higher ratio indicates a greater domain gap with respect to static features. The UCF-SCUBA and K400-SCUBA used in our experiments consist of 4,550 and 10,190 videos with domain gaps of 20.49 and 6.09, respectively, whose backgrounds are replaced by the test set of Place365.

HAT is a more "realistic-looking" mixed-up benchmark for quantitative evaluation of the background bias by automatically generating synthetic counterfactual validation videos with different visual cues. It provides four Action-Swap sets with distinct characteristics: *Random* and *Same* refer to the swap of actions and backgrounds from different and same classes, respectively, while *Close* and *Far* denote the swap of videos from a class with similar and very different backgrounds, respectively. The UCF-HAT benchmark used in our experiments consists of Action-Swap videos in *Close* and *Far* sets from 5 closest and 30 farthest action categories, respectively, following the literature [4]. Note that we only consider videos from the first test split of UCF where all frames have human masks taking up 5% to 50% of the pixels to ensure that sufficient human and background cues are present in each generated Action-Swap video.

A.2. Evaluation Protocols

For the experimental settings, we follow the previous works [12, 13, 16] for in-context generalization evaluations and perform the newly proposed out-of-context generalization evaluations described below.

In-context base-to-novel generalization. Under this setting, we divide the entire set of action categories into two equal halves: base and novel, with the most frequently occurring classes designated as the base classes. We conduct generalization evaluations on four in-context datasets, *i.e*. K400, HMDB, UCF and SSv2, where the models are initially trained on the base classes within the training splits of the dataset, and evaluated on both base and novel classes within the validation splits. Every training split consists of 16 video clips of each base class. During inference within HMDB and UCF datasets, only the novel class samples in the first validation splits are used for evaluation. For K400 and SSv2 datasets, the full validation split of each is used for evaluation here. We report the results of the average top-1 accuracies for both base and novel classes as well as the harmonic mean.

In-context cross-dataset generalization. Under this setting, the models are fine-tuned on the training set of K400, and evaluated on three in-context cross-datasets, *i.e*. UCF, HMDB and K600. We report top-1 average accuracies with performance variances on the three validation splits in case of UCF and HMDB. For K600, the models are evaluated on non-overlapping 220 categories with K400, and we report top-1 average accuracies over three randomly sampled splits of 160 categories.

Out-of-context cross-dataset generalization. Under the more challenging out-of-context cross-dataset setting, the models are also trained on K400, and then evaluated on two out-of-context datasets based on UCF, *i.e*. UCF-SCUBA and UCF-HAT. We report the top-1 and top-5 average accuracies over the synthetic counterfactual validation splits from UCF's first validation split. We further conduct the closed-set out-of-context evaluation based on the K400-SCUBA benchmark and report the harmonic mean of the accuracies under in-context and out-of-context settings to comprehensively analyze the generalization of the models.

A.3. Implementation Details

Each training video clip is sampled with 8 frames uniformly, and each sampled frame is spatially scaled in the shorter side to 256 pixels and is processed with basic augmentations like color jittering, random flipping and random cropping of 224×224. We leverage multi-view inference with 3 temporal and 1 spatial views per video and linearly aggregate the recognition results. For our Gaussian Weight Average scheme, we use $\mu = 7$ and $\sigma^2 = 10$ for in-context base-to-novel generalization and $\mu = 15$ and $\sigma^2 = 10$ for in-context and out-of-context cross-dataset generalization.

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

Method	HMDB	HMDB-SCUBA $\star$	HM
X-CLIP	44.6 ± 5.2	22.5	31.0
Open-VCLIP $*$	53.8 ± 1.5	25.9	35.0
FROSTER †	53.4 ± 1.2	23.7	32.8
Ours	**54.6 ± 1.1**	**32.5**	**40.7**

Table 2. Effect of the meta-optimization and Gaussian weight average (GWA) components in Open-MeDe. Δ denotes the performance gains of different schemes over the baseline. Our Open-MeDe is highlighted in gray.

Meta optimization	GWA	UCF	Δ_{UCF}	UCF-SCUBA	$\Delta_{\text{UCF-SCUBA}}$
✗	✗	78.5	-	28.3	-
✓	✗	83.2	+ 4.7	32.1	+ 3.8
✗	✓	82.3	+ 3.8	30.7	+ 2.4
✓	✓	**83.9**	+ 5.4	**33.5**	+ 5.2

We also adopt decision aggregation with pre-trained CLIP with the video learner for in-context evaluations. The experiments are conducted on two computing clusters with four NVIDIA RTX 24G 4090 GPUs.

B. Additional Experimental Results

B.1. Additional Evaluations and Ablation Studies

Out-of-context cross-dataset evaluation on HMDB dataset. Regarding results shown in Tab. 1, our method achieves the highest accuracy of 32.5% on HMDB-SCUBA, and builds up an impregnable lead of $+5.7\%$ of HM results over the nearest competitor, enabling a superior balance for open-vocabulary generalization.

Effect of individual strategies in Open-MeDe. In Tab. 2, we conduct ablation experiments to study effects of the core strategies in Open-MeDe *i.e*. the cross-batch meta-optimization and GWA stabilization. Using only meta-optimization or GWA yields improvements of $+4.7\%/3.8\%$ and $+3.8\%/2.4\%$ over the plain learner on UCF / UCF-SCUBA, respectively. This indicates that meta-optimization substantially enhances generalization across both open-vocabulary settings in improving model's robustness, compared to GWA. These two components complement each other effectively, achieving substantial gains of $+5.4\%/5.2\%$. Their integration leads to consistent improvements across diverse scenarios.

Effect of the learning rate δ. As shown in Tab. 3, we conduct experiments by setting the learning rate δ to different magnitudes. It can be observed that as δ decreases, the general performance remains stable, which validates the

Table 3. Effect of the learning rate δ for meta-optimization. We choose $\delta = 1.67 \times 10^{-3}$ as the default setting.

δ	UCF	HMDB	K600	UCF-SCUBA
1.67×10^{-1}	83.7	54.3	73.5	33.2
1.67×10^{-2}	83.7	54.5	73.6	33.4
1.67×10^{-3}	83.7	54.6	73.7	33.5
1.67×10^{-4}	83.6	54.3	73.6	33.0

Table 4. Effect of cross-batch meta-optimization.

Method	UCF	HMDB	K600	UCF-SCUBA
Plain	78.5	50.3	65.9	28.3
Grad Accumulation	78.9	50.5	66.5	28.9
Meta Cross-batch	**83.7**	**54.6**	**73.7**	**33.5**

Table 5. Effect of the randomness of the batch sampler for cross-batch meta-optimization. The "*similar*" sampler denotes the usage of the most semantically similar classes across adjacent batches. We evaluate both in-context and out-of-context recognition (marked with $\star$) performances. HM: harmonic mean. Our default settings and results are highlighted in gray.

Method	Sampler	UCF (%)	UCF-SCUBA$\star$ (%)	HM (%)
Plain	*shuffle*	78.5	28.3	41.6
	initial	77.7 (↓ 0.8)	28.2 (↓ 0.1)	41.4 (↓ 0.2)
Meta Cross-batch	***shuffle***	**83.7**	**33.5**	**47.8**
	initial	82.5 (↓ 1.2)	28.9 (↓ 4.6)	42.8 (↓ 5.0)
	similar	82.7 (↓ 1.0)	30.9 (↓ 2.6)	45.0 (↓ 2.8)

robustness of our cross-batch meta-optimization. However, a further reduction to 1.67×10^{-4} slightly decreases performance across most datasets, suggesting that the optimal value for δ lies at 1.67×10^{-3}, which is chosen as the default setting. This value achieves a balanced performance with the highest or nearly highest scores in each dataset, particularly noticeable on UCF-SCUBA benchmark.

Effect of cross-batch meta-optimization. To investigate the efficacy of our cross-batch meta-optimization complementing the main paper, we further evaluate the performance using the scheme of gradient accumulation. To ensure a fair comparison of the total gradient steps with cross-batch meta-optimization, we accumulate the gradients over two steps before performing a single parameter update. As shown in Tab. 4, the gradient accumulation technique demonstrates modest improvements over the plain method for both in-context and out-of-context benchmarks. This indicates that the strength of our meta-optimization approach lies in its ability to enhance known-to-open generalization, rather than doubling the batch size for a single parameter update.

Effect of randomness of the batch sampler for cross-batch meta-optimization. To verify the efficacy of constructing tasks across batches with different inherent label distributions, we further conduct several additional studies about the sampling randomness during cross-batch meta-optimization. As shown in Tab. 5, the randomness of the batch sampler is indeed an important factor to bring out the best of our cross-batch meta leaner, which improves the overall generalization greatly ($+5.0\%$ of harmonic mean) especially for out-of-context performance ($+4.6\%$ on UCF-SCUBA). However, Plain learner shows insensibility to the sampling randomness, experiencing negligible growth of generalization performance. Without shuffling the batch sampler, our method still outperforms the non-shuffle plain learner by $+1.4\%$ of HM results. By using the most semantically similar classes across support and query batches, it brings a relative performance *decline* of 0.99% and 2.64% on UCF and UCF-SCUBA, respectively. We speculate that it amplifies inter-task semantic distribution shifts hindering cross-task generalization. In contrast, the default ensures consistent and balanced distributions of both inter- and intra-task variance.

Table 6. Effect of the batch size of tasks and samples for cross-batch meta-optimization. Our default settings and results are highlighted in gray.

Batchsize		UCF	HMDB	K600	UCF-SCUBA
Task	Sample				
2	8	83.5	54.3	73.2	33.5
4	4	83.5	54.3	73.3	33.4
4	8	83.7	54.6	73.7	33.5
4	16	83.8	54.6	73.9	33.6
8	8	83.8	54.8	73.9	33.6

Effect of the batch size of tasks and samples. In Tab. 6, we evaluate the performance with different batch sizes of the task and data for cross-batch meta-optimization. Each task consists of two data batches, one for the support set and one for the query set. From the results, we observe that increasing the batch size leads to slight improvements in performance, especially for K600. While larger batch sizes provide marginal improvements, they may not justify the increased computational cost. Thus, the default setting provides an effective balance between performance and computational efficiency.

Effect of the CLIP ensemble. In Tab. 7, we evaluate the effectiveness of the CLIP ensemble in the weight space and decision space, with the ensemble ratios all set to 0.5. The results demonstrate that both types of CLIP ensemble improve performance in in-context evaluations, with the prediction-based ensemble yielding the most consistent gains across all methods. This suggests that integrating CLIP predictions effectively leverages the strengths of CLIP, leading to significant performance enhancements, particularly over the naive approach. However, there is a noticeable drop on UCF-SCUBA for the out-of-context generalization, indicating that the static generalization derived

Table 7. Effect of the CLIP ensemble. We evaluate the performance of integrating the CLIP ensemble within the weight and decision spaces. *Naive* denotes applying only the video learners for evaluations without further CLIP ensemble.

Method	CLIP ensemble	UCF	HMDB	K600	UCF-SCUBA
VCLIP	Naive	78.5	50.3	65.9	28.3
	Weight	80.1	51.9	71.0	26.6
	Prediction	80.3	52.1	71.2	27.0
Open-VCLIP	Naive	81.4	53.2	71.5	30.0
	Weight	83.3	53.8	73.0	28.9
	Prediction	83.4	54.0	73.2	29.9
Open-MeDe	Naive	83.3	54.3	73.5	**33.5**
	Weight	83.6	54.4	73.6	29.9
	Prediction	**83.7**	**54.6**	**73.7**	32.0

Table 8. Effect of mitigating static bias in action recognition with various training strategies. We report the Top-1 Acc. (%) and harmonic mean (HM) of both in-context (IC) and out-of-context (OC) generalization performance for closed-set and zero-shot action recognition. ✗ indicates that the methods are not capable of zero-shot action recognition.

Method	Pretrain	Training Strategy	K400 (closed-set)			UCF (zero-shot)		
			IC	OC	HM	IC	OC	HM
BE [15]	ImageNet	Debiasing	73.9	41.9	53.5	✗	✗	✗
FAME [6]	K400	Debiasing	73.8	49.0	58.9	✗	✗	✗
StillMix [11]	ImageNet	Debiasing	73.9	43.4	54.7	✗	✗	✗
DEVIAS [1]	VideoMAE	Disentangle	77.3	**51.8**	**62.0**	✗	✗	✗
VCLIP	CLIP	Plain	80.1	42.4	55.4	78.5	28.3	41.6
Open-MeDe	CLIP	Meta-optimization	**81.5**	46.6	59.3	**83.9**	**33.5**	**47.9**

from the CLIP ensemble can adversely affect the model's robustness and overall generalizability.

Effect of static debiasing strategies. In Tab. 8, we compare Open-MeDe with several baselines especially designed for mitigating static bias in action recognition, including three scene-debiasing methods (BE [15], FAME [6] and StillMix [11]) and a state-of-the-art action-scene disentanglement method (DEVIAS [1]). Note that DEVIAS leverages additional scene labels for disentangled video representation. As can be seen from the results, while FAME and DEVIAS perform well in the K400 closed-set out-of-context evaluation against static bias, they fall short in in-context performance and lack zero-shot inference capability. In contrast, our Open-MeDe, despite not employing explicit debiasing or disentangled action modeling, achieves favorable out-of-context generalization with a balanced harmonic mean. This highlights its robust generalizability across both in-context and out-of-context scenarios, particularly excelling in zero-shot generalization.

Analysis of class-wise performance. In Fig. 1, we further present the improvements of our Open-MeDe over Open-VCLIP on out-of-context UCF-SCUBA across 22 novel classes. It can be observed that Open-MeDe wins across 19 of the 22 classes. The improved categories involve localized motions, where most of the static content is misinterpreted by irrelevant context noise on UCF-SCUBA. We attribute these gains primarily to its effectiveness in static debiasing and capturing fine-grained dynamics. However, its performance is slightly compromised in cases involving team sports or rapid shifts in spatial locations.

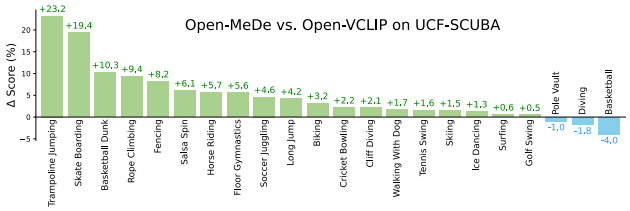

Figure 1. Comparison between Open-MeDe and OpenVCLIP across 22 classes on UCF-SCUBA.

Table 9. Comparison of the training cost. We report the results of K400 training on four GPUs (24G RTX 4090). We maintain an equal batch size of 8 videos per GPU across all models.

Method	Params (M)	FLOPs (G)	CUDA mem. (GB)	Epoch time (min)
VCLIP	149.62	152.11	14.14	110.10
Open-VCLIP	149.62	152.11	20.09	109.26
FROSTER	299.77	152.11	21.07	80.95
Open-MeDe	149.62	152.11	16.59	74.33

B.2. Training cost analysis

In Tab. 9, we show the training cost analysis of our approach and compare it with other methods under identical training conditions. All approaches utilize the same video learner, ensuring equal GFLOPs. Our Open-MeDe achieves the lowest CUDA memory usage at 16.59 GB and a significantly reduced epoch time of 74.33 minutes, compared to other methods. This demonstrates its efficiency in terms of training time and memory consumption, providing a cost-effective solution without compromising on computational complexity.

B.3. Visualization Results

As shown in Figs. 2 to 6, we present additional visualization comparisons of Open-VCLIP and the proposed framework under in-context and out-of-context scenarios. Overall, our approach effectively attends to more motion-relevant regions, achieving higher confidence scores and correct predictions in most cases. This demonstrates its greater reliability, and robust generalizability in open-vocabulary action recognition tasks.

C. Discussions

In this part, we further elucidate the core distinction between the proposed method and similar paradigms through comparative analysis.

Meta learner *vs*. Plain learner. As discussed in the main paper, Open-MeDe formulates the video learner into a meta

learner by employing the cross-batch meta-optimization scheme that mimics sequences of known-to-open generalization tasks, enhancing adaptability to unseen data through iterative virtual evaluations during training. Plain learners, such as those employing standard fine-tuning paradigms on CLIP-based video learners, are typically straightforward and focus on in-distribution class-specific knowledge. Following a traditional gradient descent over a single objective function can lead to a narrower optimization landscape prone to overfitting. Therefore, plain learners can gain reasonable in-context performance but struggle to generalize to novel and out-of-context scenarios due to the tendency to overfit in training data and short-cutting static cues.

In contrast, our meta learner is designed to derive the training towards learning more generalizable features by optimizing not just for class-specific knowledge but for adaptability across diverse known-to-open tasks. It explicitly counteracts inherent known and static biases by leveraging feedback from virtual evaluations, ensuring the video learner does not over-rely on vulnerable static cues. By alternating between *meta training* (*w.r.t*. support data) and *meta testing* (*w.r.t*. query data), the meta learner ensures smoother optimization trajectories and enhanced robustness in a cost-effective manner. The episodic training of the meta learner fosters adaptability across varying class distributions, making it highly effective for open-vocabulary tasks.

Meta-optimization *vs*. Train-validation. In our meta-optimization framework, training involves two key stages: *meta training* (on support data) and *meta testing* (on query data). The query data evaluation provides generalization feedback via loss gradients, enabling the learner to adjust the learning trajectory to prioritize generalizable features. This iterative approach inherently targets learning to generalize and mitigates overfitting by encouraging robust learning across diverse distribution shifts. Conversely, the train-validation paradigm typically partitions data into training and validation subsets, optimizing model parameters by minimizing errors on the training data while evaluating performance on a held-out validation set for hyper-parameter tuning or early stopping. This paradigm monitors the generalization performance indirectly by balancing the performance between training and validation data without explicitly improving the open-vocabulary generalization capability toward novel data.

Both paradigms leverage the feedback to refine model training, where the feedback of meta-optimization comes from query evaluations, while in train-validation, it arises from validation performance. Additionally, the feedback of train-validation is aggregated at coarser intervals, limited to hyper-parameter adjustment on constant training-validation splits. It is worth noting that the meta-optimization provides granular, iterative feedback during training, manifesting as loss gradients to refine generalizable representation learning by dynamically constructing tasks with support-query splits. Therefore, the proposed meta-optimization framework provides a more robust and explicit mechanism for adapting to novel data, setting a new baseline for open-vocabulary action recognition.

Cross-batch meta-optimization *vs*. Gradient accumulation. As introduced in the Open-MeDe framework, the proposed cross-batch meta-optimization takes inspiration from meta-learning with minimal modification to the standard training setup, which leverages adjacent mini-batches in training, treating one as the support batch (*meta training*) and the subsequent as the query batch (*meta testing*). It aims to explicitly promote generalization by evaluating how well the model can adapt its learned parameters to open or dynamically different data distributions, thereby mitigating inherent and static biases in the video learner. When it comes to the gradient accumulation technique, by simulating large batch training, it aggregates gradients over multiple mini-batches and applies the update after a predefined number of steps, emphasizing the efficiency of stabilizing training and improving convergence on hardware-constrained scenarios. However, it primarily improves training stability without inherently targeting adaptability and enhanced generalization. Therefore, cross-batch meta-optimization differs fundamentally from gradient accumulation in its goal and methodology, which achieves a superior balance between specialization and generalization.

Meta-debiasing with MVSGG [17]. 1) Objective *w.r.t*. mitigating biases. MVSGG addresses certain conditional biases within video scene generation tasks, targeting long-tailed data issues. Here, we tackle a ubiquitous challenge for video understanding, *i.e*. mitigating static bias present in video learners. 2) Methodology *w.r.t*. meta-optimization. MVSGG emphasizes on constructing various types of conditional biases within data at each training epoch, with its meta-optimization employed per epoch against specific biases. We perform meta-optimization densely in iterations with a diverse distribution of tasks. The evaluation on a subsequent batch explicitly simulates known-to-open generalization and mitigates static bias implicitly. 3) Application scope *w.r.t*. generalization. MVSGG enhances model's generalization under closed-set settings against conditional biases within training data. Notably, we achieve more robust open-vocabulary generalization beyond training data, where MVSGG is insufficient to our requirements. 4) Computational cost *w.r.t*. task construction. MVSGG requires careful organization of training data, significantly increasing computational cost. Remarkably, our method incurs no additional computational overhead compared to standard training by effortlessly utilizing cross-batch data.

Gaussian self-ensemble with PromptSRC [9]. Our GWA is related to PromptSRC with two key differences: 1) Ob-

jective *w.r.t.* implementation. We aim to achieve a generic optimal solution for video learners by assigning different weights to learner's parameters during optimization, while PromptSRC focuses on regularizing prompt learning to reduce overfitting with frozen backbones. 2) Patching strategy *w.r.t.* start point. Our GWA starts after fine-tuning the pre-trained weights of the learner (*e.g.*, CLIP weights), which exhibits substantial static-related knowledge. With the purpose of mitigating static bias, the initial patching weights are sampled from low Gaussian probabilities. However, the start point of PromptSRC is randomly initialized, given the prompt learning framework, where lower weight assignments guarantee the task-specific knowledge.

D. Broader Impacts and Limitations

Broader Impacts. The proposed Open-MeDe framework for open-vocabulary action recognition introduces substantial advancements in several key aspects, underscoring its broader impact on both research and real-world applications: 1) By addressing the overfitting to static cues inherent in pre-trained models like CLIP, Open-MeDe introduces innovative solutions for robust generalization. Its combination of meta-optimization and Gaussian self-ensemble stabilization enables robust performance in challenging out-of-context scenarios, providing a pathway for video learners to bridge the gap between image and video modalities effectively. 2) Unlike previous approaches reliant on CLIP regularization, Open-MeDe reduces computational overhead and efficiently balances class-specific learning with generalization capabilities, leveraging a cross-batch meta-optimization approach. 3) Open-MeDe demonstrates remarkable adaptability across diverse scenarios, including base-to-novel, cross-dataset, and out-of-context evaluations. Its model-agnostic design enables seamless integration with various CLIP-based video learners, enhancing performance across parameter-efficient fine-tuned, partially-tuned, and fully-tuned video learners. This flexibility significantly broadens its utility, making it a versatile tool for tasks requiring robust generalization without extensive domain-specific tailoring. 4) Extensive experiments demonstrate the state-of-the-art results achieved by our Open-MeDe, highlighting its role in advancing general video understanding. Our framework can empower many downstream applications, such as video-based surveillance and security, autonomous vehicles, human-computer interaction, *etc*.

Limitations. Despite achieving promising open-vocabulary generalization with our framework, the out-of-context scenarios remain challenging and constrained by the reliance on temporal and static feature alignment. Specifically, scenarios with extreme domain shifts (*e.g.*, SCUBA and HAT benchmarks) show significant performance gaps. However, the residual influence of static visual cues persists, particularly in complex video backgrounds and more compact foregrounds. Incorporating stronger, explicitly targeted debiasing strategies, such as adversarial learning or counterfactual data augmentation, may further enhance robustness, which will be explored in our future work.

References

[1] 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. 4

[2] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 1

[3] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[4] 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. 1

[5] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *European Conference on Computer Vision*, pages 88–105. Springer, 2022. 1

[6] Shuangrui Ding, Maomao Li, Tianyu Yang, Rui Qian, Haohang Xu, Qingyi Chen, Jue Wang, and Hongkai Xiong. Motion-aware contrastive video representation learning via foreground-background merging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9716–9726, 2022. 4

[7] 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. 1

[8] 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

[9] 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

[10] 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. 1

[11] 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, 4

[12] 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. 2

[13] 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

[14] 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

[15] Jinpeng Wang, Yuting Gao, Ke Li, Yiqi Lin, Andy J Ma, Hao Cheng, Pai Peng, Feiyue Huang, Rongrong Ji, and Xing Sun. Removing the background by adding the background: Towards background robust self-supervised video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11804–11813, 2021. 4

[16] 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

[17] 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. 5

[18] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 1

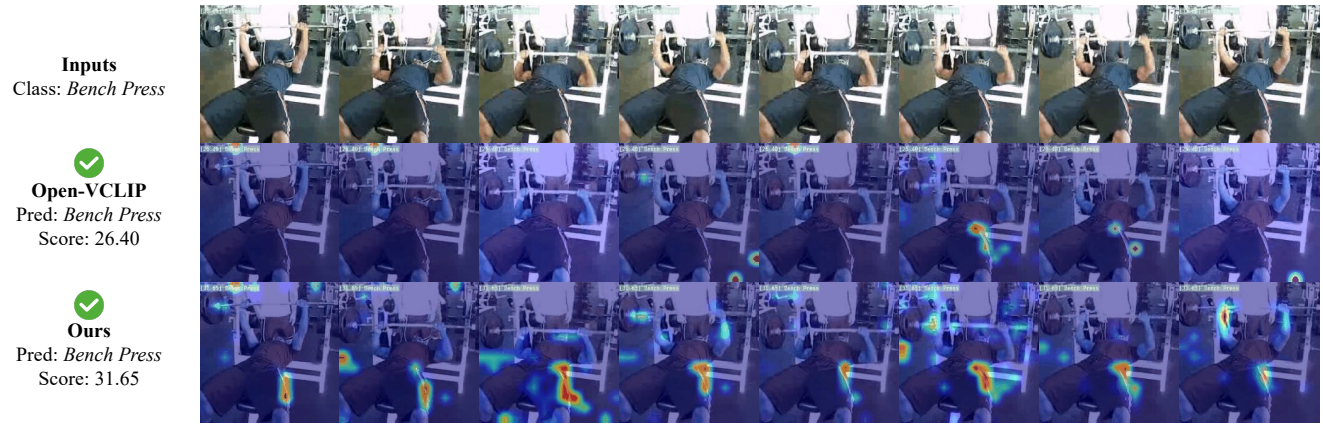

Figure 2. Visualizations of attention maps and predictions for "*Bench Press*" in the in-context setting. Both Open-VCLIP and our proposed framework correctly predict the action, while ours achieves a higher score. Additionally, our framework demonstrates enhanced attention to the key elements associated with the action, which highlights its effectiveness in capturing nuanced and discriminative features, leading to more confident predictions.

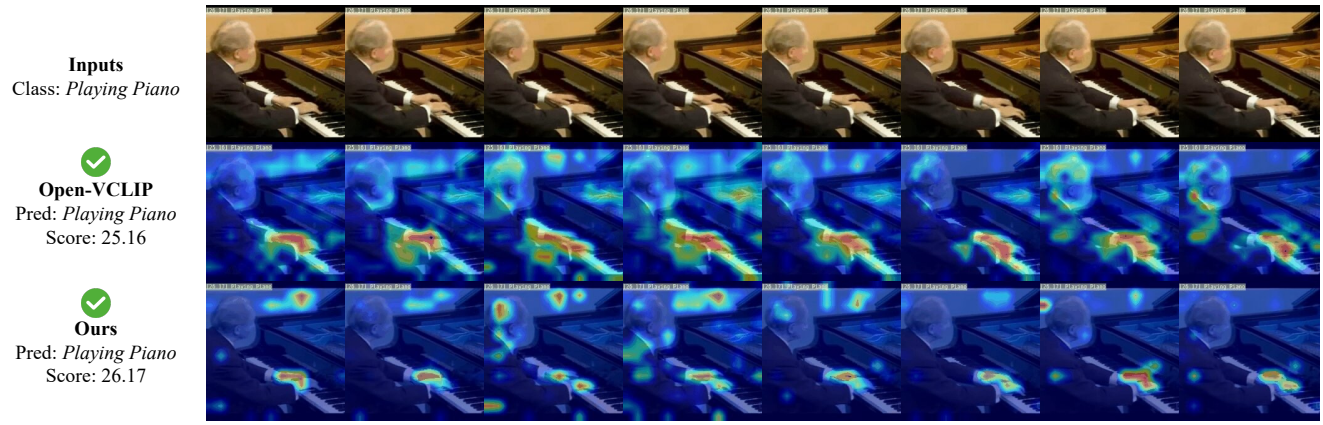

Figure 3. Visualizations of attention maps and predictions for "*Playing Piano*" in the in-context setting. Our method emphasizes the subtle movements of the action rather than redundant visual appearances, demonstrating its effectiveness of capturing critical motion cues.

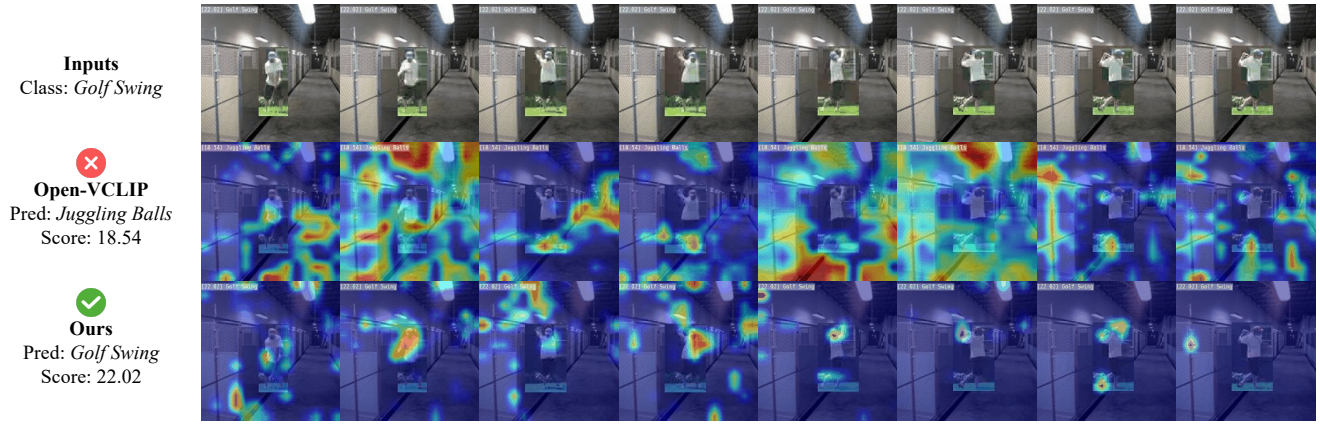

Figure 4. Visualizations of attention maps and predictions for "*Golf Swing*" in the out-of-context setting. Our method successfully classifies the action and effectively captures key visual cues associated with the motion, demonstrating the improved robustness. However, Open-VCLIP misclassifies the action as "*Juggling Balls*" due to its large static bias.

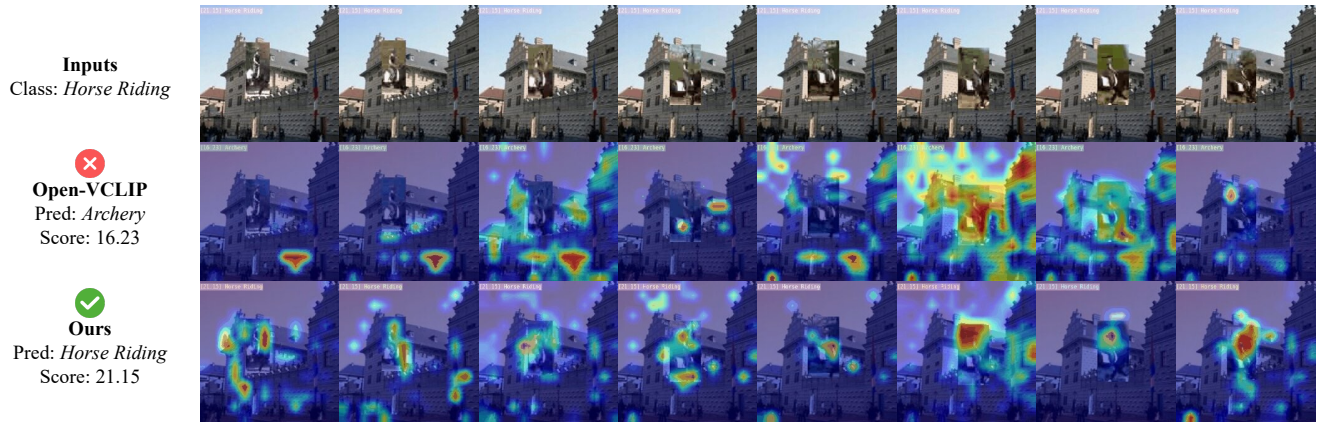

Figure 5. Visualizations of attention maps and predictions for "*Horse Riding*" in the out-of-context setting. Our method outperforms Open-VCLIP by accurately attending to critical dynamic information specific to the true action, showcasing its robustness and reliability in discerning action-relevant features under challenging out-of-context scenarios.

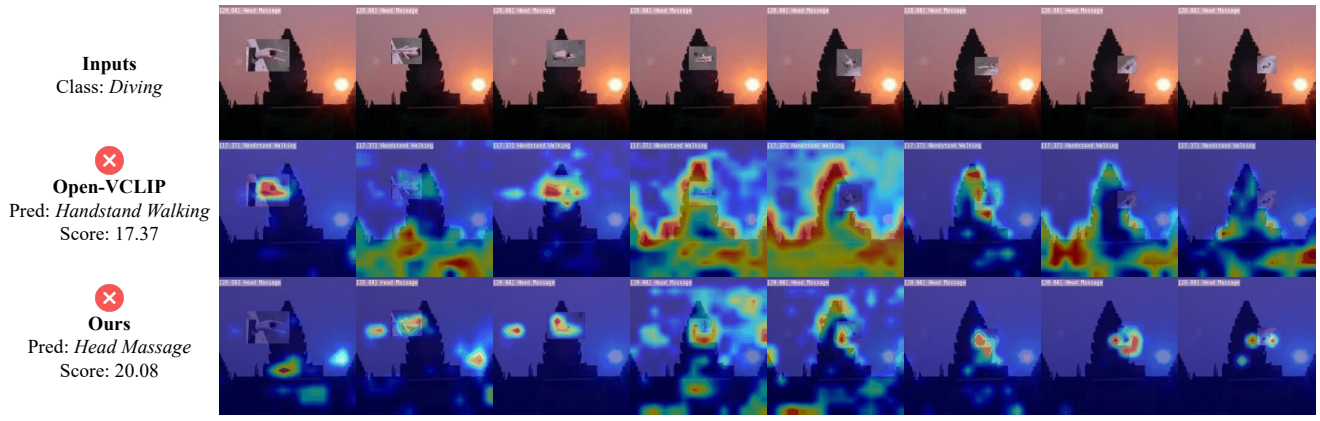

Figure 6. Visualizations of attention maps and predictions for "*Diving*" in the out-of-context setting. Both methods struggle to classify the action correctly, suggesting more room for improvement under this challenging scenario. Despite the incorrect prediction, our method reflects a better focus on motion-relevant areas, which indicates its effectiveness of mitigating static bias.