# Robust Zero-Shot Generalization for Open-Vocabulary Action Recognition via Task Arithmetic

Francesca Morandi*[1,2], Omayma Moussadek*[1], Federico Venturini[3], Mauro Suardi[3], Alessandro Banzatti[3], Francesco Cannarile[3], Angelo Porrello[1], Simone Calderara[1]

[1]University of Modena and Reggio Emilia
[1]{name.surname}@unimore.it

[2]University of Pisa
[2]{name.surname}@phd.unipi.it

[3]Eni S.p.A.
[3]{name.surname}@eni.com

*Equal contribution.

***Abstract*—Open Vocabulary Action Recognition (OVAR) enables the recognition of novel actions by leveraging vision–language representations, overcoming the limitations of traditional closed-set approaches. However, achieving robust performance in real-world scenarios typically requires domain-specific fine-tuning, which is often costly and raises privacy and regulatory concerns. In this work, we propose an alternative paradigm that bypasses target-domain training and recombines knowledge from existing datasets and models. Leveraging model merging and task arithmetic, we extract and combine task vectors from models fine-tuned on diverse public OVAR datasets. We show that, in out-of-distribution settings, the resulting merged model achieves superior zero-shot generalization to the pre-trained base model. Code is available at https://github.com/omaymaMoussadek/robust-ovar**

## I. Introduction

Action recognition is a fundamental task in computer vision that aims to identify human actions from video data, with applications in surveillance, human-computer interaction, autonomous driving, and video retrieval. Traditionally, it is formulated as a *closed-set classification problem*, where models are trained on a fixed set of predefined action classes and cannot generalize to unseen ones. In contrast, **Open Vocabulary Action Recognition (OVAR)** leverages multimodal vision-language representations [1] to align videos with textual descriptions, enabling the recognition of novel actions at inference time by matching visual inputs with semantic concepts expressed in language. This multimodal approach effectively shifts the paradigm from a closed-set to an *open-dictionary setting*, where actions can be specified through structured prompts that encode richer information than simple predicates, while also allowing the recognition of actions not observed during training, thus providing greater flexibility.

Although the **zero-shot** capabilities of OVAR models theoretically allow them to operate in new domains without further training, achieving robust performance in real-world scenarios often requires a **fine-tuning phase** on domain-specific data. This is particularly evident in complex and heterogeneous environments characterized by poor lighting conditions, multiple interacting actors and objects, and distant or low-resolution camera views. In these contexts, a reliable recognition system requires data-driven adaptation.

This requirement poses a significant challenge, as collecting and annotating video data for action recognition and video analytics is inherently **sensitive** and **difficult**. In many applications, data acquisition involves continuous monitoring of individuals in physical spaces, raising **serious privacy concerns**. Moreover, such systems fall under strict regulatory frameworks: for instance, the AI Act classifies many of these applications as **high-risk**, due to their reliance on biometric data, video surveillance, and the potential for behavioral profiling. As a result, acquiring in-domain data for fine-tuning is often not only costly, but also legally and ethically constrained.

To overcome the intractability of standard fine-tuning for OVAR, we explore an alternative direction: instead of adapting models to the target domain, we recombine and fuse the knowledge already embedded in existing, safely trained models. This strategy aligns with a broader trend in artificial intelligence based on the reuse and combination of existing models, that has become particularly relevant in fields such as image classification and natural language processing [2], [3], where the increasing scale and complexity of deep architectures makes fine-tuning progressively impractical.

Central to this line of research is task arithmetic [4], [5], which enables model editing and merging through simple algebraic operations in parameter space. This approach relies on the concept of **task vector**, formally defined as the exact displacement in parameter space between a pre-trained base model and its task-specific fine-tuned counterpart. As demonstrated in recent literature [6]–[8], these vectors isolate learned knowledge so it can be directly manipulated: subtracting them enables the selective removal of behaviors, while summing them allows models sharing the same initialization to be merged into a single multi-task network (as shown in Fig. 1).

The main objective of our work is to deploy a modular and scalable approach for open-vocabulary action recognition that operates entirely without target-domain fine-tuning. By leveraging task arithmetic and advanced model merging techniques, we achieve robust zero-shot generalization, main-

taining high performance even in out-of-distribution (OOD) settings characterized by severe distribution shifts with respect to the pre-trained model. To achieve this, we leverage a collection of fine-tuned models trained on diverse, publicly available datasets [9]–[12], from which we extract the corresponding task vectors. These task vectors are then combined through addition, resulting in a unified multi-task model that aggregates knowledge from multiple sources. In this way, the model acquires more refined and diverse action understanding capabilities, improving generalization across a broader range of action classes without further in-domain training.

While the effectiveness of model merging techniques has been widely demonstrated in Large Language Models (LLMs) [6], [13], image classification [4], and medical imaging [14], their evaluation in the video domain remains unexplored. This leaves a significant gap in understanding the impact of subspace merging and task arithmetic on the complex spatiotemporal representations required for open-vocabulary action recognition. We show that, even in the absence of a specialized target dataset, it is possible to successfully merge models fine-tuned on diverse, publicly available OVAR datasets. Crucially, the resulting models outperform the original zero-shot baseline, even when the source datasets are entirely out-of-distribution with respect to the target application, mitigating the need for sensitive, domain-specific data collection. To sum up, our main contributions are:

1) **Novel application to video:** We present the first systematic evaluation of existing model merging and task arithmetic techniques (including Task Arithmetic (TA) [4], TSV-M [8], and Iso-C [15]) applied specifically to Open Vocabulary Action Recognition.
2) **Privacy-preserving OOD adaptation:** We provide empirical evidence that parameter-space merging of models fine-tuned on public datasets reliably yields constructive interference and generalized performance gains in the video domain, even under out-of-distribution shifts. By achieving effective target-domain adaptation entirely through task vector aggregation, our approach completely eliminates the need to collect or train on sensitive, application-specific video data.

## II. Background

### A. Open-Vocabulary Video Recognition

Contrastive vision-language models like CLIP [1] exhibit strong zero-shot classification ability by aligning visual and textual representations within a shared embedding space. This allows the model to recognize categories beyond the training set by matching visual features with text descriptions.

However, extending this paradigm from images to videos is not straightforward, as action recognition requires modeling temporal dynamics in addition to visual semantics. **Open-VCLIP** [16] addresses this limitation by adapting CLIP to video recognition while preserving its open-vocabulary nature. A key aspect of Open-VCLIP is the introduction of lightweight temporal modeling directly inside the visual transformer.

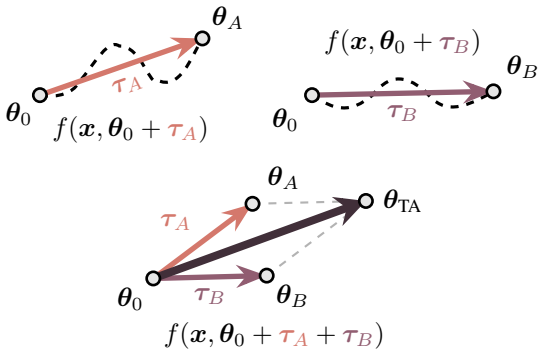


Fig. 1. Given two models fine-tuned from the same weights, task addition [4] produces a multi-task model by summing their corresponding task vectors.

Instead of processing each frame independently, the model expands the self-attention operation so that visual patches can attend not only to patches from the same frame, but also to those in adjacent frames, including both preceding and subsequent ones. In this way, CLIP's original spatial attention is transformed into a spatio-temporal mechanism able to capture short-range temporal dependencies without adding extra parameters.

### B. Model Soups and Task Arithmetic

In recent years, model soups [2] have shown that averaging the weights of fine-tuned models originating from the same pre-trained initialization can improve both accuracy and robustness without increasing inference cost. This approach was originally studied in a setting where multiple models are fine-tuned on the same downstream task and dataset, differing only in training hyperparameters such as learning rate, data order, or regularization. Empirically, weight averaging in this setting yields a model that matches in-domain performance while significantly improving robustness under distribution shifts.

Task arithmetic [4] represents updates as the parameter difference between a fine-tuned model and its pre-trained initialization. These parameter differences, called task vectors, can be manipulated through simple operations such as addition, subtraction, and scaling, enabling controlled modifications of model behavior without retraining from scratch.

A key difference with respect to model soups is that, in task arithmetic, each model is fine-tuned on a different dataset or objective, and the resulting task vectors encode distinct capabilities. Prior work [5], [17], [18] has shown that, under a linearized regime, *i.e.*, when models are approximated via a first-order Taylor expansion around the pre-trained parameters, task vectors exhibit structured compositionality. Complementarily, recent second-order analyses of model compositionality suggest that composable task-specific modules can also emerge in standard non-linear networks, provided that fine-tuning remains within the pre-training basin [19]. Related work [20] bridges linearized and standard non-linear fine-tuning by distilling the representations of a linearized teacher into a non-linear student.

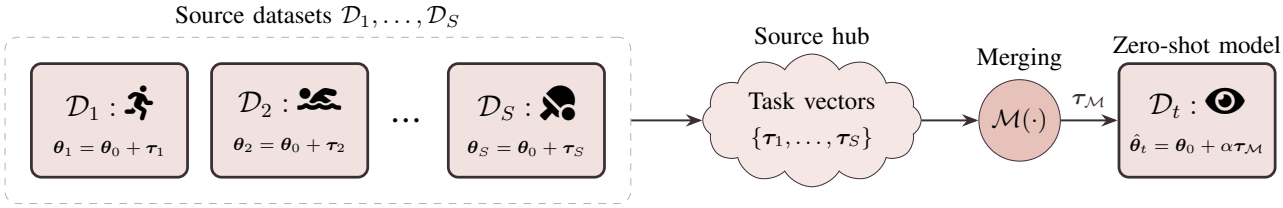


Fig. 2. Experimental protocol for zero-shot evaluation. Source models are fine-tuned on individual datasets $\mathcal{D}_1, \ldots, \mathcal{D}_S$. Task vectors are aggregated in the *Source Hub*, merged, and then evaluated on the held-out target $D_t$.

In this paper, we build on these insights and employ task arithmetic to fuse knowledge from multiple datasets by combining their task vectors. Compared to centralized retraining, this approach offers a more scalable and modular alternative, enabling incremental integration of new capabilities without access to all training data or full model retraining.

### C. Advanced Merging Strategies

While task vector addition offers a simple and efficient way to combine models, its effectiveness can be limited when the source models are specialized for different domains or tasks. In such cases, interactions between parameter updates may attenuate useful task-specific directions. Prior work shows that this can hinder the preservation of performance achieved by individually trained models, as conflicting updates may lead to degradation across tasks [8], [15]. To overcome this issue, more advanced model merging strategies leverage the internal structure of weight updates instead of merging the network as a single flat vector. Among them, **Task Singular Vector (TSV-M)** [8] and **Iso-C** [15] tackle the problem from two complementary perspectives:

- **TSV-M** [8] performs a structured merge based on singular value decomposition. For each matrix-shaped parameter, it analyzes task updates at the layer level, extracts the most informative low-rank directions, and recombines them into a merged parameter update. Non-matrix parameters are merged by direct averaging.
- **Iso-C** [15] starts from the average task vector and then applies an isotropization step, replacing the original singular value spectrum with a flatter one. This reduces dominant directions that may bias the merged updates toward a subset of source tasks, leading to a more balanced combination of shared and task-specific information.

Driven by these findings, we analyze whether advanced merging strategies, such as TSV-M and Iso-C, can improve performance in out-of-domain settings, *i.e.*, when transferring to datasets that differ from those used during fine-tuning.

## III. Task Arithmetic for Robust Open-VOC Action Recognition

Given a publicly available dataset $\mathcal{D} = \{(V_i, T_i)\}_{i=1}^N$, where each video clip is paired with a textual description, we initialize the model from a pre-trained Open-VCLIP checkpoint, denoted by $f_{\boldsymbol{\theta}_0}$ with parameters $\boldsymbol{\theta}_0$ [16], and fine-tune it to learn a shared embedding space in which each video is aligned with its corresponding text. Let $f_{\boldsymbol{\theta}^{(V)}}(\cdot)$ and $f_{\boldsymbol{\theta}^{(T)}}(\cdot)$ denote the visual and text encoders, respectively. For a video-text pair $(V, T)$, the corresponding embeddings are

$$v = f_{\boldsymbol{\theta}^{(V)}}(V), \qquad t = f_{\boldsymbol{\theta}^{(T)}}(T), \tag{1}$$

and their alignment is encouraged by maximizing the inner-product similarity $\mathrm{sim}(v, t) = \langle v, t \rangle$. During fine-tuning, the text encoder is kept frozen [4]; therefore, the optimization primarily adapts the visual encoder. At inference time, class predictions are obtained by matching visual video representations against text-derived class prototypes built from a fixed set of CLIP-style prompt templates.

After fine-tuning on a source dataset $\mathcal{D}$, we extract the **task vector** of the visual encoder as

$$\boldsymbol{\tau} = \boldsymbol{\theta}^{(V)} - \boldsymbol{\theta}_0^{(V)}, \quad \text{where } \boldsymbol{\theta}^{(V)} = \mathcal{A}(\boldsymbol{\theta}_0^{(V)}, \mathcal{D}), \tag{2}$$

where $\boldsymbol{\theta}_0$ denotes the pre-trained model (*initialization*) and $\mathcal{A}(\cdot)$ is a gradient-based adaptation procedure.

### A. Source Model Adaptation

A common limitation of standard fine-tuning is the lack of access to target-domain training data, often due to privacy constraints or data-sharing restrictions. A natural alternative is to leverage existing datasets from related domains. In principle, multiple publicly available action recognition datasets could be combined into a single large training set for joint fine-tuning.

However, this approach is often impractical in real-world scenarios, since datasets typically differ in data formats, annotation protocols, class vocabularies, video lengths, frame sampling strategies, and dataset-specific preprocessing pipelines (*e.g.*, data loaders, augmentations, and normalization schemes). Harmonizing these aspects into a unified training framework requires engineering effort and may still result in suboptimal performance due to domain inconsistencies.

To address these challenges, we adopt a **modular strategy**. We assume access to a collection of *source* action recognition datasets $\{\mathcal{D}_s\}_{s=1}^S$, while the *target* dataset $\mathcal{D}_t$ is not available for training. Instead of merging all datasets at the data level, we independently fine-tune the same pre-trained model $\boldsymbol{\theta}_0$ on each source dataset $\mathcal{D}_s$, obtaining a set of source-specific parameters $\{\boldsymbol{\theta}_s\}_{s=1}^S$. Each of these models captures domain-specific adaptations, which we represent through the corresponding hub of task vectors $\mathcal{T}_{\texttt{Source-Hub}}$:

$$\mathcal{T}_{\texttt{Source-Hub}} = \{\boldsymbol{\tau}_s \,|\, \boldsymbol{\tau}_s = \boldsymbol{\theta}_s^{(V)} - \boldsymbol{\theta}_0^{(V)}\}_{s=1}^S \tag{3}$$

TABLE I
OVERVIEW OF THE DATASETS USED IN OUR EXPERIMENTS.

| Dataset | #Classes | #Videos | Description |
|---|---|---|---|
| K700 [9] | 700 | 650,317 | Daily activities, YouTube clips |
| UCF101 [10] | 101 | 13320 | Sports and human actions |
| HMDB51 [11] | 51 | 6766 | Cinematic and facial motions |
| XD-Violence [12] | 7 | 4754 | Audio-visual violent events |

A schematic overview of the proposed method is provided in Fig. 2, which illustrates the extraction of task vectors from source-specific models, their subsequent merging, and zero-shot evaluation on the held-out target dataset. The extracted source task vectors are then combined into a single update, as described in the following subsection.

### B. Task-Vector Merging

Given the set of source task vectors $\mathcal{T}_{\texttt{Source-Hub}} = \{\boldsymbol{\tau}_s\}_{s=1}^{S}$, we define a merging function $\mathcal{M}(\cdot)$ that aggregates them into a single update vector $\boldsymbol{\tau}_{\mathcal{M}}$. As a baseline, we consider standard Task Arithmetic (TA), i.e.,

$$\boldsymbol{\tau}_{\mathcal{M}} = \mathcal{M}\big(\mathcal{T}_{\texttt{Source-Hub}}\big), \qquad \textit{e.g., } \boldsymbol{\tau}_{\mathcal{M}} = \sum_{s=1}^{S} \boldsymbol{\tau}_s \tag{4}$$

In addition to TA, we assess the advanced merging strategies Iso-C [15] and TSV-M [8] introduced in Sec. II-C. The resulting merged update is used to construct the target model as $\hat{\boldsymbol{\theta}}_t = \boldsymbol{\theta}_0 + \alpha\, \boldsymbol{\tau}_{\mathcal{M}}$ where $\alpha$ is a scaling coefficient controlling the contribution of the merged task vector to the base model.

## IV. EXPERIMENTS

In the following, we evaluate the ability of different merging strategies (TA, Iso-C, TSV-M) to generalize in an out-of-distribution setting, *i.e.*, on a held-out external dataset, while also assessing their capacity to preserve in-domain performance, *i.e.*, on the source datasets used during adaptation.

### A. Experimental setting

**Backbones.** For all experiments, we adopt Open-VCLIP [16] as the base model and consider two visual backbones, ViT-B/16 and ViT-L/14. Both architectures are initially adapted to the video domain via pre-training on Kinetics-400 (K400) [21]. We employ the official public checkpoints provided by the Open-VCLIP authors [16] and maintain their standard training and inference protocols across all experiments.

**Datasets.** We evaluate our approach on four action recognition benchmarks. Kinetics-700 (K700) [9] is a large-scale collection of curated YouTube clips covering a broad range of daily activities and social interactions. UCF101 [10] features realistic videos across five macro-groups, including human and object interactions and sports in unconstrained environments. HMDB51 [11] is a cinematic dataset composed of clips extracted from movies and public databases, focusing on facial and body movements. Lastly, XD-Violence [12] is a massive multi-modal benchmark for anomaly detection, capturing complex violent events through both RGB and audio signals. Table I summarizes their main characteristics.

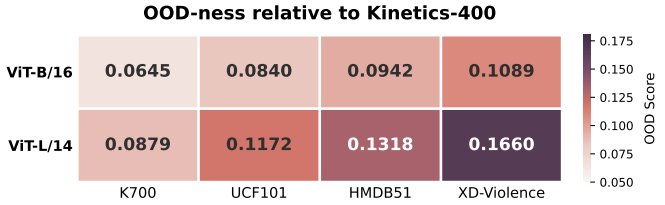


Fig. 3. For each dataset, we report its out-of-distribution (OOD) shift relative to K400, used to pre-train the base model. Larger values (*e.g.*, for XD-Violence) indicate a stronger semantic mismatch with the source domain.

**Evaluation Protocol.** To rigorously assess zero-shot generalization, we evaluate all merging methods under a **leave-one-dataset-out protocol** across HMDB51, UCF101, K700, and XD-Violence. In each of the four iterations, one dataset is held out as the target (test) domain. The merged model is then constructed by extracting and aggregating task vectors from the models fine-tuned on the remaining three source datasets. Crucially, the training split of the target dataset is never observed during this process, ensuring a pure out-of-distribution evaluation.

**Implementation Details.** All models are optimized using AdamW [22] with a fixed learning rate of $3.33 \times 10^{-6}$. The performance of the merged model is governed by a scaling factor $\alpha$, which modulates the contribution of the task-specific vectors. Following the original setups of [4], [5], we determine the optimal $\alpha$ by performing a grid search in the range $[0.1, 2.0]$ with a step size of $0.1$, selecting the value that yields the highest accuracy on the target dataset's validation split.

**Metrics.** To assess overall performance, we report Top-1 and Top-5 accuracy of the models on the target dataset.

### B. Quantifying Distribution Shifts

To evaluate the ability of different merging strategies to generalize in an out-of-distribution (OOD) setting, we must first quantitatively assess the semantic gap exhibited by each target dataset with respect to the base model. We achieve this by adopting a text-based metric derived from the class labels of the source domain (*i.e.*, K400) and those of the target domain.

Specifically, we extract the embeddings of the class labels of K400 and each target dataset using the CLIP text encoder, and compute all pairwise cosine similarities between the corresponding textual embeddings. For each target label, we retain the maximum similarity to any K400 label and average these values over the entire set of target labels. We then define the OOD score as one minus this mean maximum similarity, such that higher values indicate a larger semantic mismatch with respect to K400:

$$\mathrm{OOD}(\mathcal{T}) = 1 - \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \max_{c \in \mathcal{C}} \ \mathrm{sim}\big(f_{\boldsymbol{\theta}^{(\mathcal{T})}}(t), f_{\boldsymbol{\theta}^{(\mathcal{T})}}(c)\big) \tag{5}$$

where $\mathcal{T}$ denotes the set of class labels in the target dataset, $\mathcal{C}$ the set of class labels in K400 and $\mathrm{sim}\big(f_{\boldsymbol{\theta}^{(\mathcal{T})}}(t), f_{\boldsymbol{\theta}^{(\mathcal{T})}}(c)\big)$ the cosine similarity between the textual embedding of target label $t$ and that of source label $c$.

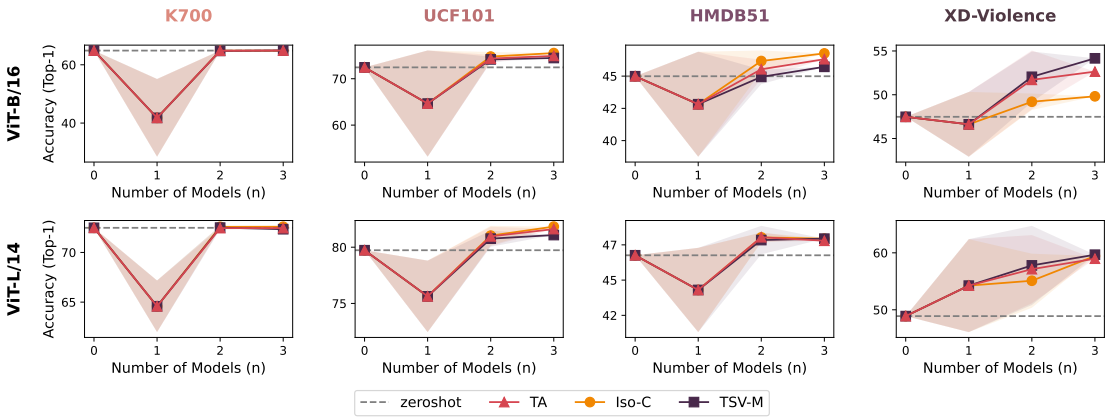


Fig. 4. Target accuracy versus number of fused source models, showing how performance scales as more task vectors are aggregated.

TABLE II
RESULTS UNDER THE LEAVE-ONE-DATASET-OUT PROTOCOL. FOR EACH DATASET, MERGING STRATEGIES ARE EVALUATED USING TASK-SPECIFIC CHECKPOINTS FROM THE REMAINING THREE SOURCE DOMAINS.

| | | K700 | | UCF101 | | HMDB51 | | XD-Violence | |
|---|---|---|---|---|---|---|---|---|---|
| Model | Strategy | Top-1 | Top-5 | Top-1 | Top-5 | Top-1 | Top-5 | Top-1 | Top-5 |
| **ViT-B/16** | | | | | | | | | |
| Fine-tuned | | 78.38 | 93.76 | 99.47 | 99.92 | 80.97 | 93.81 | 83.35 | 99.18 |
| Zero-shot | | 64.87 | 86.37 | 72.45 | 89.56 | 44.99 | 69.76 | 47.48 | 91.44 |
| Merged | TA | 64.89 | **86.19** | 74.92 | 89.79 | 46.31 | 69.32 | 52.64 | 94.72 |
| | Iso-C | 64.85 | 86.08 | **75.53** | **90.09** | **46.76** | 69.76 | 49.82 | 92.61 |
| | TSV-M | **64.91** | 86.00 | 74.47 | 89.94 | 45.72 | **70.80** | **54.16** | **94.73** |
| **ViT-L/14** | | | | | | | | | |
| Fine-tuned | | 82.77 | 95.47 | 99.77 | 100.00 | 83.92 | 95.87 | 84.64 | 99.41 |
| Zero-shot | | 72.48 | 90.18 | 79.73 | 93.60 | 46.76 | 71.98 | 48.89 | 92.03 |
| Merged | TA | 72.45 | **90.31** | 81.61 | 93.32 | 47.79 | 72.86 | 58.97 | 94.61 |
| | Iso-C | **72.59** | 90.26 | **81.83** | 93.62 | **47.94** | 73.01 | 59.44 | **96.48** |
| | TSV-M | 72.50 | 90.27 | 81.08 | **93.84** | **47.94** | **73.60** | **59.67** | 95.43 |

We report the OOD scores of each target dataset in Fig. 3 for both ViT-B/16 and ViT-L/14. Overall, K700 is the closest dataset to K400, whereas **XD-Violence is the most distant**. In contrast, HMDB51 and UCF101 remain relatively close to K400, indicating a broader semantic alignment. We leverage these observations to contextualize the following results.

### *C. Zero-Shot Generalization in Open-Voc Action Recognition*

To understand how varying degrees of distribution shift impact actual performance, we evaluate our approach under the leave-one-dataset-out protocol detailed in Sec. IV-A. By iteratively holding out each dataset as the unseen target, we analyze how performance scales when merging different combinations of 1, 2, and 3 source models fine-tuned on the remaining domains, reporting average accuracy across ensembles. As illustrated in Fig. 4, increasing the pool of merged models positively correlates with performance on OOD distributions, yielding results consistently superior to those of the original zero-shot model. This suggests that merging multiple models integrates complementary knowledge, a factor that is beneficial for overcoming the base model's limitations.

TABLE III
HYPERPARAMETER-DIVERSE POOLS (XD-VIOLENCE): "MERGED" AND "ALL-8 MERGED" FUSE 1 AND 8 MODELS PER DATASET, RESPECTIVELY.

| | | ViT-B/16 | | ViT-L/14 | |
|---|---|---|---|---|---|
| Model | Strategy | Top-1 | Top-5 | Top-1 | Top-5 |
| Fine-tuned | | 83.35 | 99.18 | 84.64 | 99.41 |
| Zero-shot | | 47.48 | 91.44 | 48.89 | 92.03 |
| Merged | TA | 52.64 | 94.72 | 58.97 | 94.61 |
| | Iso-C | 49.82 | 92.61 | 59.44 | **96.48** |
| | TSV-M | **54.16** | **94.73** | **59.67** | 95.43 |
| ALL-8 Merged | TA | 53.81 | 94.49 | 64.24 | 96.72 |
| | Iso-C | 49.82 | 93.02 | 65.53 | **97.30** |
| | TSV-M | **54.75** | **95.43** | **66.59** | 96.13 |

Furthermore, performance gains appear to be correlated with the "OOD-ness" of the dataset: the more distant the downstream task is from the pre-trained model (*e.g.*, XD-Violence), the greater the benefits of increasing the number of merged models. Conversely, when the task is closely related to the pre-training domain, the improvements are negligible, as the base model already exhibits strong inherent proficiency.

Next, we evaluate the performance obtained when utilizing the maximum number of source models per target dataset, specifically by merging the checkpoints fine-tuned on all the remaining datasets. As presented in Tab. II, more complex fusion strategies, such as Iso-C and TSV-M, provide tangible benefits when the target dataset is distant from the pre-trained model, whereas they perform comparably to simple task vector summation when the knowledge gap is minimal. Furthermore, the efficacy of model merging remains consistent across different backbones (ViT-B/16 and ViT-L/14).

**Leveraging Hyperparameter-Diverse Model Pools.** Inspired by prior work on model soups [2], which shows that averaging checkpoints obtained from different hyperparameter configurations can improve robustness, we extend this idea to a multi-task merging setting to further expand knowledge diversity. Instead of relying on a single fine-tuned model per non-held-out source dataset under the same leave-one-dataset-out protocol, we generate **eight distinct source checkpoints** per

TABLE IV
COMPARISON UNDER STANDARD MERGING EVALUATION PROTOCOL.

| Model | Strategy | K700 | | UCF101 | | HMDB51 | | XD-Violence | |
|---|---|---|---|---|---|---|---|---|---|
| | | Top-1 | Top-5 | Top-1 | Top-5 | Top-1 | Top-5 | Top-1 | Top-5 |
| **ViT-B/16** | | | | | | | | | |
| Fine-tuned | | 78.38 | 93.76 | 99.47 | 99.92 | 80.97 | 93.81 | 83.35 | 99.18 |
| Zero-shot | | 64.87 | 86.37 | 72.45 | 89.56 | 44.99 | 69.76 | 47.48 | 91.44 |
| Merged | TA | 67.01 | 88.03 | 99.55 | 100.00 | 71.39 | 91.30 | 78.43 | 98.59 |
| | Iso-C | **68.96** | **89.87** | **99.77** | **100.00** | 72.42 | 91.15 | 75.62 | 99.06 |
| | TSV-M | 67.28 | 88.50 | **99.77** | **100.00** | **74.48** | **91.59** | **80.19** | **99.06** |
| **ViT-L/14** | | | | | | | | | |
| Fine-tuned | | 82.77 | 95.47 | 99.77 | 100.00 | 83.92 | 95.87 | 84.64 | 99.41 |
| Zero-shot | | 72.48 | 90.18 | 79.73 | 93.60 | 46.76 | 71.98 | 48.89 | 92.03 |
| Merged | TA | 79.45 | 94.60 | 98.05 | 99.85 | 71.09 | 91.30 | **84.29** | 99.18 |
| | Iso-C | **80.13** | **95.18** | 99.02 | 99.92 | 73.89 | 92.04 | 83.59 | **99.41** |
| | TSV-M | 80.11 | 94.85 | **99.77** | **100.00** | **74.50** | **93.36** | 84.06 | 99.06 |

dataset to be used for subsequent merging. These checkpoints are obtained by varying fine-tuning hyperparameters, including learning rate, weight decay, and data augmentation.

The resulting checkpoints are first merged within each dataset via simple averaging, yielding a single consolidated checkpoint per dataset. These consolidated checkpoints are then combined using TA, Iso-C, and TSV-M, as in the previous subsection. We evaluate this two-level approach on XD-Violence, using the validation split to tune the scaling factor $\alpha$ and the test split to report the final results. As shown in Tab. III, increasing source diversity at the hyperparameter level produces more transferable merged models under distribution shifts, yielding substantial gains over the single-model merging approach discussed previously. This further corroborates our intuition that leveraging model diversity is essential for composing a model that is robust against distribution shifts.

### D. Model Merging Performance

Finally, we evaluate the performance of the merged models on the union of the source datasets (using their respective test sets). This is consistent with standard evaluation practices in model merging literature [4], where the objective is to assess whether the fused model preserves the knowledge embedded in each fine-tuned model comprising the ensemble.

We report the results in Tab. IV. As shown, except for UCF101, the performance of the merged model is generally inferior to that of individual fine-tuning. Nevertheless, the application of more complex merging strategies is generally rewarding, suggesting that there remains scope for improving merging strategies for open-vocabulary action recognition.

## V. CONCLUSION

We investigate the application of task arithmetic to open-vocabulary action recognition as a means to address performance degradation under distribution shifts. We demonstrate that the efficacy of these merging strategies is positively correlated with the semantic distance (OOD-ness) of the target task, with advanced fusion techniques providing tangible benefits. Furthermore, we increase model diversity by extending this approach to pools of models trained with different hyperparameters, showing that such diversity enhances model transferability and robustness. Collectively, our results suggest that model merging offers a promising paradigm for efficient, adaptable, and generalizable action recognition systems.

## REFERENCES


[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[2] M. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. Gontijo-Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith *et al.*, "Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time," in *ICML*, 2022.

[3] M. S. Matena and C. A. Raffel, "Merging models with fisher-weighted averaging," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17 703–17 716, 2022.

[4] G. Ilharco, M. T. Ribeiro, M. Wortsman, S. Gururangan, L. Schmidt, H. Hajishirzi, and A. Farhadi, "Editing models with task arithmetic," in *ICLR*, 2022.

[5] G. Ortiz-Jimenez, A. Favero, and P. Frossard, "Task arithmetic in the tangent space: Improved editing of pre-trained models," *NeurIPS*, 2023.

[6] P. Yadav, D. Tam, L. Choshen, C. A. Raffel, and M. Bansal, "Ties-merging: Resolving interference when merging models," *Advances in neural information processing systems*, vol. 36, pp. 7093–7115, 2023.

[7] K. Wang, N. Dimitriadis, G. Ortiz-Jimenez, F. Fleuret, and P. Frossard, "Localizing task information for improved model merging and compression," *arXiv preprint arXiv:2405.07813*, 2024.

[8] A. A. Gargiulo, D. Crisostomi, M. S. Bucarelli, S. Scardapane, F. Silvestri, and E. Rodola, "Task singular vectors: Reducing task interference in model merging," in *CVPR*, 2025.

[9] J. Carreira, E. Noland, C. Hillier, and A. Zisserman, "A short note on the kinetics-700 human action dataset," *arXiv preprint arXiv:1907.06987*, 2019.

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

[11] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: A large video database for human motion recognition," in *ICCV*, 2011.

[12] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang, "Not only look, but also listen: Learning multimodal violence detection under weak supervision," in *ECCV*. Springer, 2020.

[13] L. Yu, B. Yu, H. Yu, F. Huang, and Y. Li, "Language models are super mario: Absorbing abilities from homologous models as a free lunch," in *Forty-first International Conference on Machine Learning*, 2024.

[14] L. Lumetti, G. Capitani, E. Ficarra, S. Calderara, C. Grana, A. Porrello, and F. Bolelli, "U-net transplant: the role of pre-training for model merging in 3d medical segmentation," in *MICCAI*, 2025.

[15] D. Marczak, S. Magistri, S. Cygert, B. Twardowski, A. D. Bagdanov, and J. Van De Weijer, "No task left behind: Isotropic model merging with common and task-specific subspaces," in *ICML*, 2025.

[16] Z. Weng, X. Yang, A. Li, Z. Wu, and Y.-G. Jiang, "Open-vclip: Transforming clip to an open-vocabulary video model via interpolated weight optimization," in *ICML*, 2023.

[17] K. Yoshida, Y. Naraki, T. Horie, R. Yamaki, R. Shimizu, Y. Saito, J. McAuley, and H. Naganuma, "Mastering task arithmetic: $\tau$jp as a key indicator for weight disentanglement," in *ICLR*, 2025.

[18] A. Porrello, P. Buzzega, F. Dangel, T. Sommariva, R. Salami, L. Bonicelli, and S. Calderara, "Dataless weight disentanglement in task arithmetic via kronecker-factored approximate curvature," in *ICLR*, 2026.

[19] A. Porrello, L. Bonicelli, P. Buzzega, M. Millunzi, S. Calderara, and R. Cucchiara, "A second-order perspective on model compositionality and incremental learning," in *ICLR*, vol. 2025, 2025.

[20] T. Sommariva, F. Morandi, S. Calderara, and A. Porrello, "Distilling linearized behavior into non-linear fine-tuning for effective task arithmetic," in *ICML*, 2026.

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

[22] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019.