Protect to Adapt: Orthogonal Subspace Control with Ranked Negative-Prompt Curriculum for Few-Shot Action Recognition

Supplementary Material

6. Datasets and Evaluation Protocols

6.1. Datasets

We evaluate on five standard few-shot action recognition benchmarks: HMDB51 [15], UCF101 [31], Kinetics-100 [4], SSv2-Small, and SSv2-Full [9].

HMDB51 [15] contains 51 action classes collected from movies and online videos. The clips are short and unconstrained, with camera motion, diverse viewpoints, and background clutter.

UCF101 [31] includes 101 action categories from YouTube videos, covering sports, daily activities, and human–object interactions in unconstrained scenes.

Kinetics-100 is a 100-class subset of Kinetics-400 [4] that is widely used for few-shot evaluation of video recognition models, containing diverse human actions in web videos.

SSv2-Small and **SSv2-Full** are few-shot benchmarks built from the Something-Something V2 dataset [9], which focuses on object-centric, fine-grained human–object interactions under controlled backgrounds. Both SSv2-Small and SSv2-Full select 100 classes from the original 174 categories. SSv2-Small retains 100 videos per class, whereas SSv2-Full keeps all available videos per class.

Taken together, these datasets span controlled, object-centric scenes (SSv2), diverse web videos (Kinetics-100, UCF101), and edited movie clips (HMDB51), providing a diverse testbed for our cross-dataset continual FSAR evaluation. We follow the train/val/test partitions defined in the main paper, and reproduce the resulting class splits for meta-training, meta-validation, and meta-testing in Tab. 11.

Table 11. Class splits used for meta-training (Train), meta-validation (Val), and meta-testing (Test) in our few-shot experiments.

Dataset	Total classes	Train / Val / Test
HMDB51	51	31 / 10 / 10
UCF101	101	70 / 10 / 21
Kinetics-100	100	64 / 12 / 24
SSv2-Small	100	64 / 12 / 24
SSv2-Full	100	64 / 12 / 24

6.2. Few-shot evaluation protocol

We adopt the standard episodic protocol commonly used in FSAR. Each episode is an N-way K-shot classification task sampled from the meta-training or meta-testing splits in Table 11. Unless otherwise stated, we set $N = 5$ and evaluate $K \in \{1, 5\}$ on all benchmarks.

Meta-training. During meta-training on each dataset, we optimize the model on episodes randomly sampled on-the-fly from the training classes using the same (N, K) configuration as in evaluation. For each dataset and (N, K) setting, we train all methods for the same fixed number of meta-training episodes, matching the episode counts used in prior work [35].

Meta-testing. During meta-testing, we generate a large number of episodes to obtain stable estimates. For each dataset and each (N, K) configuration, we randomly sample $E = 10{,}000$ episodes from the meta-test classes and report the mean classification accuracy over all episodes.

6.3. Cross-dataset continual protocol

To evaluate knowledge preservation across domains, we instantiate a cross-dataset continual few-shot protocol with three training tasks and five evaluation benchmarks. Following the notation in the main paper, we denote SSv2-Small as $\mathcal{D}_{\mathrm{S}}$, Kinetics-100 as $\mathcal{D}_{\mathrm{K}}$, and HMDB51 as $\mathcal{D}_{\mathrm{H}}$. We further denote UCF101 as $\mathcal{D}_{\mathrm{U}}$ and SSv2-Full as $\mathcal{D}_{\mathrm{Sf}}$. We construct a sequence of three tasks:

$$\mathcal{D}_{\mathrm{S}} \rightarrow \mathcal{D}_{\mathrm{K}} \rightarrow \mathcal{D}_{\mathrm{H}},$$

which moves from object-centric, controlled scenes (SSv2-Small) to web videos (Kinetics-100) and finally to movie clips (HMDB51). In this protocol, SSv2-Full and UCF101 are only used as held-out evaluation domains and are never seen during continual training.

At each stage t, we adapt the model on the training split of the current dataset $\mathcal{D}_t$ using the same episodic configuration as in the standalone FSAR experiments, and we adopt the 5-way 5-shot setting for all continual-learning results. We reuse the same meta-training schedule as in the single-dataset setting for each task and use the parameters obtained after training on $\mathcal{D}_t$ to initialize adaptation on $\mathcal{D}_{t+1}$. We keep the set of trainable parameters and all optimization hyperparameters identical to those used in the single-task FSAR experiments.

After adaptation on $\mathcal{D}_t$, we evaluate the model on meta-test episodes from all five datasets $\{\mathcal{D}_{\mathrm{H}}, \mathcal{D}_{\mathrm{U}}, \mathcal{D}_{\mathrm{Sf}}, \mathcal{D}_{\mathrm{S}}, \mathcal{D}_{\mathrm{K}}\}$ under the 5-way 5-shot protocol. This yields an accuracy matrix R whose entry $R_{i,t}$ records the accuracy on dataset

$\mathcal{D}_i$ after finishing training on task t. Following the main paper, we report the final Average Accuracy (AvgAcc), computed as the mean accuracy over all five evaluation datasets $(\mathrm{H}, \mathrm{U}, \mathrm{Sf}, \mathrm{S}, \mathrm{K})$ after completing the last task in a sequence. We also report Backward Transfer (BWT) and Forward Transfer (FWT) as defined in the main paper. Intuitively, higher BWT indicates better retention of past knowledge, and higher FWT indicates better zero-shot generalization to future tasks.

6.4. Implementation Details

Architecture and Input. We use CLIP ViT-B/16 [28] as the frozen backbone for both visual and textual encoders. For video input, we uniformly sample $T = 8$ frames (unless specified otherwise in ablation studies) and resize them to 224×224 pixels. For the textual input, we use the standard CLIP tokenizer with a context length of 77.

P2A Configuration. For Orthogonal Subspace Control (OSC), we insert LoRA modules with rank $r = 2$ and dropout 0.25 into the self-attention blocks of the visual backbone and the text encoder, following the layer selection described in the main paper. Crucially, the principal subspace basis U_k is pre-computed offline using the frozen CLIP weights and cached, so it incurs zero computational overhead during meta-training and inference. The effective rank k is determined automatically via entropy-based estimation as described in the main paper.

Table 12. Default hyperparameters for P2A training.

Hyperparameter	Value
Backbone	CLIP ViT-B/16
Input Resolution	224×224
Number of Frames (T)	8
Optimizer	AdamW
Learning Rate	1×10^{-4}
Weight Decay	1×10^{-4}
LoRA Rank (r)	2
Negatives per level (M)	3
Loss Weight (α)	0.1
Total Epochs	10

Training Hyperparameters. We use the AdamW optimizer with a weight decay of $1e^{-4}$. The learning rate is initialized to $1e^{-4}$ and decays following a cosine annealing schedule. The balance hyperparameter α in Eq. 12 is set to 0.1. During the Ranked Negative-prompt Curriculum (RNC), we train for 10 epochs in total and set the curriculum phase boundaries to $T_e = 4$ and $T_m = 6$ epochs, corresponding to easy (epochs 1–4), medium (epochs 5–6), and hard (epochs 7–10) negatives. All experiments are conducted on NVIDIA RTX 3090 GPUs with Automatic Mixed Precision (AMP). Tab. 12 summarizes the default settings.

6.5. Generator-Verifier loop

For each dataset d and each class $a_i \in \mathcal{C}_d$, we run an offline generator–verifier loop, consistent with the main paper. Let M be the number of negatives per difficulty level and $R_{\max}$ the maximum number of rounds.

Algorithm 1 Offline generation of difficulty-ranked negatives

Require: class set $\mathcal{C}_d$, forbidden set $\mathcal{F}_d$, rounds $R_{\max}$, negatives per level M

for each class $a_i \in \mathcal{C}_d$ **do**
  $r \leftarrow 1$
  while $r \leq R_{\max}$ **do**
    $J_i^{\text{gen},r} \leftarrow \text{Generator}(a_i, \mathcal{F}_d, M)$
    $(\text{status}, \mathcal{S}_i^{\text{e}}, \mathcal{S}_i^{\text{m}}, \mathcal{S}_i^{\text{h}}) \leftarrow \text{Verifier}(a_i, J_i^{\text{gen},r})$
    if status $=$ OK **then**
      cache $\mathcal{S}_i^{\ell}$ for $\ell \in \{\text{easy}, \text{medium}, \text{hard}\}$ on disk
      break
    else
      $r \leftarrow r + 1$ *// revise and try again*
    end if
  end while
  if $r > R_{\max}$ **then**
    restart for class a_i with a fresh generator call
  end if
end for

7. Additional Ablation Studies

7.1. Effect of Input Frame Count

In the main paper, all few-shot experiments use $T = 8$ input frames per video. To study the robustness of P2A to the temporal sampling budget, we vary the number of frames T on HMDB51 under the 5-way 1-shot protocol. We keep all other hyperparameters fixed and report results averaged over $E = 10{,}000$ meta-test episodes.

Tab. 13 shows that increasing T from 1 to 8 frames consistently improves accuracy for both the CLIP-FSAR baseline and P2A, as the model observes more temporal context. Using more than 8 frames yields saturated or slightly lower performance, which we attribute to redundant frames and the fixed training schedule. Across all temporal budgets, P2A consistently outperforms the baseline, and the absolute gain remains substantial even in the low-frame regime (e.g., $T = 1$ and $T = 2$), where temporal cues are scarce. This suggests that sharpening decision boundaries in the text space helps compensate for reduced visual evidence.

HMDB51 consists of short trimmed clips collected from movies and web videos, with substantial camera motion and background clutter. In this setting, sparsely sampling $T = 8$ frames already covers most of the action duration. Using more frames mainly adds redundant views under a fixed training budget, which explains the slight drop beyond $T = 8$.

Table 13. Effect of the number of input frames T on HMDB51 under the 5-way 1-shot protocol. We report mean accuracy (%) over $E = 10{,}000$ episodes.

T (frames)	CLIP-FSAR	P2A
1	57.7	62.6
2	63.6	68.5
4	69.7	74.3
8	**75.8**	**82.3**
12	74.3	79.8
16	73.6	79.1

We choose $T = 8$ as our default setting, which provides a good trade-off between accuracy and computational cost. Notably, P2A preserves its advantage even when T is reduced to 1 or 2 frames, indicating that our protection of domain-general semantics and difficulty-ranked negatives are effective under tighter temporal budgets.

7.2. Scaling to Larger N-way Tasks

Few-shot action recognition is often evaluated under the 5-way setting. To assess how P2A scales to more challenging tasks, we further evaluate both the CLIP-FSAR baseline and P2A on HMDB51 under N-way 1-shot protocols with $N \in \{5, 6, 7, 8, 9, 10\}$. As shown in Tab. 14, we reuse the same trained models and only change the meta-testing episodes, keeping the total number of evaluation episodes fixed at $E = 10{,}000$.

Table 14. Scaling to larger N on HMDB51. We compare CLIP-FSAR and P2A under N-way 1-shot protocols with $N \in \{5, \ldots, 10\}$. Results are mean accuracy (%) over $E = 10{,}000$ episodes.

Setting	CLIP-FSAR	P2A
5-way 1-shot	**75.8**	**82.3**
6-way 1-shot	72.1	74.8
7-way 1-shot	69.1	70.7
8-way 1-shot	65.7	66.9
9-way 1-shot	62.6	65.4
10-way 1-shot	60.7	63.5

As N increases from 5 to 10, the accuracy of both methods decreases, since each query must be classified among more candidate classes. Across all values of N, P2A consistently outperforms CLIP-FSAR, and the absolute improvement remains non-trivial, indicating that the decision boundaries induced by Orthogonal Subspace Control and Ranked Negative-prompt Curriculum continue to provide benefits as the task difficulty increases. Minor non-monotonic fluctuations are within the variance of episodic evaluation with $E = 10{,}000$ episodes.

7.3. Sensitivity to Contrastive Loss Weight α

We study the impact of the weight α on the Curricular Contrastive loss $\mathcal{L}_{\text{CC}}$ in the composite objective while keeping all other hyperparameters fixed. Setting $\alpha = 0$ removes $\mathcal{L}_{\text{CC}}$ and reduces P2A to an OSC-only variant optimized with the alignment classification loss $\mathcal{L}_{\text{AC}}$.

Table 15. Sensitivity of P2A to the Curricular Contrastive loss weight α under 5-way 1-shot settings. "–" denotes a variant without $\mathcal{L}_{\text{CC}}$ (i.e., $\alpha = 0$).

α	HMDB51	Kinetics-100
–	79.9	91.8
0.05	79.2	92.3
0.1	**82.3**	**93.7**
0.25	<u>81.4</u>	<u>93.5</u>
0.5	79.2	92.4
1	79.3	91.6
2	78.2	90.1

As shown in Tab. 15, performance follows a unimodal trend on both HMDB51 and Kinetics-100, with the best results obtained at $\alpha = 0.1$. Small positive values of α provide consistent gains by leveraging RNC-generated hard negatives to sharpen decision boundaries. In contrast, larger weights (e.g., $\alpha \geq 0.5$) slightly degrade accuracy, suggesting that over-emphasizing $\mathcal{L}_{\text{CC}}$ can destabilize optimization and overfit to a few difficult negatives.