Adaptive Parameter Selection for Tuning Vision-Language Models

Supplementary Material

A. Appendix

This appendix presents supplementary experimental analyses and technical details supporting the main findings. Sec. A.1 demonstrates CLIP-AST's architectural scalability through comparisons with larger ViT-L/14 models, showing consistent performance gainsacross 11 benchmarks. Comprehensive ablation studies reveal: 1) optimal K values balancing in-distribution accuracy and OOD robustness, 2) the training efficiency of our selective fine-tuning approach and 3) superiority over parameter-efficient methods. Sec. A.2 establishes theoretical connections between our parameter selection strategy and Fisher Information Matrix principles. Sec. A.3 include prompt templates for various datasets and details of SCL loss coefficients.

A.1. More Experiments

Compare With Larger Model. To investigate the architectural scalability of CLIP-AST, we systematically evaluate its stability when scaled to CLIP's larger ViT-L/14 backbone under the 16-shot paradigm, comparing against three representative adaptation approaches: 1) vanilla CLIP [38] as the foundational baseline, 2) cache-based adaptation via Tip-Adapter [55], and 3) self-regulatory prompt tuning through PromptSRC [21]. As demonstrated in Tab. 6, CLIP-AST consistently maintains superior performance on the scaled architecture, achieving a 2.29% average accuracy improvement over PromptSRC across 11 benchmarks. This performance persistence manifests most notably in fine-grained recognition tasks, where CLIP-AST attains 68.34% on FGVC Aircraft and 90.36% on StanfordCars, validating that CLIP-AST effectively preserves model stability during architectural scaling.

Ablation of K in the OOD setting. To examine how the selection of K affects both the generalization of the in-distribution and out-of-distribution, we conduct systematic experiments varying K values during training on ImageNet, followed by evaluation on two OOD benchmarks: ImageNet-V2 and ImageNet-Sketch. Our empirical analysis demonstrates that increasing K generally enhances in-distribution accuracy while progressively improving OOD robustness. However, we observe a marginal performance degradation on ImageNet-V2 at $K = 10$ compared to $K = 9$, suggesting potential over-parameterization effects. These findings indicate that while increasing K benefits model generalization, excessively large K values may induce overfitting risks.

Ablation of training step of transformer fine-tuning stage. The transformer fine-tuning stage estimates parameter importance through gradient second-moment statistics. To investigate its training step impact, we conduct systematic ablation studies on the StanfordCars dataset under 1-shot, 16-shot, and base-to-novel generalization settings. Our experiments compare full transformer training against CLIP-AST variants with varying Stage 1 durations ($1/100$ to 1 epoch) followed by 10-epoch selective fine-tuning. As shown in Tab. 7, some key observations emerge: First, reducing Stage 1 from 1 epoch to $1/100$ epoch preserves 99.3% 16-shot accuracy, demonstrating estimation robustness. Second, full transformer training catastrophically fails in 1-shot and base-to-novel settings. The proposed method has been proven to be effective in terms of both effect and training efficiency.

Comparison with selective fine-tuning methods. We conduct a comprehensive comparison with other selective fine-tuning approaches using the ViT-B/16 model in the 16-shot learning setting. Specifically, we evaluate two representative methods: 1) BitFit [53], which selectively updates bias parameters (we default to selecting all bias terms in Transformer layers), and 2) GPS [57], a gradient-based parameter selection approach where we employ the moving average of gradients from AdamW optimization as the parameter importance metric. Both approaches are implemented within the transformer module of the CLIP architecture. In the case of GPS, only the gradients are utilized as the selection strategy along with the corresponding hyperparametric, while other enhancement factors discussed in the paper have not been incorporated. As demonstrated in Tab. 8, our method achieves superior performance across all 11 benchmark datasets. These results substantiate that our adaptive selection strategy enables more effective identification of mission-critical parameters compared to existing parameter selection.

A.2. Theoretical Connection to Fisher Information

The Fisher Information Matrix (FIM) provides a principled framework for understanding the sensitivity of parameters in probabilistic models. For a parameterized distribution $p(x|\theta)$, the FIM is defined as:

$$F(\theta)_{ij} = \mathbb{E}_{x\sim p(x|\theta)}\left[\frac{\partial \log p(x|\theta)}{\partial \theta_i}\frac{\partial \log p(x|\theta)}{\partial \theta_j}\right]. \quad (12)$$

The diagonal elements F_{ii} quantify the sensitivity of parameter θ_i - large values indicate parameters whose perturbations significantly affect the model's output distribution.

Method	Caltech101	DTD	EuroSAT	FGVC Aircraft	Flowers102	Food101	ImageNet	OxfordPets	StanfordCars	SUN397	UCF101	Avg Acc(%)
CLIP [38]	93.75	52.22	60.50	32.91	78.51	90.88	73.46	93.75	76.33	67.76	76.92	72.45
Tip-Adapter [55]	97.61	75.95	90.84	56.95	98.34	91.70	78.74	95.31	89.13	79.22	88.26	85.64
PromptSRC [21]	97.20	78.10	90.10	54.80	98.50	92.00	79.20	95.10	86.80	80.40	88.90	85.55
CLIP-AST(Ours)	98.29	78.60	94.45	68.34	99.06	91.71	79.20	96.12	90.36	80.23	89.95	87.84

Table 6. Comparison with larger ViT-L/14 backbone under 16-shot setting.

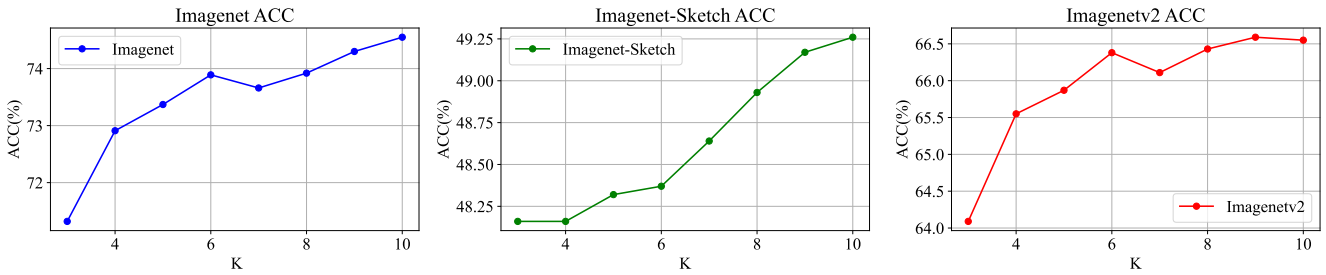

Figure 7. Ablation study of the parameter K in the OOD setting, trained on in-distribution (ImageNet) and evaluated on out-of-distribution target datasets (ImageNet-Sketch and ImageNet-V2).

Method	Stage1	Stage2	1-shot	16-shot	B2N-Base	B2N-Novel	B2N-HM	16-shot Cost
Transformer Full Training	-	-	51.65	86.97	84.01	52.54	40.02	113s
CLIP-AST	1 epoch	-	-	-	-	-	-	11s
CLIP-AST	1/100 epoch	10 epoch	68.72	85.87	84.11	74.05	78.76	101s
CLIP-AST	1/10 epoch	10 epoch	69.25	85.81	83.03	74.20	78.36	101s
CLIP-AST	1 epoch	10 epoch	69.28	86.05	84.16	74.30	78.92	111s

Table 7. Compared with the previous SOTA methods in the out-of-distribution setting, where the model is trained on the ImageNet dataset with 16-shot and evaluated on the ImageNet-V2 and ImageNet-Sketch benchmarks.

Method	Caltech101	DTD	EuroSAT	FGVC Aircraft	Flowers102	Food101	ImageNet	OxfordPets	StanfordCars	SUN397	UCF101	Avg Acc(%)
CLIP [38]	93.51	43.88	48.39	24.66	85.77	66.92	70.06	88.72	66.12	63.32	65.96	65.21
BitFit [53]	96.19	68.56	86.12	44.07	87.36	72.46	94.19	93.68	79.59	74.21	83.14	79.96
GPS [57]	96.22	69.70	85.64	54.12	87.51	73.58	97.27	94.46	84.75	76.50	85.75	82.31
CLIP-AST(Ours)	97.16	75.65	94.51	60.67	87.64	73.91	98.50	94.52	88.45	77.71	87.1	85.07

Table 8. Comparison of our adaptive selection strategy (CLIP-AST) with selective fine-tuning methods, including BitFit [53] and GPS [57], on 11 benchmark datasets in the 16-shot learning setting.

In classification tasks, the empirical FIM can be approximated through gradient statistics:

$$F(\theta) \approx \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\nabla\mathcal{L}(\theta)\nabla\mathcal{L}(\theta)^\top\right], \tag{13}$$

where $F_{ii} \approx \mathbb{E}[g_i^2]$ corresponds to the second moment of gradients. This establishes a direct connection to the AdamW optimizer's second-moment estimate v_i in Eq. (5). Our importance scores $v_i' = \text{Avg}(1/\sqrt{\hat{v}_i})$ are inversely correlated with F_{ii}, prioritizing parameters with lower Fisher information for adaptation.

The theoretical justification emerges from two perspectives: 1) Parameters with low F_{ii} (equivalently small v_i) exhibit stable gradient directions, permitting larger updates without destabilizing pre-trained knowledge. 2) High F_{ii} parameters correspond to "anchors" in the pre-trained feature space - freezing them preserves zero-shot capabilities while allowing task-specific adaptation through flexible parameters.

A.3. More Details of Experimental Setting

Prompt Templates. Following prior work [55, 59], we adopt the following prompt templates for different datasets. Notably, while previous works typically employ multiple prompt templates for ImageNet, we use a single simplified template across all variants. The complete template configurations are as follows:

- `Caltech101`: `a photo of a {}.`
- `DTD`: `{} texture.`
- `EuroSAT`: `a centered satellite photo of {}.`
- `FGVC Aircraft`: `a photo of a {}, a type of aircraft.`
- `Food101`: `a photo of {}, a type of food.`
- `ImageNet`, `ImageNetV2`, `ImageNet Sketch`, `StanfordCars`, `SUN397`: `a photo of a {}.`
- `OxfordFlowers`: `a photo of a {}, a type of flower.`

- `OxfordPets: a photo of a {}, a type of pet.`
- `UCF101: a photo of a person doing {}.`

Coefficients for the SCL loss. The coefficients controlling the SCL loss exhibit task-aware adaptation based on distinct learning objectives across settings. For the few-shot setting where overfitting mitigation must be carefully balanced with preserving discriminative power in low-data regimes, we employ moderate coefficients (typically around 1.0) to maintain stable gradient signals from the primary classification objective. In contrast, for base-to-novel generalization tasks where feature-space regularization plays a more critical role, we amplify the SCL coefficients (ranging from 10 to 100, and even higher for some datasets) to reduce overfitting on base classes.