Concept-Guided Fine-Tuning: Steering ViTs away from Spurious Correlations to Improve Robustness

Supplementary Material

6. Implementation Details

Hyperparameters. In Table 8, we summarize the hyperparameter configurations used across all experiments. Our method is highly stable and relies on a consistent set of hyperparameters, with the exception of the learning rate. In contrast, RRR requires model-specific hyperparameter tuning and is notably sensitive to these choices. GradMask also demands careful learning rate tuning for each model. Its performance varies substantially with small adjustments to this parameter, and achieving convergence of the background loss proved challenging in both cases. RRDA shares no common hyperparameters with the other methods aside from the learning rate. CFT uses fixed loss weights $\lambda_{\text{non-concept}} = 1.2$, $\lambda_{\text{concept}} = 0.5$, $\lambda_{\text{align}} = 0.8$, and $\lambda_{\text{cls}} = 0.2$ across all models and datasets. We heavily weight the $L_{\text{non-concept}}$ loss, as reliance on spurious cues (areas without concepts in this case) is the primary issue to be corrected.

Ablation on Concept Validation Thresholds. Following the initial label-free concept discovery procedure of [39], we further refined the resulting pool to obtain a high-quality concept set. To this end, we enforced minimum thresholds of an occurrence rate of at least 15% and spatial coverage of at least 20%. Applying these criteria to IN produced 1,852 validated concepts. Across the dataset, concepts appeared in 29% of images on average, and those satisfying the filtering criteria covered roughly 35% of the relevant region. We examined the effect of varying the occurrence-rate and spatial-coverage thresholds on Top-1 accuracy for ViT-B evaluated on IN-A and IN-R. The best results were obtained using our default thresholds of 15% and 20%. Increasing the thresholds to 40%/40% reduced the number of concepts to 694 and led to a noticeable drop in performance (IN-A: 24.59, IN-R: 44.23), presumably because many informative concepts were discarded. Relaxing the thresholds to 5%/10% increased the concept count to 2,435 but introduced substantial noise, which similarly harmed performance (IN-A: 25.13, IN-R: 44.92).

Concept Set Creation. For concept set construction, we used $P = 30$ samples per class, guided by the occurrence rate and spatial coverage feedback. Using thresholds of occurrence rate $\geq 15\%$ and spatial coverage $\geq 20\%$, this process yielded a total of 1852 concepts across 500 classes (half of ImageNet-1K [17]). The filtering (occurrence rate and spatial coverage feedback) proceeded in two stages: we first applied the occurrence-rate threshold, and then evaluated the remaining candidates using the spatial-coverage criterion. In our experiments, all concepts that passed the occurrence-rate filter also satisfied the 20% coverage threshold. While not required for our study, this procedure could be extended with an iterative refinement step — potentially assisted by an LLM to identify additional concepts that jointly satisfy both constraints.

Clarification on the P parameter. The parameter P is used exclusively during the validation of the initial concept sets. We first generate the initial concept sets using the procedure of Oikarinen et al. [39]. Then, for each class, we examine $P = 30$ images to compute the occurrence rate and spatial coverage. These measurements are subsequently used to filter and refine the initial concept sets.

7. Concept Validation Effect

Effect of the optional concept validation step. We further compared CFT performance with and without the concept validation stage, evaluating Top-1 accuracy on both IN-A and IN-R. Without validation, CFT achieves 26.01 on IN-A (vs. 27.92 with validation) and 47.19 on IN-R (vs. 48.51 with validation). Although the validation step provides a consistent performance boost, the non-validated variant remains competitive and continues to outperform several robustness-oriented baselines (Tab. 1). Yet, using the validation step provides state-of-the-art performance, outperforming all other approaches.

Ablation on Concept Validation Thresholds. Following the initial label-free concept discovery procedure of [39], we further refined the resulting pool to obtain a high-quality concept set. To this end, we enforced minimum thresholds of an occurrence rate of at least 15% and spatial coverage of at least 20%. Applying these criteria to IN produced 1,852 validated concepts. Across the dataset, concepts appeared in 29% of images on average, and those satisfying the filtering criteria covered roughly 35% of the relevant region. We examined the effect of varying the occurrence-rate and spatial-coverage thresholds on Top-1 accuracy for ViT-B evaluated on IN-A and IN-R. The best results were obtained using our default thresholds of 15% and 20%. Increasing the thresholds to 40%/40% reduced the number of concepts to 694 and led to a noticeable drop in performance

Table 8. Hyperparameter selection for all methods.

	Model	λ_{align}	λ_{cls}	$\lambda_{\text{non-concept}}$	λ_{concept}	Learning rate
CFT	ViT-B	0.8	0.2	1.2	0.5	5e−7
	DINOv2	0.8	0.2	1.2	0.5	6e−7
	DeiT	0.8	0.2	1.2	0.5	8e−7
	CNv2	0.8	0.2	1.2	0.5	3e−6
RRR	ViT-B	-	2e−6	1e−10	-	2e−6
	DINOv2	-	2e−8	1e−10	-	1e−5
	DeiT	-	2e−6	1e−10	-	5e−6
	CNv2	-	2e−6	1e−8	-	3e−6
GradMask	ViT-B	-	0.1	50	-	0.001
	DINOv2	-	0.1	50	-	0.005
	DeiT	-	0.1	50	-	0.001
	CNv2	-	0.1	50	-	0.05
RRDA	ViT-B	-	-	-	-	2e-6
	DINOv2	-	-	-	-	1e-5
	DeiT	-	-	-	-	5e-6
	CNv2	-	-	-	-	3e-6

(IN-A: 24.59, IN-R: 44.23), presumably because many informative concepts were discarded. Relaxing the thresholds to 5%/10% increased the concept count to 2,435 but introduced substantial noise, which similarly harmed performance (IN-A: 25.13, IN-R: 44.92).

8. Main evaluation - full results

The results in Table 1 are averaged over five random seeds, where the subset of ImageNet classes used for fine-tuning is varied while keeping all other parameters fixed. Table 9 reports the corresponding standard deviations for this experiment.

9. Limitations and Future Work

9.1. Failure Cases

Despite strong overall performance, CFT exhibits identifiable failure modes:

Abstract or non-visual concepts. GPT-4o-mini rarely generates concepts that are semantically appropriate but not visually grounded (e.g., "aggressive behaviour" for a lion). GroundedSAM cannot localize such concepts, resulting in empty high-confidence masks.

Very small object parts. For parts occupying $< 2\%$ of image area (e.g., the beak of a distant bird), GroundedSAM's hit rate decreases. The impact on final accuracy is limited, as the remaining concepts provide sufficient coverage, but fine-grained part-level reasoning may be impaired.

Domain mismatch between LLM and target domain. In specialized domains (medical imaging, satellite imagery), GPT-4o-mini's concept vocabularies may be imprecise or incomplete. In such settings, domain-specific LLMs or expert-curated concept lists are recommended.

9.2. Limitations

While our proposed CFT framework demonstrates improvements in model robustness across multiple benchmarks, several limitations warrant discussion.

Dependency on Vision-Language Models. Our approach relies on the quality and capabilities of GroundedSAM for concept localization. While this eliminates the need for manual annotations, it introduces a dependency on the grounding model's performance. In cases where GroundedSAM fails to accurately segment concepts, particularly for abstract or fine-grained semantic attributes, the quality of guidance masks may degrade, potentially limiting CFT's effectiveness.

Computational Overhead. While CFT is designed as a lightweight fine-tuning procedure requiring only 1,500 images, the initial concept creation and validation stage involves processing 30 samples per class through GroundedSAM, which introduces non-negligible computational costs. For datasets with thousands of classes, this preprocessing step could become a practical bottleneck. Moreover, computing relevance maps via AttnLRP during training adds overhead compared to standard gradient-based methods, though this cost is amortized across the fine-tuning procedure.

Table 9. Evaluation over 5 different seeds.

Model	Metric	IN-V	IN-V2	IN-A	ObjectNet	IN-R	IN-Sketch
ViT-B	R@1	81.35 ±0.28	69.19 ±0.51	27.76 ±0.14	54.28 ±0.82	48.47 ±0.39	37.06 ±0.66
	R@5	95.51 ±0.47	84.77 ±0.19	62.75 ±0.73	75.46 ±0.32	70.50 ±0.56	62.59 ±0.21
DINOv2	R@1	81.44 ±0.61	71.91 ±0.34	27.71 ±0.80	53.89 ±0.17	48.53 ±0.54	44.74 ±0.42
	R@5	95.65 ±0.42	88.15 ±0.77	62.36 ±0.54	75.58 ±0.59	70.73 ±0.18	68.90 ±0.69
DeiT	R@1	82.61 ±0.44	73.11 ±0.34	27.72 ±0.57	54.24 ±0.71	48.33 ±0.23	44.83 ±0.36
	R@5	95.77 ±0.68	88.58 ±0.49	62.20 ±0.31	75.46 ±0.15	70.65 ±0.75	69.55 ±0.53
CNv2	R@1	87.27 ±0.43	75.25 ±0.63	27.93 ±0.41	54.19 ±0.50	48.37 ±0.78	46.14 ±0.27
	R@5	95.71 ±0.52	89.50 ±0.38	62.40 ±0.65	75.62 ±0.22	70.68 ±0.46	70.81 ±0.60

Architecture Specificity. Although we demonstrate CFT's applicability to both ViTs and CNNs (ConvNeXt-V2), the primary design and optimization were conducted with transformer architectures in mind. The adaptation to CNNs, while successful, required modifications to the relevance computation procedure. Extending CFT to other emerging architectures may require additional architectural considerations.

9.3. Future Work

Several promising directions emerge from this work that could further advance concept-guided robustness in vision models.

Adaptive Concept Weighting. Our current approach treats all validated concepts equally during fine-tuning. However, different concepts may contribute unequally to robustness for specific distribution shifts. Developing methods to dynamically weight concepts based on their discriminative power or relevance to particular OOD scenarios could yield more targeted robustness improvements. This could be achieved through simple masking response-based approaches or by using concept activation vectors (CAVs) to produce concept-class importance weights.

Hierarchical and Compositional Concepts. Our framework currently treats concepts as independent entities. However, real-world objects exhibit hierarchical structure and compositional semantics. Incorporating compositional reasoning, where complex concepts are built from simpler primitives, could enhance both interpretability and robustness.

Application to Other Domains. Although this work focuses on image classification, the underlying principle of aligning model reasoning with semantically meaningful concepts extends naturally to other computer vision tasks (e.g., object detection, semantic segmentation, video understanding) and potentially to non-vision domains where structured, interpretable representations are valuable. Exploring these extensions could validate the generality of concept-guided learning as a robustness paradigm.