Interpretable Adversarial Prompt Tuning via Semantic Concepts

Pedram MohajerAnsari[1] Zongxi Liu[2] Yi Zhu[2] Amir Salarpour[1] Mert D. Pesé[1]
[1]Clemson University [2]Wayne State University
pmohaje@clemson.edu, zongxiliu@wayne.edu, yzhu39@wayne.edu
amir.salarpour@gmail.com, mpese@clemson.edu

Abstract

Adversarial prompt tuning adapts vision-language models efficiently but suffers from poor few-shot performance and lack of interpretability. We propose concept-enhanced adversarial prompt tuning, which replaces abstract context vectors with structured semantic concepts. Our approach augments base text embeddings with weighted concept combinations, optimizing only scalar weights while keeping concept representations fixed. This provides semantic structure for few-shot learning and interpretability through learned weights. Across six benchmarks, we achieve substantial improvements: +19.58pp clean accuracy on EuroSAT 1-shot and substantial robustness gains against PGD-100 attacks. Our method uses 98.7% fewer parameters than class-specific approaches (5,086 vs. 393K). Analysis reveals semantically meaningful patterns: kitchens emphasize cabinets ($\alpha = +0.92$) while suppressing sinks ($\alpha = -0.88$) to avoid bathroom confusion. Our approach successfully balances adversarial robustness, few-shot generalization, and interpretability.

1. Introduction

Vision-Language Models (VLMs) such as CLIP [4] have revolutionized multi-modal learning through contrastive pre-training on large-scale image-text datasets, demonstrating remarkable zero-shot generalization. To efficiently adapt pre-trained VLMs to downstream tasks, prompt tuning, which optimizes learnable prompts while keeping model weights frozen, has emerged as a parameter-efficient alternative to full fine-tuning [17, 18]. Recent work shows that adversarial prompt tuning substantially improves both clean accuracy and robustness by learning context vectors through adversarial training [9, 16].

Despite their effectiveness, existing adversarial prompt tuning methods face two critical limitations. ❶ *Abstract embeddings struggle in few-shot scenarios*: Current methods optimize prompts as abstract, unstructured vectors in continuous embedding space. Without semantic grounding in human-interpretable visual concepts, these abstract representations must learn robustness patterns entirely from scratch. In few-shot regimes (1-4 examples per class) where supervision is severely limited, this approach exhibits degraded performance, as abstract optimization lacks the inductive biases needed to generalize effectively from scarce data. ❷ *Learned prompts lack interpretability*: The optimized prompt vectors offer no connection to semantic attributes or visual concepts. Practitioners cannot understand which properties (e.g., texture, shape, color) contribute to adversarial robustness for specific classes, limiting both scientific understanding and practical refinement of defense strategies.

We address these limitations by replacing abstract context vectors with structured semantic concepts. Our approach represents each class as a base text embedding ("a photo of a [CLASS]") augmented by weighted combinations of human-interpretable concept embeddings (e.g., "furry," "metallic," "outdoor"). Learnable scalar weights α determine each concept's contribution, optimized through adversarial training. This design provides: (1) *semantic structure*, pre-existing concept knowledge supplies strong inductive biases that improve few-shot learning when abstract optimization struggles; (2) *interpretability*, learned α weights reveal which semantic attributes contribute to robustness for each class. We obtain 6-10 diverse concepts per class through structured LLM prompting [1], then optimize only the combination weights ($C \times K$ scalar parameters), maintaining efficiency while incorporating semantic priors.

We evaluate on 6 benchmark datasets (Caltech101 [5], EuroSAT [7], OxfordPets [12], StanfordCars [8], SUN397 [15], Food101 [3]) across 4 data regimes (1-shot, 4-shot, 16-shot, full training data), measuring clean accuracy and adversarial robustness against PGD-100 attacks [10]. All models use adversarially pre-trained CLIP backbones [11] and identical training procedures (on-the-fly PGD-3 during training) [10], enabling controlled comparison with baseline APT [9]. Results show concept-driven prompting substantially outperforms abstract context learning. On clean accuracy, our method achieves consistent gains across all 6 datasets, with dramatic few-shot improvements: +19.58% on EuroSAT

(1-shot), +7.46% on StanfordCars (1-shot), and +10.73% on SUN397 (4-shot). For adversarial robustness, we improve on 4 out of 6 datasets, with notable gains on Caltech101 (+4.05% robust accuracy at 1-shot), StanfordCars (+1.52%), and particularly strong improvements as training data increases. These results validate that structured semantic concepts provide essential inductive biases for few-shot adversarial robustness.

Contributions.

- We propose concept-enhanced adversarial prompt tuning that replaces abstract context vectors with interpretable semantic concepts and learns class-specific combination weights through adversarial training, addressing poor few-shot performance and lack of interpretability in existing methods.
- Through experiments across 6 datasets and 4 data regimes, we demonstrate that semantic concepts substantially improve few-shot performance while providing interpretable insights into robustness mechanisms through learned concept weights.

2. Related Works

Zhou *et al.* [18] introduce CoOp, a prompt-learning method for CLIP-style vision–language models that replaces handcrafted context words with a small set of learnable continuous vectors while keeping the pretrained encoders frozen. They study both a unified context shared across classes and class-specific contexts, showing strong few-shot improvements over manual prompts and competitive transfer under domain shift. Zhou *et al.* [17] propose CoCoOp to address CoOp's limited generalization to unseen classes by making the prompt instance-conditional. A lightweight conditioning network produces an image-dependent token that is combined with learnable context vectors, yielding dynamic prompts per input and improved base-to-new class generalization across diverse datasets.

Gao *et al.* [6] present CLIP-Adapter, which adapts CLIP using small feature-adapter modules rather than changing text prompts. The method inserts lightweight bottleneck adapters on top of frozen encoders and blends adapted features with original CLIP features via residual mixing, enabling efficient training and strong few-shot performance. Shu *et al.* [13] introduce Test-Time Prompt Tuning (TPT), which tunes prompts online using only unlabeled test inputs. By optimizing a marginal-entropy objective over multiple augmentations of each test image (with confidence-based selection), TPT improves zero-shot generalization under distribution shifts without requiring labeled downstream data.

Zhang *et al.* [16] propose Adversarial Prompt Tuning (AdvPT) to improve robustness of vision–language models by learning soft prompts while keeping model weights fixed. The approach leverages adversarial image features (e.g., via an embedding bank) and tunes prompts to better align text and adversarial image representations, strengthening performance against adversarial attacks. Bar *et al.* [2] formulate visual prompting as an image inpainting problem by constructing multi-panel grids containing example input–output pairs plus a query. Training large inpainting models on a large corpus of figure grids enables the same model to perform diverse image-to-image tasks (e.g., segmentation, edges, colorization) at test time without task-specific fine-tuning.

Li *et al.* [9] show that adversarial robustness in pretrained vision–language models is highly sensitive to prompts and that even a single learned prompt token can significantly improve robustness. They propose adversarial prompt tuning that learns a robust soft prompt (unified or class-specific) using adversarial training while freezing the encoders, improving both clean and adversarial accuracy in low-shot settings. Wang *et al.* [14] propose TAPT, a test-time adversarial prompt tuning approach for robust inference that adapts prompts during evaluation to counter adversarial perturbations. By updating prompt parameters on-the-fly from unlabeled test data and incorporating robustness-aware objectives, TAPT improves adversarial performance while maintaining strong clean accuracy.

3. Threat Model

We consider a white-box adversary with full access to the vision-language model, including the frozen encoders (f_v, f_t) and learned prompt parameters. The attacker's objective is to generate visually imperceptible perturbations that cause misclassification.

Attack Model. The adversary crafts ℓ_∞-bounded perturbations with budget $\epsilon = 4/255$ using Projected Gradient Descent (PGD). Following standard protocols [10], we evaluate against PGD-100 at test time, providing a strong approximation of worst-case adversarial examples. The attacker has complete knowledge of model architecture, parameters, and training procedure, the strongest threat scenario.

Constraints and Scope. Perturbations satisfy $\|x^{\text{adv}} - x\|_\infty \leq \epsilon$, ensuring adversarial images remain visually similar to originals. We focus on standard PGD attacks rather than adaptive attacks specifically targeting our concept structure, as our goal is demonstrating improved robustness over baseline prompt tuning methods.

Defense Objective. The defender maintains high accuracy on both clean and adversarial inputs while providing interpretable robustness insights. Unlike traditional defenses that sacrifice interpretability for robustness, our concept-based approach achieves both, enabling analysts to identify which semantic attributes contribute to model robustness and potential class-specific vulnerabilities.

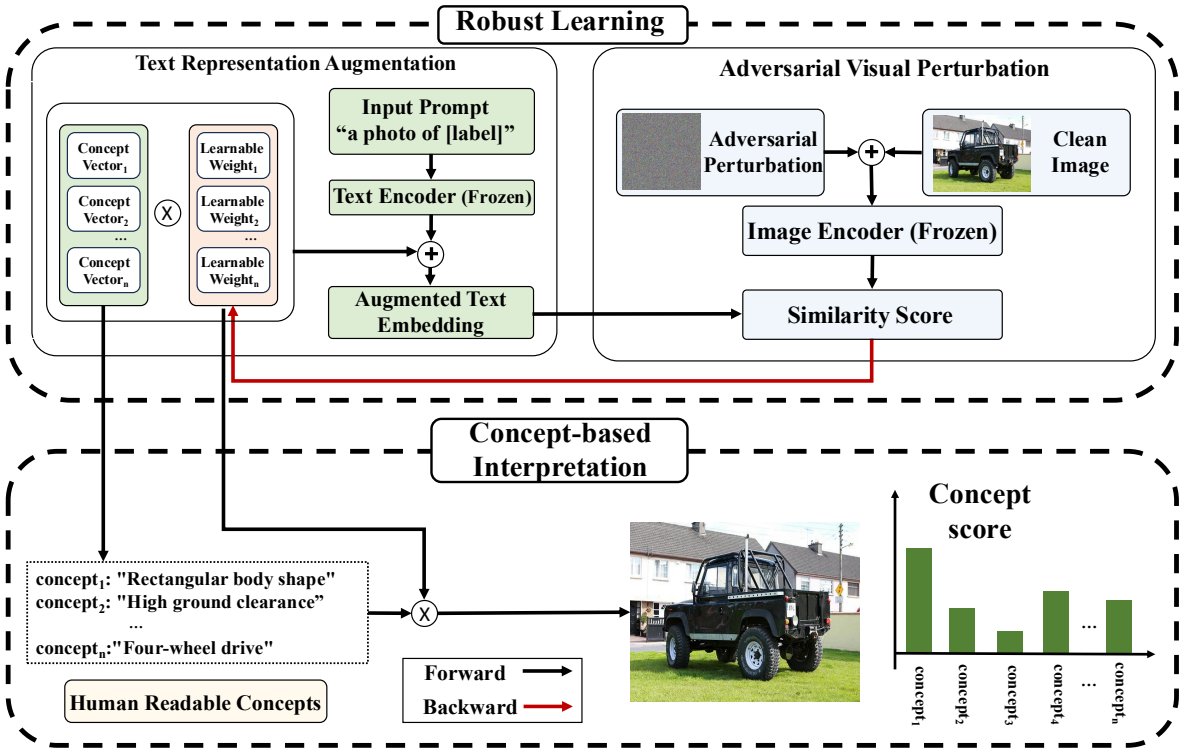

Figure 1. **Concept-enhanced adversarial prompt tuning framework.** *Top (Robust Learning):* During training, learnable scalar weights α combine pre-defined semantic concept embeddings with base text prompts. Adversarial training optimizes these weights against perturbed images while keeping encoders frozen. *Bottom (Concept-based Interpretation):* The learned α weights directly reveal which concepts contribute to robustness e.g., "rectangular body shape" emphasizes vehicle structure.

4. Methodology

Problem Formulation. We consider the standard vision-language model setup where a frozen image encoder $f_v : \mathcal{X} \to \mathbb{R}^d$ and frozen text encoder $f_t : \mathcal{T} \to \mathbb{R}^d$ map images and text to a shared d-dimensional embedding space. Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ with C classes, the zero-shot classification score for image x and class c is computed via cosine similarity:

$$s(x, c) = \frac{f_v(x)^T f_t(t_c)}{\|f_v(x)\| \|f_t(t_c)\|}, \tag{1}$$

where t_c is the text prompt for class c, typically instantiated as "a photo of [CLASS]". Our goal is to learn robust text representations that maintain high classification accuracy under bounded ℓ_∞ adversarial attacks while providing interpretable insights into robustness mechanisms.

Adversarial Prompt Tuning with Concepts. Our approach extends adversarial prompt tuning by replacing abstract learnable vectors with structured semantic concepts. Figure 1 illustrates our framework, which consists of robust learning during training and concept-based interpretation after training.

We represent each class c through a base text embedding augmented by weighted combinations of concept embeddings. Let $\mathcal{C}_c = \{c_1, c_2, \ldots, c_K\}$ denote K human-readable concepts for class c (e.g., "rectangular body shape", "high ground clearance" for vehicles). We obtain concept embeddings by encoding these descriptions through the frozen text encoder: $e_{c,k} = f_t(c_k) \in \mathbb{R}^d$ for $k = 1, \ldots, K$.

The augmented text representation for class c is:

$$\tilde{t}_c = t_c^{\text{base}} + \sum_{k=1}^{K} \alpha_{c,k} \cdot e_{c,k}, \tag{2}$$

where $t_c^{\text{base}} = f_t(\text{"a photo of [CLASS]"})$ is the base embedding and $\alpha_{c,k} \in \mathbb{R}$ are learnable scalar weights. Only the weights $\{\alpha_{c,k}\}$ are optimized during training—concept embeddings $\{e_{c,k}\}$ remain fixed. This reduces parameters from $C \times d$ (for class-specific context) to $C \times K$ scalars, where typically $K \ll d$. The final text embedding is normalized: $\hat{t}_c = \tilde{t}_c / \|\tilde{t}_c\|$.

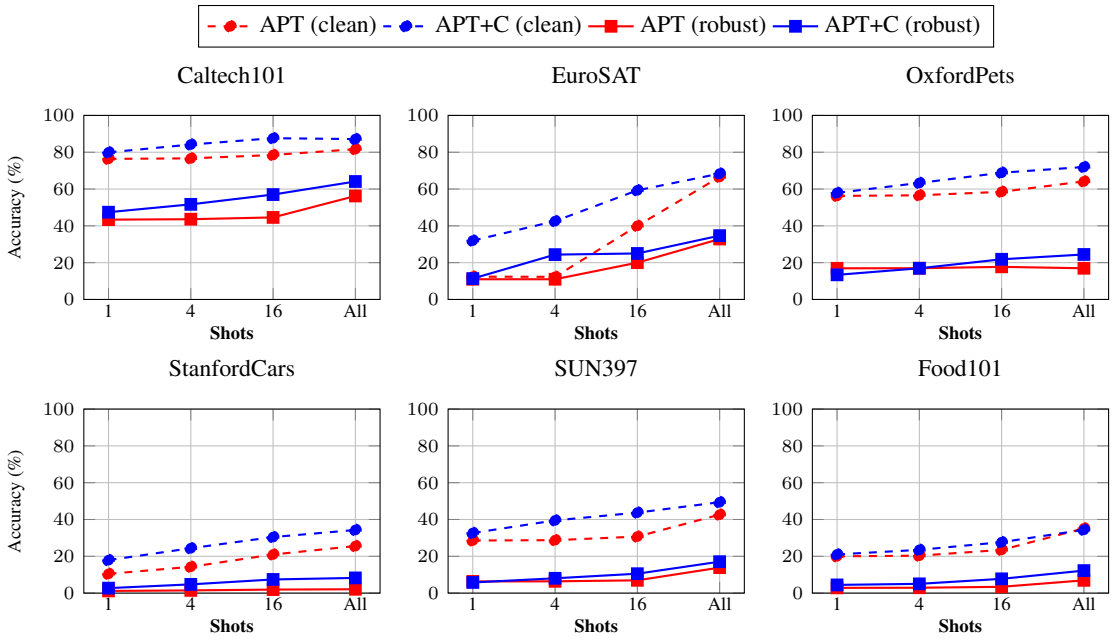

Figure 2. **Clean and robust accuracy across data regimes.** Dashed lines show clean accuracy; solid lines show robustness under PGD-100 attacks ($\epsilon = 4/255$). Our method (blue) consistently outperforms APT baseline (red) on both metrics. Clean accuracy shows dramatic few-shot gains (+19.58pp EuroSAT 1-shot, +10.73pp SUN397 4-shot). Robust accuracy improves on most datasets, with notable gains on Caltech101 (+7.87pp full-shot) and StanfordCars (4× improvement). Circles denote clean accuracy; squares denote robust accuracy.

We train the concept weights $\{\alpha_{c,k}\}$ to maximize robustness against adversarial perturbations. For each training image (x, y), we generate adversarial examples using T-step Projected Gradient Descent (PGD). The training objective minimizes cross-entropy loss on adversarial examples:

$$\min_{\{\alpha\}} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathcal{L}_{\text{CE}}(\text{PGD}_T(x), y; \{\alpha\})\right], \tag{3}$$

where the loss is computed using augmented text embeddings $\{\hat{t}_c\}$ from Equation 2. The image and text encoders remain frozen; only concept weights are updated.

We generate semantic concepts using GPT-4, prompting it to produce K visually distinctive attributes for each class. For example, for "jeep", we obtain concepts like "rectangular body shape", "high ground clearance", "four-wheel drive appearance". These are encoded via f_t to obtain concept embeddings. We set $K = 6$ for most datasets and $K = 10$ for EuroSAT.

Post-Training Interpretability. After training, learned weights $\{\alpha_{c,k}\}$ directly reveal which concepts contribute to adversarial robustness. Positive weights ($\alpha_{c,k} > 0$) indicate emphasized concepts; negative weights ($\alpha_{c,k} < 0$) indicate suppressed concepts. Unlike abstract vectors in $\mathbb{R}^d$, each $\alpha_{c,k}$ quantifies the importance of a human-readable attribute (e.g., "cabinets" for kitchens, "bacon layers" for club sandwiches).

Optimization Details. We use AdamW optimizer with learning rate 0.01 and weight decay 10^{-4}. For adversarial training, we use PGD with $T = 3$ steps, step size $\eta = 2.67/255$, and perturbation budget $\epsilon = 4/255$. At test time, we evaluate against PGD-100 ($T = 100$ steps). We train for 50 epochs with seed 0. Concept weights are initialized to zero.

5. Experiment

Performance Analysis. We evaluate our concept-enhanced adversarial prompt tuning across six benchmark datasets: object recognition (Caltech101), satellite imagery (EuroSAT), pet breed classification (OxfordPets), vehicle identification (StanfordCars), scene understanding (SUN397), and food classification (Food101). For each dataset, we train models using four data regimes (1-shot, 4-shot, 16-shot, and full training data) and evaluate both clean accuracy and robustness against PGD-100 attacks with $\epsilon = 4/255$.

Figure 2 shows results comparing our APT+Concepts method against the APT baseline. Our approach demonstrates consistent improvements across nearly all dataset-shot combinations, with particularly pronounced gains in data-scarce settings. EuroSAT 1-shot improves from 12.40% to 31.98% (+19.58pp), StanfordCars 1-shot gains +7.46pp, and SUN397 4-shot improves +10.73pp. These gains show that semantic structure helps the model generalize from minimal examples. As the number of shots increases, the performance gap generally narrows but remains positive in most cases. Caltech101 shows consistent improvements across all shot settings (+3.53pp to +5.48pp), while Food101 exhibits

Table 1. **Learned concept weights reveal interpretable semantic patterns.** Each row shows one representative class with top-5 weighted concepts. Positive weights ($\alpha > 0$) emphasize class-defining attributes, while negative weights ($\alpha < 0$) suppress confusing features.

Dataset	Class	Top Positive Weights (emphasize)	Top Negative Weights (suppress)
EuroSAT	Industrial buildings	Paved parking lots ($\alpha = +1.12$) Boxy, rectangular shapes ($\alpha = +0.69$) Regular human activity ($\alpha = +0.29$) Visible structures ($\alpha = +0.24$) Open paved areas ($\alpha = +0.19$)	Industrial equipment ($\alpha = -0.70$) Smokestacks, ventilation ($\alpha = -0.53$) Manufacturing machinery ($\alpha = -0.42$) Dense infrastructure ($\alpha = -0.35$) Heavy vehicles ($\alpha = -0.28$)
Food101	Club sandwich	Bacon strip layers ($\alpha = +1.03$) Sliced poultry/turkey ($\alpha = +0.22$) Lettuce and tomato ($\alpha = +0.02$) Triple-layer bread ($\alpha = +0.01$) Toasted appearance ($\alpha = +0.01$)	Toothpick hold ($\alpha = -0.33$) Serving presentation ($\alpha = -0.17$) Plate arrangement ($\alpha = -0.12$) Side garnish ($\alpha = -0.08$) Condiments visible ($\alpha = -0.05$)
SUN397	Kitchen	Cabinets ($\alpha = +0.92$) Countertops ($\alpha = +0.29$) Appliances ($\alpha = +0.21$) Food storage ($\alpha = +0.18$) Cooking areas ($\alpha = +0.14$)	Sink and faucet ($\alpha = -0.88$) Stove or range ($\alpha = -0.55$) Dining furniture ($\alpha = -0.42$) Bathroom fixtures ($\alpha = -0.38$) Natural lighting ($\alpha = -0.31$)
OxfordPets	Bombay (cat)	Short, sleek fur ($\alpha = +0.64$) Rounded ears ($\alpha = +0.51$) Round copper eyes ($\alpha = +0.05$) Compact body ($\alpha = +0.03$) Smooth coat ($\alpha = +0.02$)	Short, thick legs ($\alpha = -0.22$) Black or brown coat ($\alpha = -0.10$) Muscular build ($\alpha = -0.08$) Long tail ($\alpha = -0.06$) Pointed muzzle ($\alpha = -0.04$)
StanfordCars	2012 Tesla Model S	Sleek aerodynamic body ($\alpha = +0.50$) Flush door handles ($\alpha = +0.31$) Large touchscreen dash ($\alpha = +0.18$) Panoramic roof ($\alpha = +0.12$) Low profile design ($\alpha = +0.09$)	Traditional grille ($\alpha = -0.46$) Chrome trim ($\alpha = -0.28$) Engine exhaust ($\alpha = -0.19$) Conventional keys ($\alpha = -0.15$) Mechanical buttons ($\alpha = -0.11$)
Caltech101	Accordion	Multiple rows of keys ($\alpha = +0.82$) Metal grille on front ($\alpha = +0.59$) Rectangular shape ($\alpha = +0.27$) Button controls ($\alpha = +0.15$) Compact design ($\alpha = +0.12$)	Straps for carrying ($\alpha = -0.40$) Bellows between keys ($\alpha = -0.20$) Leather handles ($\alpha = -0.15$) Decorative trim ($\alpha = -0.12$) Size variations ($\alpha = -0.08$)

comparable performance in the full-data setting (34.38% vs. 35.18%). OxfordPets maintains advantages of +7.82pp in full-data (64.16% to 71.98%), particularly for fine-grained classification where semantic distinctions are subtle.

For adversarial robustness under PGD-100 attacks, our method maintains consistent advantages on most datasets. Caltech101 demonstrates strong robustness gains across all settings (+4.05pp to +7.87pp), while StanfordCars exhibits the most dramatic improvement, with robust accuracy more than doubling in 1-shot (1.19% to 2.71%) and nearly quadrupling at full-shot (2.09% to 8.26%). However, not all datasets show uniform improvements. OxfordPets exhibits a decrease in 1-shot robust accuracy (16.93% to 13.38%), though this reverses in higher-shot settings with eventual gains of +7.50pp at full-shot. Similarly, SUN397 shows a slight 1-shot decrease (6.35% to 5.79%) that reverses by 4-shot and strengthens in higher-shot settings. These patterns suggest dataset-specific sensitivity to concept quality in 1-shot settings where single examples are insufficient to calibrate the α weights.

The results demonstrate that incorporating semantic concepts addresses poor few-shot generalization and lack of interpretability. Combined with the interpretability analysis in Table 1, these results establish that concept-enhanced adversarial prompt tuning balances performance, parameter efficiency, and semantic transparency.

Interpretability of Learned Concept Weights. To demonstrate that learned concept weights provide human-interpretable insights, we analyze the α parameters across all six datasets. Our method optimizes only 5,086 scalar α weights, achieving 98.7% parameter reduction compared to 393K parameters for class-specific context (CSC) approaches [18], where each class learns separate prompt token embeddings. Table 2 shows that the proportion of significant weights ($|\alpha| > 0.1$) varies substantially across datasets, ranging from 14.7% for Caltech101 to 79.0% for EuroSAT. Satellite imagery classification shows both higher mean absolute weights (0.345 vs. 0.086 for Caltech101) and maximum weight magnitudes (1.12 vs. 0.82), reflecting the need for stronger semantic distinctions in overhead imagery.

Table 1 reveals that learned weights capture semantically meaningful patterns. For satellite imagery (EuroSAT, *industrial buildings*), the model emphasizes "paved parking lots" ($\alpha = +1.12$) and "boxy shapes" ($\alpha = +0.69$) while suppressing "industrial equipment" ($\alpha = -0.70$) and "smokestacks" ($\alpha = -0.53$). For food classification (Food101, *club sandwich*), the model emphasizes ingredient composition—"bacon layers" ($\alpha = +1.03$) and "sliced poultry" ($\alpha = +0.22$), while suppressing presentation features like "toothpick hold" ($\alpha = -0.33$), showing that adversarial robustness benefits from focusing on intrinsic characteristics rather than serving methods.

Table 2. **Learned concept weight statistics.** EuroSAT's high mean weight (0.345) and 79% significant weights reflect the need for strong semantic distinctions in satellite imagery.

Dataset	Params	Mean $\|\alpha\|$	Max $\|\alpha\|$	Active (%)
Caltech101	600	0.086	0.82	14.7
EuroSAT	100	0.345	1.12	79.0
Food101	606	0.199	1.03	57.3
OxfordPets	222	0.090	0.64	26.1
StanfordCars	1,176	0.088	0.50	25.9
SUN397	2,382	0.143	1.00	29.8
Total	**5,086**	**0.142**	**1.12**	**38.8**

For scene classification (SUN397, *kitchen*), the model demonstrates contextual discrimination: cabinets receive high positive weight ($\alpha = +0.92$) while sinks receive strong negative weight ($\alpha = -0.88$). This pattern is interpretable: while both kitchens and bathrooms contain sinks, cabinets more distinctively identify kitchen environments. For breed recognition (OxfordPets, *bombay cat*), anatomical features like "sleek fur" ($\alpha = +0.64$) and "rounded ears" ($\alpha = +0.51$) are emphasized, while generic features like "black coat" ($\alpha = -0.10$) are suppressed. For object recognition (Caltech101, *accordion*), structural components like "multiple rows of keys" ($\alpha = +0.82$) and "metal grille" ($\alpha = +0.59$) receive high positive weights, while accessories like "straps" ($\alpha = -0.40$) are suppressed.

Unlike abstract prompt representations [9], our concept-based approach reveals which semantic attributes contribute to adversarial robustness through explicit weight values. Positive weights identify features the model emphasizes to maintain correct predictions under attack, while negative weights reveal attributes that are non-discriminative or vulnerable to adversarial manipulation. This interpretability enables targeted analysis of defense strategies based on domain-specific visual characteristics.

6. Ablation Study

To validate the necessity and sufficiency of learned concept weights, we conduct ablation experiments on all six datasets using full-shot trained models, analyzing (1) whether both positive and negative weights are essential, (2) how many concepts are required, and (3) whether learned magnitudes reflect genuine importance.

Q1: Are both positive and negative concept weights necessary for adversarial robustness?

We ablate concept weights by type: retaining only positive weights ($\alpha_{c,k} > 0$) while zeroing negative weights, and vice versa. Figure 3 shows the contribution of each concept type to robust accuracy across all datasets. Positive concepts consistently dominate, contributing 0.3-8.4pp to robust accuracy, while negative concepts contribute 0.1-3.4pp. On average, positive concepts contribute 2-4× more than negative concepts, validating that emphasis mechanisms are the primary drivers of robustness.

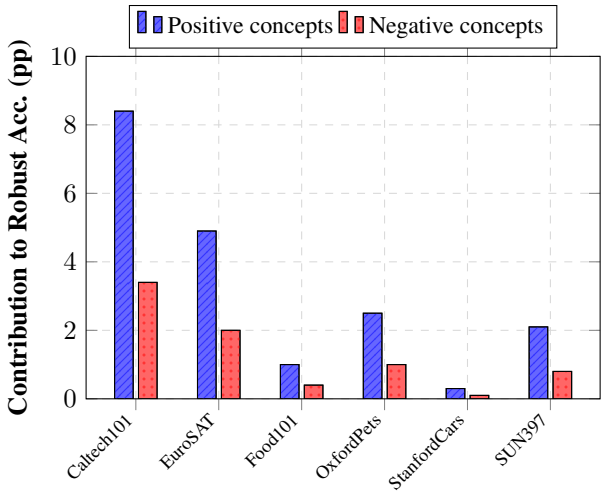

Figure 3. **Positive vs. negative concept contributions (Strategy C).** Robust accuracy contribution (in percentage points) when removing each concept type. Positive concepts (blue, diagonal pattern) are primary drivers, contributing 2-4× more than negative concepts (red, dotted pattern). However, negative concepts provide 0.4-3.4pp improvement, validating that suppression mechanisms are essential for robustness refinement.

However, negative concepts remain essential. Removing negative weights costs 0.4-3.4pp robust accuracy across datasets, with particularly notable losses on Caltech101 (-3.4pp), EuroSAT (-2.0pp), and OxfordPets (-1.0pp). This demonstrates that suppression mechanisms, while secondary to positive emphasis, are critical for refining decision boundaries. The model cannot achieve full robustness through positive concepts alone; it must also learn which features to suppress. Negative weights enable contextual disambiguation (e.g., suppressing sinks in kitchen scenes to avoid bathroom confusion, as shown in Table 1). These results validate our design choice to learn both positive and negative weights rather than constraining weights to be non-negative.

Q2: How many concepts are required to achieve competitive performance?

We systematically reduce concepts by retaining only the top-K by $|\alpha|$ magnitude per class while zeroing others. Figure 4 shows performance drops as K decreases from full model (K=6 for most datasets, K=10 for EuroSAT) to K=1. Remarkably, retaining only the top-3 concepts limits performance drops to $\leq$7pp for clean accuracy and $\leq$6pp for robust accuracy across all datasets, representing 90-95% of full model performance with only half the concepts.

The pattern is consistent across datasets despite varying baseline performance. Caltech101 experiences a 6.12pp clean drop and 5.62pp robust drop at K=3, while Food101

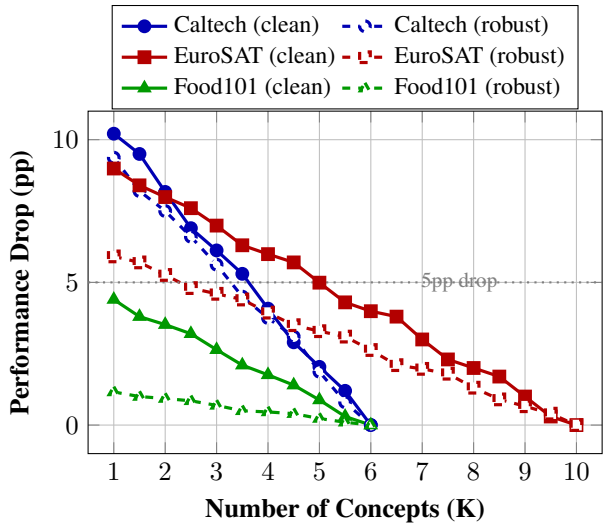

Figure 4. **Concept sufficiency analysis.** Performance drop (in percentage points) when reducing concepts from full model. Solid lines: clean accuracy. Dashed lines: robust accuracy. Top-3 concepts limit drop to $\leq$7pp across datasets. K=6 (or K=10 for EuroSAT) represents zero drop (full baseline).

shows only 2.64pp and 0.69pp drops respectively. EuroSAT, with K=10, requires slightly more concepts (K=5-6) to reach the 95% threshold, reflecting the greater semantic complexity of satellite imagery where fine-grained land-use distinctions demand more explicit semantic guidance. However, even for EuroSAT, K=3 retains 93-95% of full performance.

These results demonstrate concept sufficiency: the model does not require all 6-10 concepts per class to achieve competitive performance. Rather, a small subset of high-magnitude concepts captures most of the robustness benefit, while additional concepts provide diminishing returns. This finding has practical implications for deployment scenarios where computational efficiency matters, practitioners could use K=3 concept weights with minimal performance loss.

Q3: Do learned concept magnitudes reflect actual conceptual importance?

To validate that learned $|\alpha|$ magnitudes encode genuine semantic importance, we analyze the correlation between concept magnitude and usage frequency across classes. For each concept, we compute its mean $|\alpha|$ and measure how many classes use it significantly ($|\alpha| > 0.1$). Figure 5 reveals distinct domain-dependent patterns.

Three datasets exhibit strong positive correlation ($r > 0.7$): OxfordPets ($r = 0.94$), EuroSAT ($r = 0.88$), and StanfordCars ($r = 0.78$), validating that high-magnitude concepts are universally important. For instance, EuroSAT's top concept ("distinct rows or patterns of crops," $|\alpha| = 0.528$) is used by all 10 classes, while its lowest-magnitude concept ("geometric planting," $|\alpha| = 0.204$) appears in only 60% of classes. Similarly, "ruddy coat color" ($|\alpha| = 0.090$) is used by 41% of pet breeds, while "agouti pattern" ($|\alpha| = 0.062$) appears in only 8%.

In contrast, Caltech101 ($r = 0.07$) and SUN397 ($r = -0.30$) show weak or negative correlations, indicating class-specific concept usage. These diverse datasets—100 object categories ranging from accordions to yin-yang symbols, and 397 scene types from abbeys to zoos—require specialized concepts rather than universal features. The negative correlation in SUN397 reflects deliberate specialization: "Gothic or Romanesque architecture" ($|\alpha| = 0.105$) has high magnitude but applies to only 29% of scenes, serving as a strong discriminator for specific scene types.

Food101 shows moderate correlation ($r = 0.48$), reflecting semi-universal concepts: its pie/pastry-related features apply to many baked goods but not all 101 food classes. This intermediate pattern demonstrates the model's ability to leverage partially relevant concepts across related classes.

Together, these ablation studies validate our approach: (1) both emphasis and suppression are necessary, with positive concepts as primary drivers (Q1), (2) top-3 concepts achieve 90-95% performance, demonstrating concept sufficiency (Q2), and (3) learned magnitudes encode genuine semantic importance, with domain-dependent usage patterns reflecting dataset structure (Q3). Across all datasets, positive and negative weights maintain balanced magnitudes (average pos/neg ratio: 1.06×), with one exception: OxfordPets shows a 1.48× ratio, reflecting that breed identification relies more on distinctive feature presence than absence.

7. Limitations and Future Work

While our method demonstrates consistent improvements across most datasets, some datasets exhibit sensitivity to concept quality in extreme low-data regimes (e.g., OxfordPets 1-shot robustness). Future work could explore automated concept refinement or adaptive concept selection based on training dynamics. Additionally, evaluating robustness against adaptive attacks specifically designed to exploit concept structure would provide further insights into the security properties of semantic-based defenses.

8. Conclusion

We presented concept-enhanced adversarial prompt tuning, addressing limitations of existing methods through structured semantic concepts. By optimizing only combination weights for fixed concept embeddings, we achieve improved few-shot generalization (+19.58pp on EuroSAT 1-shot), enhanced adversarial robustness (up to 4× improvement on StanfordCars), and 98.7% parameter reduction (5,086 vs. 393K). Learned weights reveal interpretable patterns: spatial layout for satellite imagery, ingredient composition for food, and contextual discrimination for scenes (kitchens emphasize cabinets while suppressing sinks). While some datasets

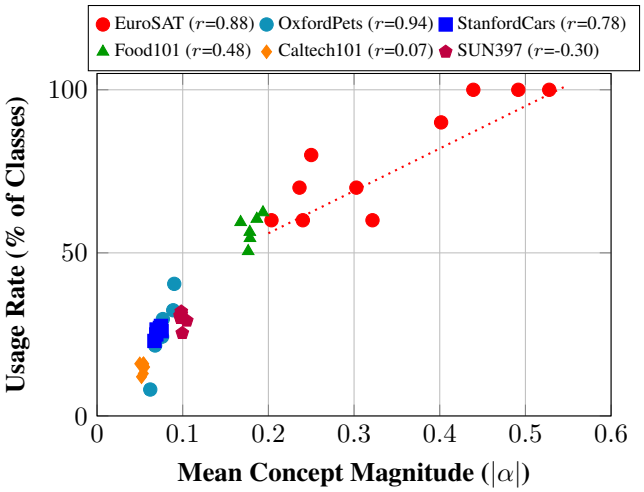

Figure 5. **Concept magnitude vs. usage correlation (Q3).** Each point represents one concept across all six datasets. Three datasets show strong positive correlation ($r > 0.7$): EuroSAT (red circles, $r = 0.88$), OxfordPets (cyan crosses, $r = 0.94$), and StanfordCars (blue squares, $r = 0.78$), validating that high-magnitude concepts are widely used. Food101 (green triangles, $r = 0.48$) shows moderate correlation. Caltech101 (orange diamonds, $r = 0.07$) and SUN397 (purple pentagons, $r = -0.30$) show weak/negative correlations: concepts are class-specific rather than universal. Dotted line shows EuroSAT's linear trend.

show sensitivity to concept quality in extreme low-data settings, our results demonstrate that incorporating interpretable semantic concepts successfully balances robustness, generalization, efficiency, and transparency, enabling security practitioners to understand and refine adversarial defenses.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, 35: 25005–25017, 2022. 2

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014. 1

[4] Hong-You Chen, Zhengfeng Lai, Haotian Zhang, Xinze Wang, Marcin Eichner, Keen You, Meng Cao, Bowen Zhang, Yinfei Yang, and Zhe Gan. Contrastive localized language-image pre-training. *arXiv preprint arXiv:2410.02746*, 2024. 1

[5] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 1

[6] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024. 2

[7] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 1

[8] Jonathan Krause, Jia Deng, Michael Stark, and Li Fei-Fei. Collecting a large-scale dataset of fine-grained cars. 2013. 1

[9] Lin Li, Haoyan Guan, Jianing Qiu, and Michael Spratling. One prompt word is enough to boost adversarial robustness for pre-trained vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24408–24419, 2024. 1, 2, 6

[10] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 1, 2

[11] Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. *arXiv preprint arXiv:2212.07016*, 2022. 1

[12] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 1

[13] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022. 2

[14] Xin Wang, Kai Chen, Jiaming Zhang, Jingjing Chen, and Xingjun Ma. Tapt: Test-time adversarial prompt tuning for robust inference in vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19910–19920, 2025. 2

[15] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 1

[16] Jiaming Zhang, Xingjun Ma, Xin Wang, Lingyu Qiu, Jiaqi Wang, Yu-Gang Jiang, and Jitao Sang. Adversarial prompt tuning for vision-language models. In *European conference on computer vision*, pages 56–72. Springer, 2024. 1, 2

[17] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022. 1, 2

[18] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1, 2, 5