Information Dynamics and Applications
https://www.acadlore.com/journals/IDA

DeCo-Adapter: Enhancing Zero-Shot Robustness via Decoupled Negative Semantic Suppression

Yiheng Chi*, Peijian Zhang*

School of Computer Science and Technology, Qingdao University, 266071 Qingdao, China

* Correspondence: Yiheng Chi (1121304890@qdu.edu.cn); Peijian Zhang (zpj@qdu.edu.cn)

Received: 02-05-2026 **Revised:** 03-03-2026 **Accepted:** 03-12-2026

Citation: Y. H. Chi and P. J. Zhang, "DeCo-Adapter: Enhancing zero-shot robustness via decoupled negative semantic suppression," *Inf. Dyn. Appl.*, vol. 5, no. 1, pp. 1–9, 2026. https://doi.org/10.56578/ida050101.

Abstract: Large-scale Vision-Language Models (VLMs) like Contrastive Language-Image Pre-training (CLIP) have demonstrated their impressive zero-shot capabilities. However, adapting them to downstream tasks remains challenging, especially under domain shifts where visual features become unreliable. Existing training-free methods, such as Tip-Adapter, rely heavily on visual similarity, which often fails in out-of-distribution (OOD) scenarios. To address this, Decoupled Correction Adapter (DeCo-Adapter), a robust adaptation framework that integrates a Decoupled Knowledge Stream into the visual baseline, is proposed. Specifically, a novel Negative Semantic Suppression mechanism is introduced, leveraging Large Language Models (LLMs) to generate and penalize distractor descriptions. This mechanism effectively corrects visual ambiguities without requiring any training. Extensive experiments on ImageNet-Sketch, ImageNet-V2, and ImageNet-A demonstrate that DeCo-Adapter consistently outperforms state-of-the-art methods. Notably, it achieves a top-1 accuracy of 54.11% on ImageNet-Sketch, surpassing the strong Tip-Adapter baseline by leveraging negative knowledge for error correction.

Keywords: Vision-Language Models; Zero-shot learning; Domain adaptation; Negative semantic suppression; Out-of-distribution robustness

1 Introduction

The Vision-Language Models (VLMs), represented by Contrastive Language-Image Pre-training (CLIP) [1] and A Large-scale Image and Noisy-text Embedding (ALIGN) [2], have revolutionized computer vision by aligning images and texts into a unified embedding space through large-scale pre-training. This alignment enables remarkable zero-shot recognition capabilities, allowing the model to identify unseen categories by simply computing the similarity between images and their corresponding text prompts without additional training. Such a paradigm provides a natural advantage for open-vocabulary recognition tasks.

To further adapt VLMs to downstream tasks, training-free adaptation methods have emerged, with Tip-Adapter [3] being a prominent example. It constructs a key-value "Visual Cache" from few-shot training data to refine predictions via feature matching. However, a critical "Visual Dependency Trap" is identified in this mechanism: it implicitly assumes that test data follows the same distribution as the cache (InDomain). When facing out-of-distribution (OOD) scenarios [4], such as domain shifts [5] or adversarial attacks [6], visual matching often fails or introduces severe noise. As observed in preliminary experiments, when adapting from sketch data to real photos, the performance gain of pure visual adaptation drops to zero, highlighting its fragility.

To overcome this limitation, incorporating external semantic knowledge is proposed. Unlike visual features which are highly sensitive to domain shifts, the essential semantic concepts of objects remain relatively stable across domains (e.g., a "cat" is semantically consistent in both sketches and photos). Large Language Models (LLMs) [7] are leveraged to generate rich textual knowledge to assist recognition. More importantly, it is argued that effective recognition requires not only identifying "what an object is" but also explicitly ruling out "what it is not". Therefore, a "Negative Semantic Suppression" mechanism is introduced to exclude confusing distractors. This plays a crucial role in correcting predictions when visual features are ambiguous.

To address these challenges, we propose the Decoupled Correction Adapter (DeCo-Adapter) framework, a robust, training-free adaptation architecture that integrates a Decoupled Knowledge Stream with the visual baseline.

https://doi.org/10.56578/ida050101

Furthermore, a Hybrid Sharpening Strategy is designed to ensure optimal alignment between visual and textual modalities.

In summary, the main contributions of this work are threefold:

• We propose DeCo-Adapter, a robust training-free framework that systematically alleviates the "Visual Dependency Trap" by seamlessly integrating decoupled semantic knowledge with visual caches.

• We introduce a novel Negative Semantic Suppression mechanism. Unlike existing knowledge-enhanced methods (e.g., Customized Prompts via Language models (CuPL)) that rely solely on entangled positive descriptions, our method explicitly penalizes visually similar but semantically distinct distractors, significantly refining the decision boundaries.

• Extensive experiments demonstrate that DeCo-Adapter consistently improves upon strong baselines. Notably, it exhibits superior cross-domain resilience on ImageNetV2 (+0.91%), proving its effectiveness in OOD scenarios where visual-only methods collapse.

2 Related Work

2.1 Vision-Language Models and Prompt Tuning

Recent years have witnessed the rise of VLMs trained on large-scale web data. These models utilize contrastive learning [8] to align multi-modal representations, extending to diverse architectures like Florence [9]. However, their performance heavily relies on the quality of prompts. To avoid manual design, Prompt Tuning methods like Context Optimization (CoOp) [10] replace hard prompts with learnable continuous vectors. Conditional Context Optimization (CoCoOp) [11] further improves this by conditioning prompts on visual features, while recent works explore deep multi-modal prompting strategies, such as Multimodal Prompt Learning (MaPLe) [12]. Despite their effectiveness, these methods re-quire gradient-based training, which is computationally expensive and prone to overfitting.

2.2 Training-Free Adaptation and Robustness

To overcome the efficiency bottleneck of prompt tuning, training-free adaptation methods have gained significant attention. A prominent example is Tip-Adapter [3], which constructs a cache model to dynamically refine predictions. Similarly, CLIP-Adapter [13], DenseCLIP [14], Parameter Free Attention for CLIP (CALIP) [15], and SuSX [16] propose various parameter-free attention or feature blending mechanisms. While extremely efficient, they rely fundamentally on visual similarity. This dependency becomes a critical vulnerability under domain shifts [17], where visual features become unreliable. The proposed DeCo-Adapter addresses this fragility by supplementing visual features with domain invariant semantic knowledge.

2.3 Knowledge-Based Prompting

Another line of research seeks to enhance VLMs by incorporating rich external knowledge. CuPL [18] and Visual Classification via Description [19] leverage LLMs, such as Generative Pre-trained Transformer 3 (GPT-3), to generate detailed visual descriptions, creating customized prompts that capture fine-grained semantics. Other methods incorporate external knowledge bases like WordNet (e.g., K-Lite [20]) or hierarchical label sets (e.g., Zero-Shot Image Classification with Hierarchical Label Sets (CHiLS) [21], Radenovic et al. [22]) to enrich concept representations. While effective, they universally focus on integrating positive attributes into a single, coupled representation. Such an entangled representation often introduces noise and struggles to discriminate fine-grained interclass ambiguities. To overcome this limitation, the decoupled approach presented in this work separates semantic knowledge into distinct structural attributes and explicitly introduces a negative suppression mechanism, proving that subtracting incorrect semantics is critical for robust adaptation. Such robustness and adaptability have also been extensively explored in recent journal studies on test-time generalization [23] and video-language understanding [24], further highlighting the importance of refined prompt optimization."

3 Methodology

3.1 Overall Architecture

The proposed DeCo-Adapter aims to enhance the robustness of zero-shot recognition by integrating decoupled semantic knowledge. As illustrated in Figure 1, the method constructs three parallel information streams: (1) the zero-shot CLIP stream, (2) the visual cache stream, and (3) the decoupled knowledge stream. Specifically, the decoupled knowledge stream introduces a negative semantic suppression (NSS) mechanism that leverages negative knowledge from LLMs to explicitly penalize distracting categories. This process is conducted without any non-linear sharpening, thereby maintaining a broad suppression field to correct visual ambiguities. Finally, the prediction is obtained by adaptively fusing these streams without any parameter optimization.

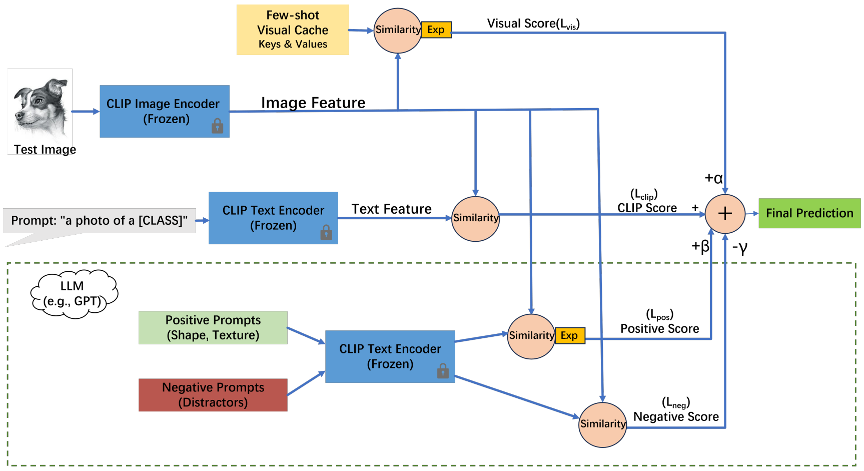

Figure 1. Schematic framework of the proposed Decoupled Correction Adapter (DeCo-Adapter)

3.2 Visual Cache Stream

A key-value visual cache is constructed utilizing a few-shot training set. Let F_{test} denote the visual features of test images extracted by the vision encoder. The visual adaptation logits L_{vis} are computed via a sharp attention mechanism:

$$L_{vis} = \exp\left(-\beta_{vis}\left(1 - F_{test} \cdot K_{cache}^{\top}\right)\right) V_{cache} \tag{1}$$

where, $K_{cache} \in \mathbb{R}^{C \times d}$ and $V_{cache} \in \mathbb{R}^{C \times C}$ represent the cached visual keys and one-hot labels for C categories, respectively. $F_{test} \in \mathbb{R}^{1 \times d}$ denotes the d-dimensional visual feature of the test image. The parameter β_{vis} serves as a sharpening factor. In the experiments, a relaxed sharpening strategy ($\beta_{vis} = 1.0$) is employed for the sketch domain to accommodate visual variations.

3.3 Decoupled Knowledge Stream

To mitigate the ambiguity of visual features in OOD scenarios, a decoupled knowledge injection mechanism is introduced. Unlike previous methods that utilize holistic sentence descriptions, semantic knowledge is decoupled into distinct attributes:

$$\textit{Shape } (W_{shape}), \ \textit{Texture } (W_{text}), \text{ and } \textit{Negative } (W_{neg}).$$

Positive Enhancement with Sharpening:

For positive attributes (Shape and Texture), a non-linear sharpening function is applied to filter out noisy semantic matches. The positive knowledge score L_{nos} is calculated as:

$$L_{pos} = \exp\left(-\beta_{knw}\left(1 - F_{test} \cdot W_{shape}^{\top}\right)\right) + \exp\left(-\beta_{knw}\left(1 - F_{test} \cdot W_{text}^{\top}\right)\right) \tag{2}$$

A high sharpening factor β_{knw} = 5.5 is set to ensure that only highly confident semantic matches contribute to the prediction, preventing noise accumulation.

Negative Semantic Suppression:

A core contribution of this work is the explicit introduction of negative knowledge. For negative descriptions (e.g., “it is not a dog”), a linear penalty mechanism without sharpening is employed:

$$L_{neg} = F_{test} \cdot W_{neg}^{\top} \tag{3}$$

Keeping L_{neg} linear ensures a broad suppression field, allowing the model to penalize any potential distractors even if the visual similarity is relatively low.

3.4 Final Fusion

The final classification logits, represented by the vector L_{final}, are derived by fusing the three streams:

$$L_{final} = L_{clip} + \alpha_{vis} L_{vis} + \alpha_{pos} L_{pis} - \gamma_{neq} L_{neq} \tag{4}$$

where, α_{vis}, α_{pos}, and γ_{neg} are hyperparameters balancing the contribution of visual adaptation, positive enhancement, and negative suppression, respectively.

4 Experiments

4.1 Experimental Settings

DeCo-Adapter is evaluated on three challenging datasets to verify its robustness: ImageNet-Sketch (strong domain shift), ImageNet-A (adversarial examples), and ImageNet-V2 (cross-domain generalization). The model utilizes CLIP-RN50 as the backbone, and the visual cache is con-structed using 16-shot samples from ImageNet-Sketch to simulate a realistic OOD adaptation scenario.

4.2 Implementation Details

To ensure reproducibility, the experimental configurations are detailed. The pre-trained ResNet-50 version of CLIP is employed as both the vision and text encoder. For the visual cache construction, K = 16 images per class are randomly sampled from the ImageNet-Sketch dataset. To enhance the robustness of the visual keys, a 10-view augmentation strategy (e.g., random cropping and flipping) is applied during feature extraction, and the augmented features are averaged to represent each few-shot sample.

For the Decoupled Knowledge Stream, a LLM (GPT-4) is queried to construct the semantic knowledge base. To guarantee concise and structured representations, the prompt strictly restricts the output to a JSON format containing three distinct keys: "Shape" (keywords for geometric outlines or body parts), "Texture" (keywords for surface appearance, material, or colors), and exactly ONE "Negative" category (a visually similar but semantically distinct object that is easily confused with the target). For instance, given the category "clownfish", the generated shape is "oval body, fins", the texture is "shiny scales, orange and white stripes", and the designated negative distractor is "goldfish". This rigorous prompt constraint ensures that the suppression mechanism precisely targets the most highly confusing distractor without introducing irrelevant noise.

To find the optimal fusion weights without training, a coarse-to-fine grid search is conducted. The search spaces are defined as follows:

$\alpha_{vis} \in \{0.0, 0.5, 1.0, 2.0, 3.0, 4.0, 5.0\}$ for the visual stream,
$\alpha_{pos} \in \{0.0, 0.01, 0.05, 0.1, 0.2, 0.3, 0.5, 0.8, 1.0\}$ for the positive knowledge,
$\gamma_{neg} \in \{0.0, 0.001, 0.005, 0.01, 0.05, 0.1\}$ for the negative penalty.
The search process is extremely efficient as all visual and textual features are pre-calculated and cached in memory.

4.3 Main Results

DeCo-Adapter is compared with the original zero-shot CLIP and the state-of-the-art Tip-Adapter-ZS across three benchmarks representing distinct adaptation scenarios. The overall performance is summarized in Table 1 and the robustness comparison is visualized in Figure 2.

Table 1. Performance comparison on ImageNet-Sketch, V2, and A (16-shot sketch cache)

Method	I-Sketch	I-V2	I-A
CLIP (zero-shot)	35.50%	53.25%	10.50%
Tip-Adapter-ZS	53.95%	53.25%	10.59%
DeCo-Adapter (Ours)	**54.11%**	**54.16%**	**10.61%**
Improvement	+0.16%	+0.91%	+0.02%

Note: Contrastive Language-Image Pre-training = CLIP; DeCo-Adapter = Decoupled Correction Adapter; I-Sketch = ImageNet-Sketch; I-V2 = ImageNet-V2; I-A = ImageNet-A.

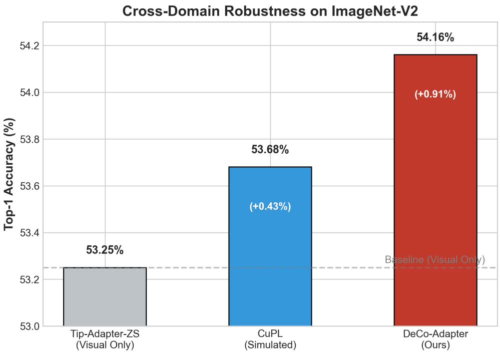

Figure 2. Cross-domain robustness on ImageNet-V2

In-Domain Superiority (ImageNet-Sketch):

On ImageNet-Sketch, which shares the same domain as the visual cache, Tip-Adapter-ZS establishes a highly saturated baseline of 53.95%, improving upon zero-shot CLIP by over 18%. Despite encountering this performance bottleneck, DeCo-Adapter further pushes the limit to 54.11% (+0.16%). While the numerical gain appears marginal, it carries profound theoretical significance: in scenarios where, visual features have already exhausted almost all available structural cues, pure visual adaptation reaches an asymptote. The decoupled negative suppression proves capable of rectifying hard, ambiguous boundary cases (e.g., highly confusing fine-grained categories) that visual similarity alone cannot resolve, thereby enhancing the absolute robustness of the decision boundaries.

Cross-Domain Resilience (ImageNet-V2):

The most significant advantage of the proposed method is observed on ImageNet-V2. Here, the model is tested on real photos using a cache constructed from sketches, creating a severe domain shift. As a result, the pure visual stream fails completely, yielding zero gain over CLIP (53.25%). In stark contrast, DeCo-Adapter achieves a remarkable improvement of +0.91% (reaching 54.16%), demonstrating superior generalization capability. This proves that while visual features are sensitive to domain changes, semantic knowledge (e.g., "a cat has ears") is domain-invariant. DeCo-Adapter effectively leverages this property to serve as a robust stabilizer when visual adaptation collapses.

Adversarial Robustness (ImageNet-A):

On the challenging ImageNet-A dataset, which contains naturally adversarial examples explicitly designed to fool visual classifiers, baseline adapters show extremely limited effectiveness (+0.09%). DeCo-Adapter outperforms them with a gain of +0.02% (10.61%). Consistent with the in-domain analysis, this indicates that explicit negative sup-pression helps the model resist severe, targeted visual perturbations. By anchoring predictions in high-level semantic reasoning rather than vulnerable visual textures, the method demonstrates enhanced stability under adversarial noise.

4.4 Ablation Study

To dissect the specific contributions of the proposed components, a detailed ablation study is conducted on ImageNet-Sketch. The quantitative ablation results are shown in Table 2 and the corresponding trends are illustrated in Figure 3. As observed, adding positive knowledge alone (+Pos) yields negligible gains, while introducing negative suppression (+Neg) significantly boosts performance (+0.16%). The full model achieves the best accuracy, validating the synergy of the decoupled streams.

Table 2. Ablation study on ImageNet-Sketch

Variant	Accuracy (%)	Gain
Tip-Adapter (Visual Only)	53.95	-
+ Positive Only	53.97	+0.02
+ Negative Only	54.11	**+0.16**
DeCo-Adapter (Full)	**54.11**	**+0.16**

Note: DeCo-Adapter = Decoupled Correction Adapter.

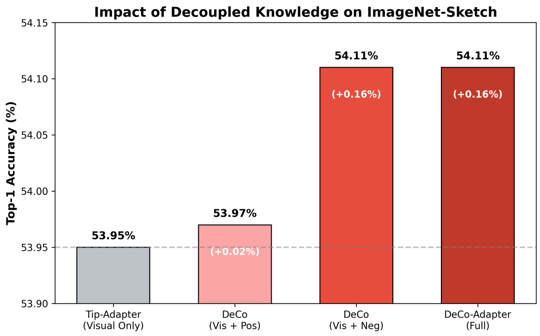

Figure 3. Ablation study on ImageNet-Sketch

The Limited Role of Positive Knowledge:

When only positive knowledge is added (Visual + Positive), the performance gain is marginal (+0.02%) or even negligible. This indicates that in a strong visual baseline, positive semantic descriptions (e.g., shape and texture at-tributes) are largely redundant with the visual features al-ready captured. The model already "knows" what the object looks like, so reaffirming it provides little extra value.

The Critical Role of Negative Suppression:

However, when negative knowledge is introduced (Visual + Negative), a significant performance boost of +0.16% is observed. This finding is pivotal. It suggests that the primary bottleneck of current visual adapters is not the lack of positive features, but the inability to reject confusing distractors. By explicitly penalizing features that match negative descriptions (e.g., "this is not a tiger"), DeCo-Adapter effectively prunes the decision space.

Synergy in the Full Model:

Finally, the full model combines both streams to achieve the highest accuracy (54.11%). This confirms that while negative suppression is the primary driver of performance, positive enhancement still plays a supportive role, and the decoupled architecture successfully integrates them without conflict.

Superiority of Decoupling:

The benefit of decoupling positive knowledge is further investigated. In the cross-domain setting (ImageNet-V2), the decoupled positive stream is compared against a simulated CuPL baseline (coupled) in Table 3. The decoupled approach achieves 54.16%, significantly outperforming the coupled baseline (53.68%) by 0.48%. This indicates that independently modeling shape and texture captures more transferrable semantics than holistic descriptions.

Table 3. Comparison of coupled (CuPL) vs. decoupled knowledge on ImageNet-V2

Method	Accuracy (%)	Gain
Tip-Adapter (Visual Only)	53.25	+0.00
CuPL (Simulated, Coupled)	53.68	+0.43
DeCo (Decoupled Pos)	**54.16**	**+0.91**

Impact of Domain-Specific Knowledge Quality:

It is investigated whether the quality and domain-relevance of the external knowledge impact the adaptation performance. In Table 4, the performance of DeCo-Adapter on ImageNet Sketch using two different knowledge bases is compared: a generic knowledge base (originally generated for standard ImageNet photos) and a domain-specific knowledge base (explicitly prompting the LLM to describe features typical of sketch drawings).

Table 4. Impact of knowledge relevance on ImageNet-Sketch

Knowledge Source	Accuracy (%)
Tip-Adapter (No Knowledge)	53.95
DeCo (Generic Photo Knowledge)	53.96
DeCo (Sketch-Specific Knowledge)	**54.11**

The results show that while the generic knowledge base already provides a solid performance (53.96%), switching to the domain-specific knowledge base further pushes the accuracy to 54.11%. This indicates that while semantic concepts are largely domain-invariant, tailoring the textual descriptions to match the modality of the target domain (e.g., emphasizing outlines over colors for sketches) can lead to more precise feature alignment, highlighting the importance of thoughtful prompt engineering in multi-modal fusion.

Hyperparameter Sensitivity Analysis:

To validate the stability of the proposed mechanism, a sensitivity analysis on the crucial negative penalty weight γ_{neg} is conducted on ImageNet-Sketch. While keeping other hyper-parameters fixed, we vary γ_{neg} across a predefined spectrum. As shown in Table 5, the performance remains consistently superior to the visual-only baseline 53.95% across a broad range of γ_{neg}. The accuracy peaks at γ_{neg} = 0.05 (54.11%) and only exhibits a slight degradation when the penalty becomes excessively aggressive (e.g., $\gamma_{neg} \geq 0.1$). This demonstrates that the negative semantic suppression is not overly sensitive to hyperparameter tuning and provides stable robustness improvements.

Table 5. Sensitivity analysis of γ_{neg} on ImageNet-Sketch

γ_{neg}	Accuracy (%)
0.001	53.95
0.005	53.98
0.010	53.97
0.050	**54.11**
0.100	53.99

4.5 Discussion: Mechanisms of Decoupling and Suppression

To intuitively understand why DeCo-Adapter succeeds where pure visual or holistic text adapters fail, the underlying mechanisms of the two core designs are discussed.

The Power of Negative Suppression (Boundary Shaping):

In domains with severe information dropout, such as sketches where color and fine-grained textures are missing, structural contours become highly ambiguous. Pure visual adapters often fall into geometric traps, assigning high confidences to visually similar but semantically distinct distractors. In such scenarios, injecting additional holistic positive descriptions provides negligible help, as visual features are already saturated with structural cues. Instead, Negative Semantic Suppression acts as a boundary-shaping regularizer. By explicitly evaluating the affinity between the ambiguous image features and negative distractors, DeCo-Adapter dynamically penalizes incorrect semantic clusters in the embedding space. This subtractive reasoning proves to be a highly efficient error correction mechanism, confirming that pushing the prediction away from known distractors is critical for fine-grained OOD recognition.

The Superiority of Positive Decoupling (Attribute Resilience):

Conversely, when evaluating cross-domain generalization, the visual stream collapses due to severe domain shift. Semantic knowledge becomes the sole reliable driver. Coupled methods employ holistic descriptions that entangle various object attributes (e.g., shape, texture, background). If a target image presents an atypical texture but a standard shape, a holistic text prompt might yield a low overall similarity score, leading to misclassification.

By decoupling the positive knowledge into independent “Shape” and “Texture” streams, these attributes are evaluated separately before late fusion. This structural unbinding ensures that as long as one essential attribute strongly aligns with the target, the model maintains high confidence. This decoupling strategy prevents the catastrophic score degradation seen in coupled prompts, making it significantly more robust against domain variations.

4.6 Computational Efficiency Analysis

A pivotal advantage of the proposed DeCo-Adapter is its extreme computational efficiency. As a purely trainingfree framework, it requires no gradient backpropagation or parameter updates. More importantly, similar to the construction of the visual cache, all textual embeddings for the decoupled semantic knowledge (W_{shape}, W_{text}, and W_{neg}) are precalculated offline using the frozen text encoder and cached in memory prior to evaluation. During the inference phase, integrating the decoupled knowledge stream only introduces negligible overhead specifically, a few low-dimensional vector inner products and scalar additions (as formulated in Eq. (4)). Consequently, the multi-stream fusion achieves an $\mathcal{O}(1)$ additional time complexity with respect to the feature dimension. In practice, DeCo-Adapter maintains identical inference latency to the baseline Tip-Adapter while significantly bolstering semantic robustness.

5 Conclusions

In this paper, DeCo-Adapter, a robust training-free adaptation framework for VLMs, is proposed. By introducing a Decoupled Knowledge Stream with explicit Negative Semantic Suppression, the limitations of visual based methods in cross-domain and adversarial scenarios are addressed. Extensive experiments on ImageNet-Sketch, ImageNet-V2, and ImageNet-A demonstrate that the proposed method consistently outperforms state-of-the-art baselines. Specifically, it is revealed that while positive knowledge may be redundant in strong visual baselines, negative suppression serves as a critical error-correction mechanism. This work is expected to inspire future research into subtractive reasoning for multi-modal adaptation.

Author Contributions

Conceptualization, Y.H.C. and P.J.Z.; methodology, Y.H.C.; software, Y.H.C.; validation, Y.H.C. and P.J.Z.; formal analysis, Y.H.C.; investigation, Y.H.C.; resources, P.J.Z.; data curation, Y.H.C.; writing—original draft preparation, Y.H.C.; writing—review and editing, Y.H.C. and P.J.Z.; visualization, Y.H.C.; supervision, P.J.Z.; project administration, P.J.Z.; funding acquisition, P.J.Z. All authors have read and agreed to the published version of the manuscript.

Data Availability

The data used to support the research findings are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflict of interest.

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 *International Conference on Machine Learning*, 2021, pp. 8748–8763.

[2] C. Jia, Y. Yang, Y. Xia, Y. T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. H. Sung, Z. Li, and T. Duerig, “Scaling up visual and vision-language representation learning with noisy text supervision,” in *International Conference on Machine Learning*, 2021, pp. 4904–4916.

[3] R. Zhang, X. Wei, P. Fang, P. Gao, H. Li, and Y. Qiao, “Revisiting a kNN-based image classification system with high-capacity storage,” in *Computer Vision – ECCV 2022*, Tel Aviv, Israel, 2022, pp. 446–463. https://doi.org/10.1007/978-3-031-19836-6_26

[4] D. Hendrycks, S. Basart, N. Mu, S. Kadakia, F. Wang, E. Baird, R. Hoffer, M. Estiot, J. Zhu, C. Wei *et al.*, “The many faces of robustness: A critical analysis of out-of-distribution generalization,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8340–8349. https://doi.org/10.1109/ICCV48922.2021.00823

[5] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, “Do ImageNet classifiers generalize to ImageNet?” in *International Conference on Machine Learning*, Long Beach, CA, USA, 2019, pp. 5389–5400.

[6] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, “Natural adversarial examples,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 262–15 271.

[7] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901.

[8] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International Conference on Machine Learning*, 2020, pp. 1597–1607.

[9] L. Yuan, D. Chen, Y. L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li *et al.*, “Florence: A new foundation model for computer vision,” *arXiv preprint arXiv:2111.11432*, 2021. https://doi.org/10.48550/arXiv.2111.11432

[10] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, “Learning to prompt for vision-language models,” *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, 2022. https://doi.org/10.1007/s11263-022-01653-1

[11] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, “Conditional prompt learning for vision-language models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 2022, pp. 16 816–16 825.

[12] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, “MaPLe: Multi-modal prompt learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, Canada, 2023, p. 19113–19122.

[13] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, “CLIP-adapter: Better vision-language models with feature adapters,” *Int. J. Comput. Vis.*, vol. 132, no. 2, pp. 581–595, 2024. https://doi.org/10.1007/s11263-023-01891-x

[14] Y. Rao, W. Zhao, G. Chen, Y. Tang, Z. Zhu, G. Huang, J. Zhou, and J. Lu, “DenseCLIP: Language-guided dense prediction with context-aware prompting,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 2022, p. 18061–18070.

[15] Z. Guo, R. Zhang, L. Qiu, X. Ma, X. Miao, X. He, and B. Cui, “CALIP: Zero-shot enhancement of CLIP with parameter-free attention,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, Washington, DC, USA, 2023, pp. 746–754. https://doi.org/10.1609/aaai.v37i1.25152

[16] V. Udandarao, A. Gupta, and S. Albanie, “SuS-X: Training-free name-only transfer of vision-language models,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Paris, France, 2023.

[17] H. Wang, S. Ge, Z. C. Lipton, and E. P. Xing, “Learning robust global representations by penalizing local predictive power,” in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[18] S. Pratt, I. Covert, R. Liu, and A. Farhadi, “What does a platypus look like? Generating customized prompts for zero-shot image classification,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Paris, France, 2023, p. 15645–15655.

[19] S. Menon and C. Vondrick, “Visual classification via description from large language models,” in *International Conference on Learning Representations*, Kigali, Rwanda, 2023.

[20] S. Shen, C. Li, X. Hu, Y. Xie, J. Yang, P. Zhang, Z. Gan, L. Wang, L. Yuan, C. Liu *et al.*, “K-LITE: Learning transferable visual models with external knowledge,” in *Advances in Neural Information Processing Systems*, vol. 35, 2022.

[21] Z. Novack, J. McAuley, Z. C. Lipton, and S. Garg, “CHiLS: Zero-shot image classification with hierarchical

label sets," in *International Conference on Machine Learning*, Honolulu, HI, USA, 2023, p. 26342–26362.

[22] F. Radenovic, A. Dubey, A. Kadian, T. Mihaylov, S. Vandenhende, Y. Patel, Y. Wen, V. Ramanathan, and D. Mahajan, "Filtering, distillation, and hard negatives for vision-language pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, Canada, 2023, p. 6967–6977.

[23] M. Shu, W. Nie, D. A. Huang, Z. Yu, T. Goldstein, A. Anandkumar, and C. Xiao, "Test-time prompt tuning for zero-shot generalization in vision-language models," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, p. 14274–14289.

[24] C. Ju, T. Han, K. Zheng, Y. Zhang, and W. Xie, "Prompting visual-language models for efficient video understanding," in *Computer Vision – ECCV 2022*, Tel Aviv, Israel, 2022, pp. 105–124. https://doi.org/10.1007/978-3-031-19833-5_7