Training-Only Heterogeneous Image-Patch-Text Graph Supervision for Advancing Few-Shot Learning Adapters

Supplementary Material

To support the findings presented in the main paper, this document details more experiments, ablation studies, and visualizations. Specifically, we cover:

1. A comparative analysis of few-shot classification performance against several state-of-the-art (SOTA) prompt-based methods. (Section S1)
2. Additional visualization of the dataset for few-shot accuracy curves comparing various adapter-based methods.
3. t-SNE visualizations comparing the feature distributions of Zero-shot CLIP logits, Tip-Adapter-F logits, and Graph logits. (Section S2)
4. More qualitative visualizations of discriminative node filtering. (Section S3)
5. Ablation studies evaluating different backbone architectures compared to Tip-Adapter-F [6]. (Section S4)
6. A detailed analysis of computational complexity, feature dimensionality, and model parameters. (Section S5)
7. Hyperparameter sensitivity analysis. (Section S6)
8. An investigation into the impact of varying the number of multiscale patches/nodes. (Section S7)
9. Quantitative visualizations of the Cache and graph model. (Section S8)
10. Ablation studies evaluating text nodes in the graph topology and MGT vs. Transformer performance. (Section S9)

S1. Comparison with SotA Prompt Learning

While the main manuscript focuses on adapter-based baselines, we provide here a comprehensive comparison against SOTA prompt learning methods, including static (CoOp [8]), conditional (CoCoOp [7], KgCoOp [5]), and gradient-aligned (ProGrad [9], MaPLe [2], PLOT++ [1]) approaches. Table S1 details the performance across 11 benchmarks.

Low-Shot Robustness (1-4 Shots). Prompt-based methods often struggle in extreme low-data regimes due to the difficulty of optimizing continuous vectors without overfitting. Our method significantly mitigates this issue.

- **1-Shot:** We achieve an average accuracy of **72.2%**, establishing a new SOTA by surpassing the optimal-transport method PLOT++ (70.7%) by **+1.5%** and the multi-modal MaPLe (69.3%) by **+2.9%**.
- **2-Shot:** The gap widens against trajectory-constrained methods like ProGrad (72.4%) and MaPLe (72.6%), with our method reaching **75.0%**. This indicates that our *Modality-aware Graph Teacher* extracts robust prototypes even from singular examples, whereas prompt gradients often settle into sharp, non-generalizable minima when supervised by sparse data.
- **4-Shot:** This performance advantage is sustained as support size increases. Our method achieves **77.9%**, maintaining a clear lead over PLOT++ (76.9%) and MaPLe (75.8%). This demonstrates that even as sufficient data becomes available to stabilize prompt optimization, our structure-aware supervision yields superior feature alignment compared to global prompt tuning.

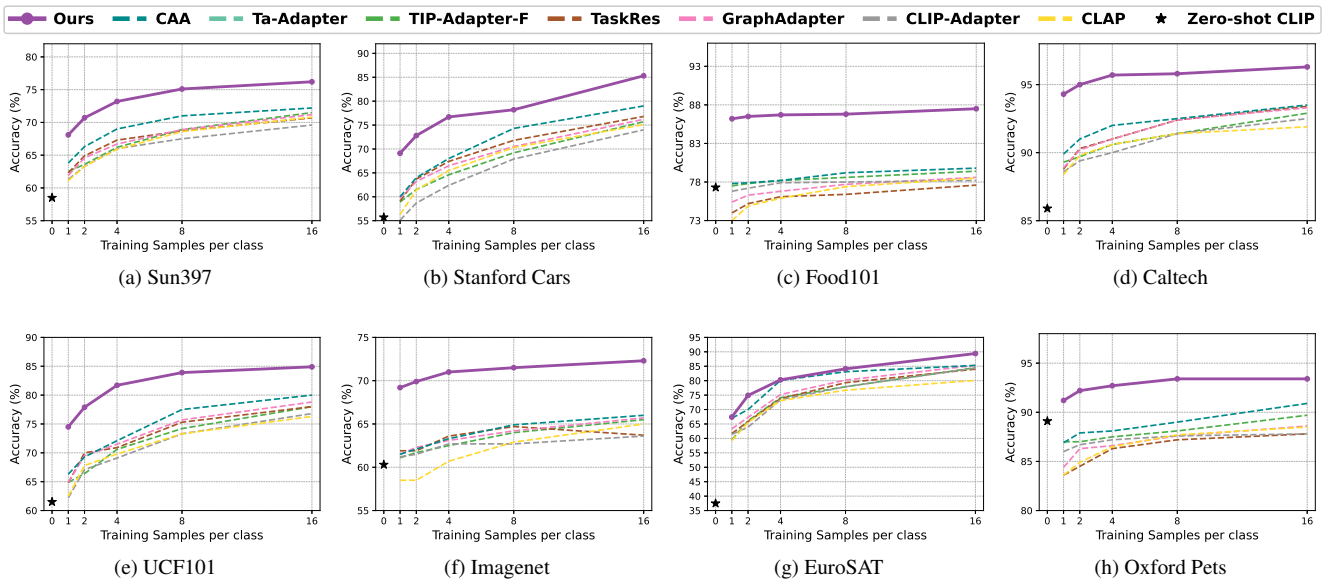
High-Shot Scalability and Fine-Grained Tasks (8-16 Shots) Our method exhibits sustained superiority in high-shot regimes, achieving an average accuracy of 82.3% at the 16-shot level, marginally leading the strongest competitor, PLOT++ (82.1%). This advantage is most pronounced in fine-grained tasks requiring high local discriminative power. For instance, on Aircraft dataset, we reach 48.4% (matching MaPLe and outperforming ProGrad's 43.0%), and on Stanford Cars, we achieve 85.3% (surpassing PLOT++'s 84.5%). This consistent edge is maintained across the full fine-grained suite, posting 98.3% on Flowers and a competitive 73.6% on the challenging task in DTD (outperforming MaPLe's 71.3

Zero-Overhead Inference Crucially, while top-performing prompt methods like MaPLe and PLOT++ incur significant overhead due to deep prompt coupling or optimal transport problems, our method distills these performance gains into a static key-value cache. This innovative design allows us to maintain the zero-overhead inference speed of a standard Tip-Adapter-F, effectively outperforming computationally heavier prompt baselines without sacrificing efficiency.

S2. Qualitative Analysis of Embedding Separability via t-SNE

To qualitatively validate the effectiveness of our asymmetric supervision, we visualize the logit distributions using t-Distributed Stochastic Neighbor Embedding (t-SNE). We compare the feature manifolds of our method against the baseline Zero-shot CLIP [4] and Tip-Adapter-F [6] in a 16-shot setting. To ensure a rigorous evaluation of decision boundary refinement, we selected a subset of 8 randomly selected classes per dataset.

The resulting visualizations (Figures S2, S3, and S4), illustrate the impact of our Training-Only modality-aware Graph Teacher. While global-feature baselines (CLIP and Tip-Adapter-F) often exhibit diffuse clustering and significant inter-class overlap among semantically similar categories, our method achieves superior intra-class compactness and inter-class separability. This improvement



(a) Sun397 (b) Stanford Cars (c) Food101 (d) Caltech

(e) UCF101 (f) Imagenet (g) EuroSAT (h) Oxford Pets

Figure S1. Performance trajectories of different adapter-based architectures. The curves show the few-shot accuracy of the model on several additional datasets, complementing the main results presented in the paper.

confirms that the fine-grained, cross-modal reasoning performed by the Modality-aware Graph Transformer (MGT) is successfully distilled into the student adapter's key-value cache. The quantitative improvement is evidenced by the Silhouette scores; for example, on the EuroSAT dataset, our approach achieves a score of 0.522 compared to the baseline's -0.043, indicating that our graph-guided teacher forcing effectively rectifies decision boundaries even in domain-shifted remote sensing tasks.

S3. Qualitative Visualization of Discriminative Node Filtering

The provided visualization further confirms the necessity of discriminative node filtering to ensure the quality of the teacher's supervisory signal. Our Top-$\mathbb{N}$ filtering mechanism is designed to mitigate the inherent feature dilution that occurs when aggregating raw patch features, which is especially critical when dealing with object-centric datasets like Caltech-101 that contain diverse object scales and cluttered backgrounds. The visualization (Figure S5) shows that the learned filtering weights, applied after cross-modal reasoning by the GNN Teacher, effectively differentiate between foreground and background elements. Patches covering the central object and defining features, such as the scorpion's body and claws, the scissors' blades and handles, and the distinct outline of the starfish, receive high activation scores (highlighted in Green). Conversely, non-informative patches encompassing uniform background areas are suppressed (highlighted in Blue). This selective mechanism consolidates contextualized information into the final visual characteristic (f_{graph}) of the teacher's classification head, concentrating primarily on discriminative foreground cues. By generating a cleaner, foreground-centric representation, the filtering enhances the discriminative power of the graph logits, leading to a more robust knowledge transfer during asymmetric supervision.

S4. Backbone Scaling and Efficiency Analysis

We validate the scalability of our approach by evaluating its few-shot classification accuracy across four distinct CLIP backbones: ResNet-101 (RN101) and the ViT-B/32, ViT-B/16, and ViT-L/14 Vision Transformer variants, comparing our method against the Tip-Adapter-F baseline (Base). As shown in Table S2, **TOGA** consistently improves accuracy over the Base model across all architectures and shot settings, demonstrating its broad applicability and robust use of features from diverse encoder structures. As the backbone capacity increases from ViT-B/32 to ViT-L/14, the overall accuracy of our method increases progressively. The highest performance is achieved by the ViT-L/14 model, which obtains an 8-shot average accuracy of $\mathbf{86.9}\%$. The gains are particularly significant on challenging tasks, such as improving Aircraft accuracy from 53.0% (Base) to $\mathbf{57.0}\%$ (Ours) in 8-shot with ViT-L/14.

S5. Computational Profile and Asymmetric Cost

The efficiency of our approach comes from its asymmetric supervision design, which minimizes the deployment cost.

Table S1. Comparison of few-shot classification accuracy (%) on 11 benchmark datasets. We evaluate our method against several SOTA prompt-based approaches. The best performance in each shot-group is marked in **bold**. Our results are highlighted in gray. **Dataset abbreviations:** INet (ImageNet), SUN (SUN397), Air (Aircraft), Euro (EuroSAT), Cars (Stanford Cars), Food (Food101), Pets (OxfordPets), Flow (Flowers102), Cal (Caltech101), DTD (Describable Textures), UCF (UCF101).

Shots	Method	Venue	INet	SUN	Air	Euro	Cars	Food	Pets	Flow	Cal	DTD	UCF	Avg
0	CLIP [3]	ICML'22	66.7	62.6	24.7	47.5	65.3	86.1	89.1	71.4	92.9	43.6	66.7	65.1
1	CoOp [8]	IJCV'22	68.0	67.3	26.2	50.9	67.1	82.6	90.3	72.7	93.2	50.1	70.7	67.2
	CoOp [8]	IJCV'22	65.7	67.0	20.8	56.4	67.5	84.3	90.2	78.3	92.5	50.1	71.2	67.6
	CoCoOp [7]	CVPR'22	**69.4**	**68.7**	28.1	55.4	67.6	84.9	**91.9**	73.4	94.1	52.6	70.4	68.8
	PLOT++ [1]	ICLR'23	66.4	66.7	28.6	65.4	68.8	86.1	91.8	80.4	94.3	54.5	74.3	70.7
	KgCoOp [5]	CVPR'23	68.9	68.4	26.8	61.9	66.7	**86.4**	92.1	74.7	94.2	52.7	72.8	69.6
	MaPLe [2]	CVPR'23	62.6	64.7	26.7	71.8	66.6	80.5	89.1	83.3	92.5	52.1	71.8	69.3
	ProGrad [9]	ICCV'23	67.0	67.0	28.8	57.0	68.2	84.9	91.4	80.9	93.5	52.8	73.3	69.5
	TOGA (Ours)	-	69.2	68.1	**31.0**	**67.4**	**69.1**	86.2	91.2	**88.2**	**94.3**	**55.2**	**74.5**	**72.2**
2	CoOp (4) [8]	IJCV'22	68.7	68.0	28.1	66.2	70.5	82.6	89.9	80.9	93.0	53.7	73.5	70.5
	CoOp (16) [8]	IJCV'22	67.0	67.0	25.9	65.1	70.4	84.4	89.9	88.0	93.1	54.1	74.1	70.8
	CoCoOp [7]	CVPR'22	**70.1**	69.4	29.3	61.8	68.4	85.9	91.9	77.8	94.4	52.3	73.4	70.4
	PLOT++ [1]	ICLR'23	68.2	68.0	31.1	76.8	**73.1**	86.3	92.2	89.8	94.6	56.7	76.7	74.0
	KgCoOp [5]	CVPR'23	69.6	69.6	28.0	69.2	68.2	**86.6**	**92.3**	79.8	94.5	55.3	74.6	71.6
	MaPLe [2]	CVPR'23	65.1	67.1	30.9	**78.3**	71.6	81.4	90.8	88.9	93.9	55.5	74.6	72.6
	ProGrad [9]	ICCV'23	69.1	69.0	31.1	66.3	72.4	84.8	91.5	87.5	93.6	56.0	75.6	72.4
	TOGA (Ours)	-	69.9	**70.7**	**34.8**	74.9	72.8	86.5	92.2	**91.9**	**95.0**	**58.4**	**77.9**	**75.0**
4	CoOp [8]	IJCV'22	69.7	70.6	29.7	65.8	73.4	83.5	92.3	86.6	94.5	58.5	78.1	73.0
	CoOp [8]	IJCV'22	70.6	69.7	30.9	69.7	74.4	84.5	92.5	92.2	94.5	59.5	77.6	74.2
	CoCoOp [7]	CVPR'22	70.8	70.4	30.6	61.7	69.5	86.3	92.7	81.5	94.8	55.7	75.3	71.8
	PLOT++ [1]	ICLR'23	70.4	71.7	35.2	83.2	76.2	86.4	92.5	92.9	95.1	62.4	79.7	76.9
	KgCoOp [5]	CVPR'23	70.2	71.5	32.2	71.8	69.5	86.9	92.6	87.0	95.0	58.7	77.6	73.9
	MaPLe [2]	CVPR'23	67.7	70.6	34.8	**84.5**	75.3	81.7	91.9	92.6	94.4	61.0	78.4	75.8
	ProGrad [9]	ICCV'23	**71.3**	71.7	34.1	69.6	75.0	85.4	92.1	91.1	94.4	59.7	77.9	73.9
	TOGA (Ours)	-	71.0	**73.2**	**38.3**	80.3	**76.7**	**86.7**	**92.7**	**96.4**	**95.7**	**64.5**	**81.7**	**77.9**
8	CoOp (4) [8]	IJCV'22	70.8	72.4	37.0	74.7	76.8	83.3	92.1	95.0	94.7	63.7	79.8	76.4
	CoOp (16) [8]	IJCV'22	70.6	71.9	38.5	77.1	79.0	82.7	91.3	94.9	94.5	64.8	80.0	76.8
	CoCoOp [7]	CVPR'22	70.8	71.5	32.4	69.1	70.4	**87.0**	93.3	86.3	94.9	60.1	75.9	73.8
	PLOT++ [1]	ICLR'23	71.3	73.9	41.4	**88.3**	**81.2**	86.5	93.0	95.4	95.5	66.4	82.8	79.6
	KgCoOp [5]	CVPR'23	70.2	72.6	34.8	73.9	72.8	**87.0**	93.0	91.5	95.1	65.6	80.0	76.0
	MaPLe [2]	CVPR'23	70.3	73.2	42.0	87.7	79.4	83.6	92.5	95.8	95.2	66.5	81.3	78.9
	ProGrad [9]	ICCV'23	71.3	73.0	37.7	77.8	78.7	86.1	92.2	95.0	94.8	63.9	80.5	77.4
	TOGA (Ours)	-	**71.5**	**75.1**	**44.2**	84.1	78.2	86.8	**93.4**	**97.3**	**95.8**	**69.6**	**83.9**	**80.0**
16	CoOp [8]	IJCV'22	71.5	74.6	40.1	83.5	79.1	85.1	92.4	96.4	95.5	69.2	81.9	79.0
	CoOp (16) [8]	IJCV'22	71.9	74.9	43.2	85.0	82.9	84.2	92.0	96.8	95.8	69.7	83.1	80.0
	CoCoOp [7]	CVPR'22	71.1	72.6	33.3	73.6	72.3	**87.4**	93.4	89.1	95.1	63.7	77.2	75.3
	PLOT++ [1]	ICLR'23	72.6	76.0	46.7	92.0	84.5	87.1	**93.5**	97.5	96.0	71.4	**85.3**	82.1
	KgCoOp [5]	CVPR'23	70.4	73.3	36.5	76.2	74.8	87.2	93.2	93.4	95.2	68.7	81.7	77.3
	MaPLe [2]	CVPR'23	**72.3**	75.5	**48.4**	**92.3**	83.5	85.3	92.8	97.0	96.0	71.3	85.0	81.8
	ProGrad [9]	ICCV'23	72.1	75.1	43.0	83.6	82.9	85.8	92.8	96.6	95.9	68.8	82.7	79.9
	TOGA (Ours)	-	**72.3**	**76.2**	**48.4**	89.4	**85.3**	87.5	93.4	**98.3**	**96.3**	**73.6**	84.9	**82.3**

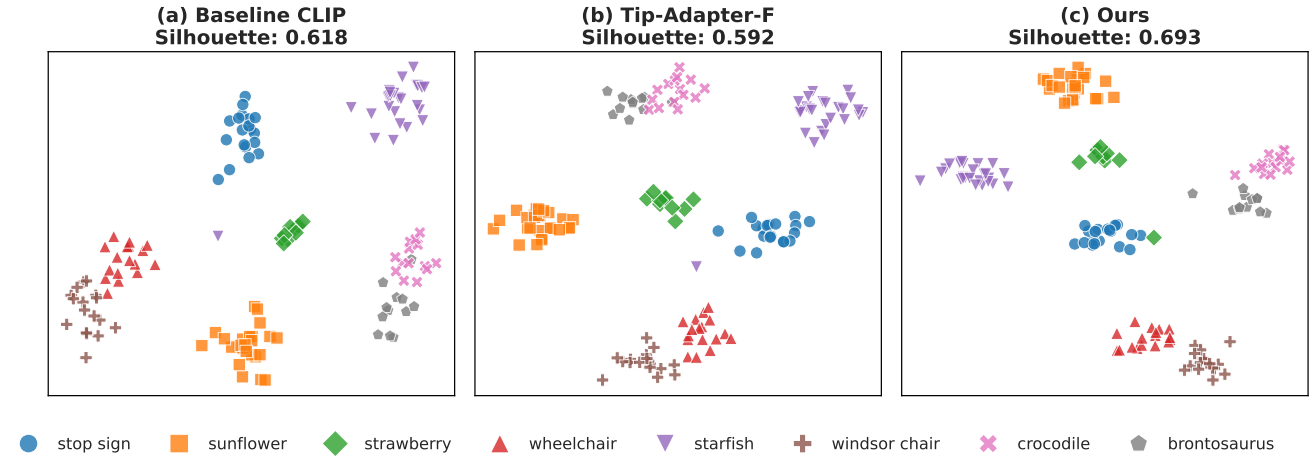


Figure S2. t-SNE visualization of logits on the Caltech101 dataset (16-shot). The plots depict the separation of 8 object categories, ranging from distinct objects like stop sign to complex shapes like windsor chair. (a) Baseline CLIP provides a strong initial separation (0.618). (b) Tip-Adapter-F introduces minor variance, slightly reducing the score to 0.592. (c) Ours further refines the manifold, tightening the clusters for difficult classes (e.g., crocodile and brontosaurus) and achieving the highest distinctiveness with a Silhouette score of 0.693. This highlights the benefit of our discriminative node filtering in suppressing background noise.

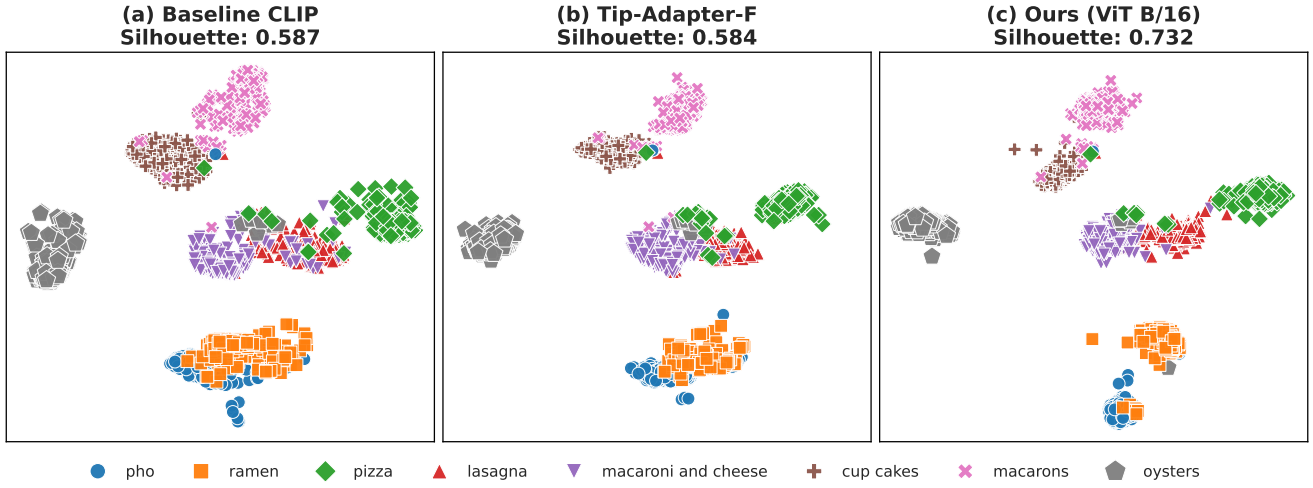


Figure S3. This visualization highlights the model's ability to distinguish between fine-grained culinary categories. (a) Baseline CLIP and (b) Tip-Adapter-F show comparable performance (≈ 0.58), with noticeable overlap between visually similar classes (e.g., macarons and cup cakes). (c) Ours (ViT-B/16) significantly enhances class purity, leading to superior clustering of features and raising the Silhouette score to 0.732. This confirms that our MGT-guided distillation effectively transfers local ingredient-level details into the student cache.

During training, the total parameter count (GNN + Cache) remains lightweight; for ViT-B/16, only 21.7 million parameters are trained, significantly reducing memory and time costs compared to full fine-tuning. Since the GNN teacher module is omitted in inference, the proposed approach does not produce additional latency or memory overhead relative to the Tip-Adapter-F baseline. As shown in Table S3 and Figure S7, the total cost of inference computation (GFLOPs) for our approach is primarily defined by the fixed cost of the frozen backbone (e.g. 155.7 GFLOPs for ViT-L/14) plus a marginal overhead introduced by the lightweight Cache Adapter. This overhead is minimal, increasing the total inference cost by only 6.1% for the high-capacity model ViT-L/14. The Accuracy vs. Efficiency graph confirms this cost-benefit trade-off, showing that high accuracy is achieved with the ViT-L/14 model at its inherent computational cost, demonstrating that graph-guided knowledge transfer effectively maximizes the performance return on the hardware investment.

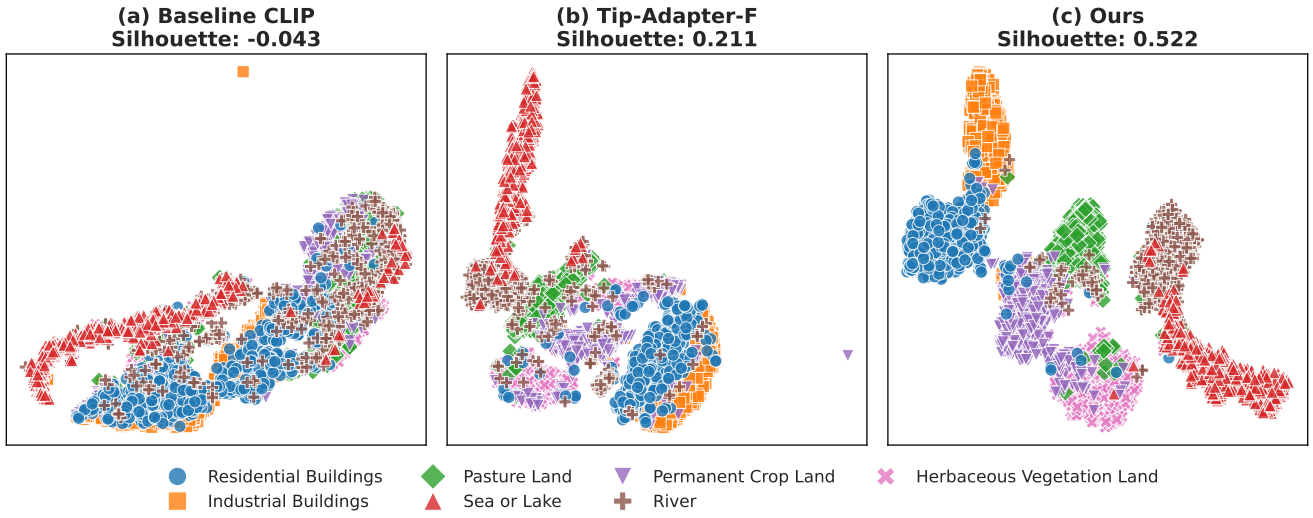


Figure S4. t-SNE visualization of logits on the EuroSAT dataset (16-shot). We visualize the embedding space for 8 randomly selected categories. (a) Baseline CLIP fails to form distinct clusters for texture-heavy classes, resulting in a negative Silhouette score of -0.043. (b) Tip-Adapter-F improves separability (0.211) but retains ambiguity between vegetation classes. (c) Ours demonstrates that the asymmetric graph supervision effectively disentangles complex spectral signatures, achieving a Silhouette score of 0.522 and showing clear separation between semantically adjacent classes like Pasture Land and Herbaceous Vegetation Land.

S6. Reproducibility and Hyperparameter Tuning

We are committed to the full reproducibility of our results. In this section, we detail the exact experimental protocols, code availability, and hyperparameter tuning strategies that ensure our method's performance is stable and reproducible across diverse visual domains.

S6.1. Unified Network Architecture

To demonstrate that our performance gains stem from topological reasoning rather than architecture engineering, we utilize a **strictly unified model architecture** across all datasets, ranging from fine-grained Flowers to satellite imagery (Eurosat). We do not alter the model depth or width per task. Specifically, we first employ a 3-layer Transformer encoder with a feed-forward multiplier of $2\times$ to refine the initial features. This is followed by our Modality-aware Graph Transformer (MGT) module with $L = 3$ layers and $H = 16$ attention heads. The final feature dimension is fixed at $D = 512$ to align perfectly with the CLIP ViT-B/16 backbone.

S6.2. Validation Protocol and Search Space

Consistent with standard few-shot learning protocols [6, 8], we strictly utilize the provided **few-shot validation set** for hyperparameter selection and explicitly avoid tuning on the test set. We acknowledge that different visual domains require different optimization dynamics. For instance, fine-grained tasks (FGVC) often require stronger contrastive regularization (λ_{con}) to separate visually similar classes, whereas satellite domains (Eurosat) may require distinct graph scaling (γ) due to the large semantic shift from the pre-training data. Therefore, rather than claiming a single scalar value works for all domains, we perform a grid search on the validation set within the ranges specified in Table S4.

S6.3. Sensitivity and Landscape Analysis

To verify the stability of our method, we analyze test accuracy trends across random hyperparameter sweeps on six dataset configurations, including Flowers (1/2/16-shot), Aircraft (2/8-shot), and Eurosat (2-shot). As visualized in Figure S8, we observe three critical technical insights:

1. **Robustness to Hardness Focus (γ_{focal}):** The performance curves for the Focal Loss parameter γ_{focal} (Figure S8a, right) remain relatively flat across all datasets. This indicates that while hard-negative mining is beneficial, the method is optimization-stable and does not require precise tuning of this parameter.
2. **Regularization vs. Shot Density (λ_{con}):** We observe a distinct shift in the need for regularization as the number of shots increases. In the extreme low-shot regime (Aircraft 2-shot), we find a strong positive correlation ($+0.63$) between λ_{con} and accuracy, suggesting that heavy regularization is required to prevent overfitting to noisy prototypes. However, as the support set grows (Aircraft 8-shot), this correlation inverts (-0.34), indicating that as prototypes become more reliable, the model benefits from relaxing the contrastive constraint.

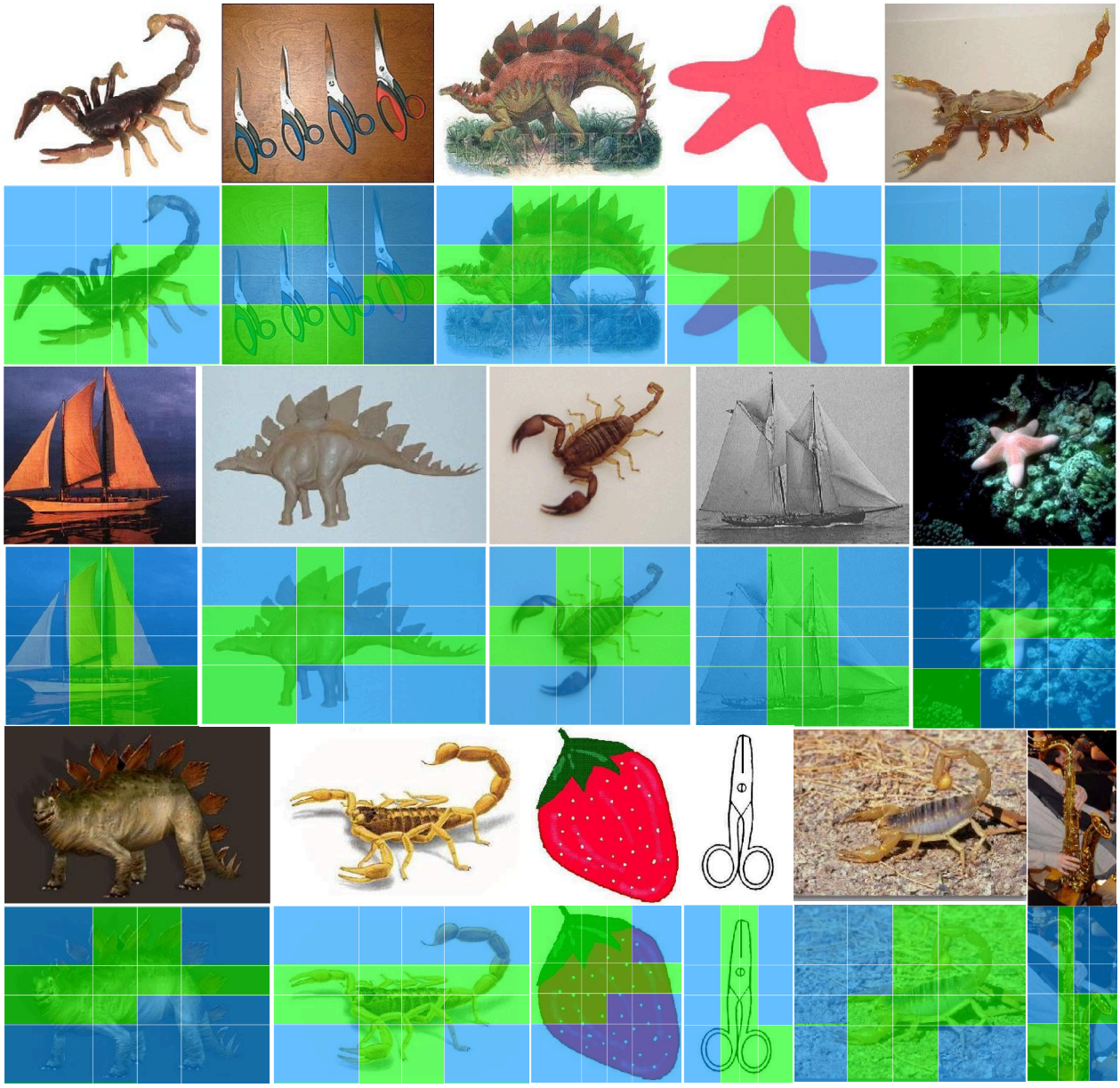


Figure S5. **Extended Qualitative Visualization of Discriminative Node Filtering.** This figure provides additional samples demonstrating the efficacy of our Top-$\mathbb{N}$ filtering across various Caltech-101 classes. The learned filter successfully identifies and retains high-scoring, discriminative foreground patches (Green) covering key object features (e.g., the segmented body of the scorpion, the silhouette of the scissors, and the spike pattern of the stegosaurus). Concurrently, it suppresses non-informative background regions (Blue). This mechanism ensures that the GNN Teacher's final visual feature is robust against feature dilution, yielding a high-fidelity relational signal for cache supervision.

3. **Domain-Specific Optimization Landscapes:** Figure S8b illustrates the bivariate optimization landscape (γ vs. λ_{con}) for different domains. While natural image datasets (Flowers, Aircraft) exhibit overlapping high-performance regions, the Eurosat dataset (satellite imagery) displays a distinct landscape topology. This confirms that the optimal balance between the pre-trained prior (CLIP) and the graph adaptation is sensitive to the semantic distance between the target domain and the pre-training data.

Table S2. Comparison of few-shot classification accuracy (%). The table compares the baseline **Tip-Adapter-F** (denoted as *Base*) against the **Proposed** method (denoted as *Ours*) across various CLIP backbones. Distinct blocks represent different shot settings ($K = \{1, 2, 4, 8\}$). Best results are highlighted in **bold**.

Shots	Backbone	Method	Datasets									
			SUN	Air	Euro	Cars	Food	Pets	Flowers	Calt	DTD	UCF
1	ResNet-101	Base	60.9	19.9	38.7	64.2	80.6	87.1	64.8	91.0	43.4	63.4
		Ours	63.5	22.7	62.1	66.1	81.0	87.8	81.1	91.9	50.5	68.0
	ViT-B/32	Base	63.8	20.8	49.9	61.1	80.3	87.2	67.8	91.5	48.2	65.9
		Ours	66.2	24.3	60.8	62.9	81.0	88.5	84.5	92.6	52.5	70.2
	ViT-B/16	Base	65.0	25.9	53.4	67.1	85.0	89.2	69.7	93.6	47.1	70.9
		Ours	68.1	31.0	67.4	69.1	86.2	91.2	88.2	94.3	55.2	74.5
	ViT-L/14	Base	70.7	37.5	65.3	79.1	**92.0**	93.1	86.1	93.8	57.0	78.5
		Ours	**75.5**	**43.4**	**71.0**	**93.1**	88.4	**93.7**	**92.1**	**97.4**	**66.5**	**78.8**
2	ResNet-101	Base	63.7	20.4	41.3	66.2	80.0	87.2	71.9	92.0	47.8	67.4
		Ours	66.5	25.0	58.8	69.8	81.4	88.8	87.7	93.0	53.9	72.6
	ViT-B/32	Base	66.4	20.6	52.6	63.0	80.2	88.0	74.9	92.7	51.7	68.6
		Ours	68.4	26.3	70.1	65.8	81.0	89.3	88.2	95.0	56.4	73.5
	ViT-B/16	Base	68.2	28.7	55.6	69.4	85.0	90.3	82.1	94.0	51.0	74.6
		Ours	70.7	34.8	74.9	72.8	86.5	92.2	91.9	95.0	58.4	77.9
	ViT-L/14	Base	73.3	37.2	70.1	80.2	**92.1**	93.4	88.8	95.4	61.1	81.0
		Ours	**76.6**	**49.0**	**75.5**	**93.6**	88.6	**94.1**	**95.2**	**97.0**	**67.0**	**81.8**
4	ResNet-101	Base	67.8	23.5	59.9	71.6	80.3	88.4	85.5	92.5	58.2	73.0
		Ours	68.9	29.8	77.0	73.4	81.3	90.0	92.5	93.4	62.0	75.2
	ViT-B/32	Base	69.3	25.7	63.4	67.0	80.5	88.7	86.5	94.5	57.8	74.6
		Ours	70.2	29.1	79.8	69.4	81.3	89.8	92.1	94.8	61.4	77.0
	ViT-B/16	Base	70.6	33.9	67.6	74.2	85.6	91.0	92.4	95.2	61.6	79.2
		Ours	73.2	38.3	80.3	76.7	86.7	92.7	96.4	95.7	64.5	81.7
	ViT-L/14	Base	76.2	45.9	76.1	84.0	**92.1**	**94.9**	96.1	96.3	68.3	83.1
		Ours	**78.2**	**52.0**	**83.2**	**93.5**	90.5	94.4	**97.7**	**97.8**	**73.6**	**85.8**
8	ResNet-101	Base	69.8	29.2	67.2	75.1	81.0	89.0	91.4	93.1	64.2	77.1
		Ours	70.9	36.1	83.0	75.9	81.7	90.2	95.1	93.9	65.4	78.5
	ViT-B/32	Base	69.9	30.6	74.2	71.8	80.8	88.9	92.0	94.7	64.3	78.1
		Ours	72.5	35.4	82.6	73.1	81.3	89.7	95.2	95.1	65.9	80.2
	ViT-B/16	Base	73.1	38.8	77.4	77.1	86.7	92.0	95.6	95.3	67.7	81.3
		Ours	75.1	44.2	84.1	78.2	86.8	93.4	97.3	95.8	69.6	83.9
	ViT-L/14	Base	78.5	53.0	78.9	86.5	**92.3**	94.5	97.6	97.2	72.0	86.2
		Ours	**79.5**	**57.0**	**83.6**	**94.3**	90.7	**94.6**	**98.2**	**97.7**	**76.0**	**86.9**

S7. Ablation of Multi-Scale Patch Hierarchy Granularity

We conduct an ablation study on the visual granularity of the Graph Teacher by varying the number and combination of multi-scale patches provided as input nodes (Table S5). This analysis validates the optimal configuration for capturing both fine-grained details and global context necessary for robust asymmetric supervision. The results confirm that the performance of the graph-guided approach is sensitive to the diversity of input views, with the established 18-Node MultiScale configuration yielding the best overall performance. This configuration is composed of a heterogeneous set of views: the global image, a 2×2 grid (4 patches),

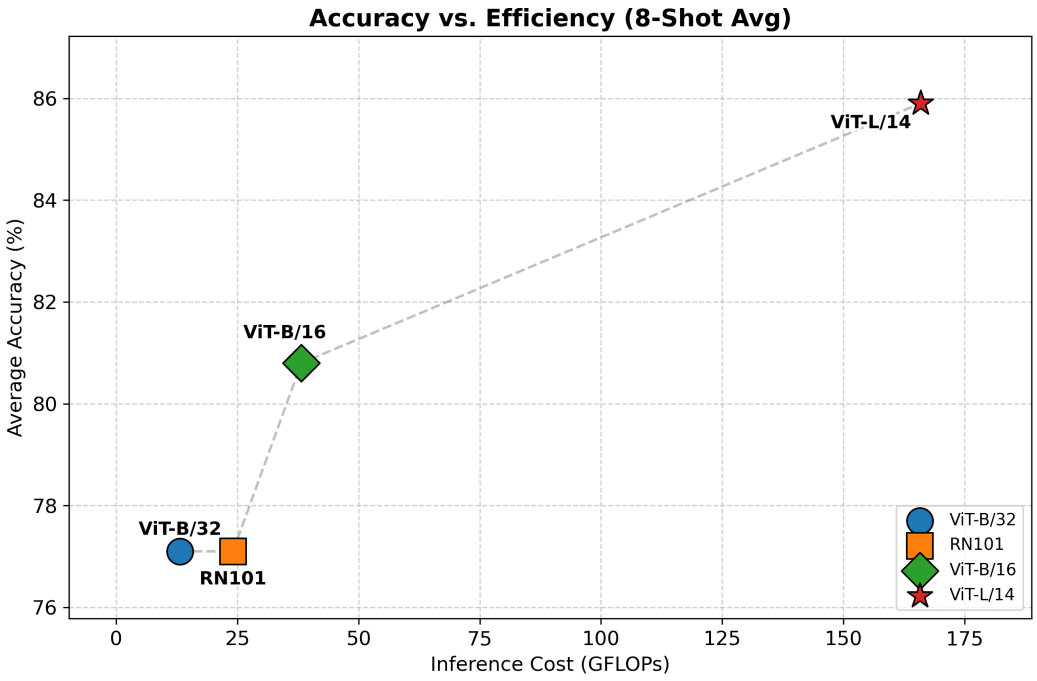


Figure S6. **Accuracy vs. Efficiency Trade-off (8-Shot Average).** We compare the total inference cost (GFLOPs) against average few-shot accuracy across 10 benchmark datasets. Our method demonstrates **strong scaling efficiency**: while the transition from ViT-B/16 to ViT-L/14 incurs higher computational cost (due to the larger backbone resolution and depth), it yields a massive **+5.1% performance gain**, reaching a state-of-the-art average accuracy of 85.9%. This highlights that our GNN-based adaptation module effectively leverages the capacity of stronger backbones without hitting a saturation point.

Table S3. **Efficiency Analysis.** We report GFLOPs (billions) and Parameters (millions) for the Backbone, our GNN module, and the Cache Adapter. Note that our GNN introduces marginal overhead compared to the backbone.

Backbone	Res.	Dim	Backbone		GNN (Ours)		Cache	
			GFLOPs	Param	GFLOPs	Param	GFLOPs	Param
RN101	224	512	19.6	56.3	4.5	13.5	0.02	8.2
ViT-B/32	224	512	8.7	87.9	4.5	13.5	0.02	8.2
ViT-B/16	224	512	33.7	86.2	4.5	13.5	0.02	8.2
ViT-L/14	224	768	155.7	304.0	10.2	30.4	0.03	12.3

Table S4. **Hyperparameter Optimization Space.** Optimization parameters are tuned on the validation set to adapt to domain granularity and shot availability.

Hyperparameter	Symbol	Search Range	Scale
Learning Rate	η	$[10^{-5}, 5 \times 10^{-2}]$	Log-Uniform
Contrastive Weight	λ_{con}	$[0.1, 1.0]$	Uniform
Graph Logit Scale	γ	$[0.1, 5.0]$	Uniform
Focal Loss Gamma	γ_{focal}	$[0.5, 3.0]$	Uniform
Tip-Adapter Alpha	α	$[1.0, 10.0]$	Uniform

a 3×3 grid (9 patches), and vertical/horizontal halves (4 patches). For the Caltech dataset, increasing the node count beyond the optimal $\mathbf{18}$ (e.g., to 21 or 30 nodes) leads to marginal performance degradation across all shot settings. The lowest performance is observed when using extremely fine-grained granularity, such as the 30-node setup (Global $+2 \times 2 + 3 \times 3 + 4 \times 4$ grids), which introduces additional visual noise or redundancy in the patches, slightly reducing accuracy in the 1-shot and 16-shot regimes.On the EuroSAT dataset, which demands reasoning over complex, heterogeneous scales, the performance variance is more pronounced. While the $\mathbf{18}$-node configuration achieves the highest accuracy in 1-shot (67.4%), 8-shot (84.1%), and 16-shot ($\mathbf{89.4}\%$), the addition of a 4×4 grid (in the 21-node configuration) offers competitive performance at 2-

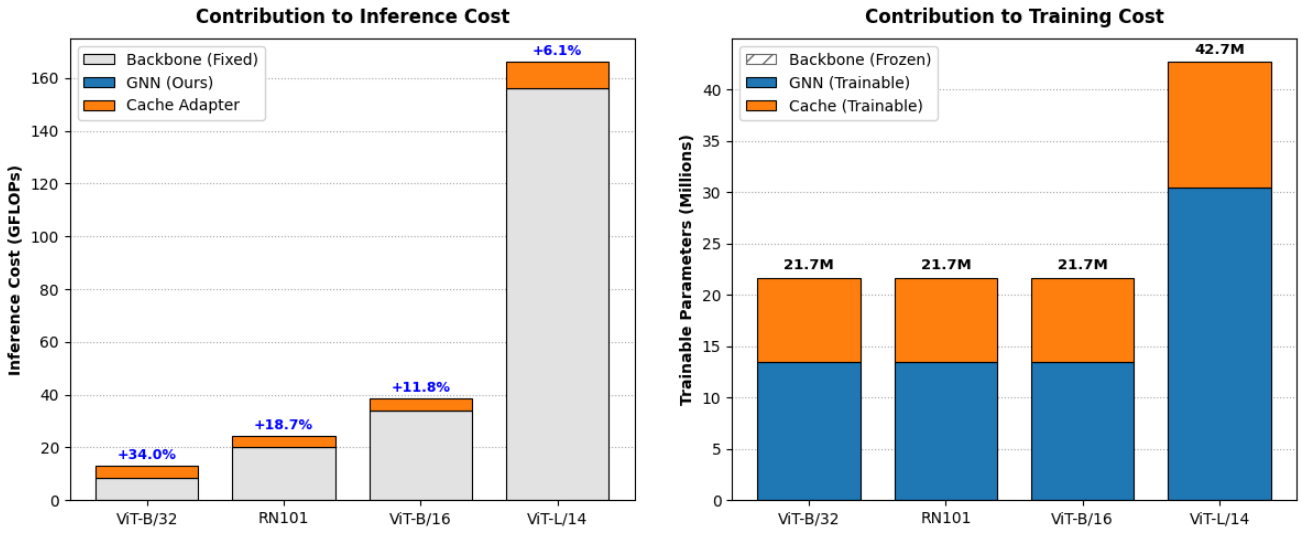


Figure S7. **Efficiency Profile: Inference vs. Training. Left:** Breakdown of inference computational cost (GFLOPs). Our method (GNN + Cache) introduces only marginal overhead compared to the pre-trained backbone, scaling efficiently to larger models (e.g., only +6.1% overhead for ViT-L/14). **Right:** Analysis of trainable parameters. By keeping the massive backbone frozen (grey), our approach requires optimizing only a lightweight set of parameters (∼21M–43M) for the GNN and Adapter, significantly reducing training memory and time costs compared to full fine-tuning.

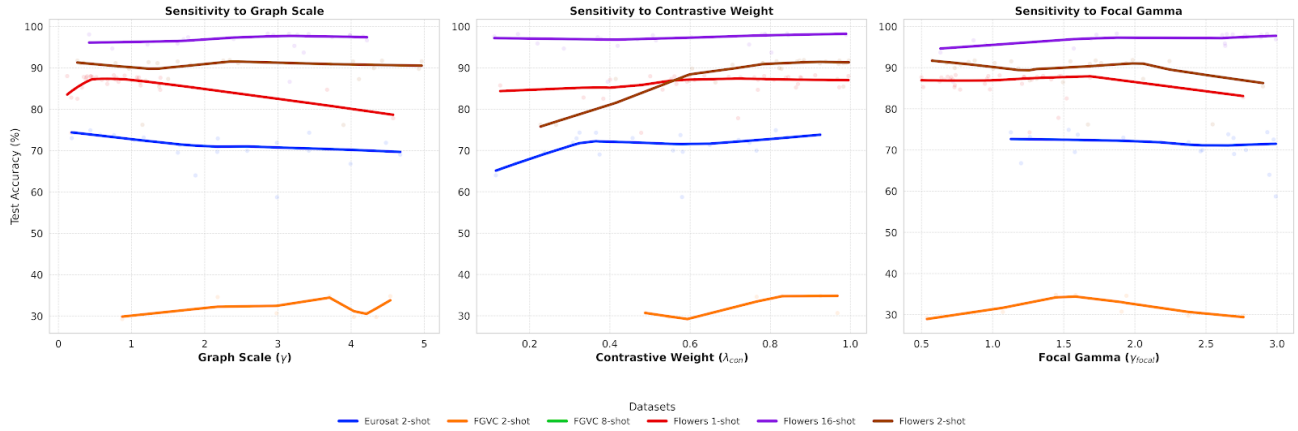
shot and 4-shot. This suggests that while fine-grained grids are beneficial for distinguishing certain local textures, the inclusion of intermediate-scale splits (halves in the 18-node setup) is more effective for maintaining larger semantic structure, particularly important for complex satellite imagery.This ablation demonstrates that a judicious balance of multi-scale views is crucial, confirming that our **18**-node MultiScale strategy provides the most comprehensive feature representation for knowledge distillation by avoiding the inherent trade-offs associated with single, fixed-scale grid strategies.

S8. Comparision between Cache and Graph model

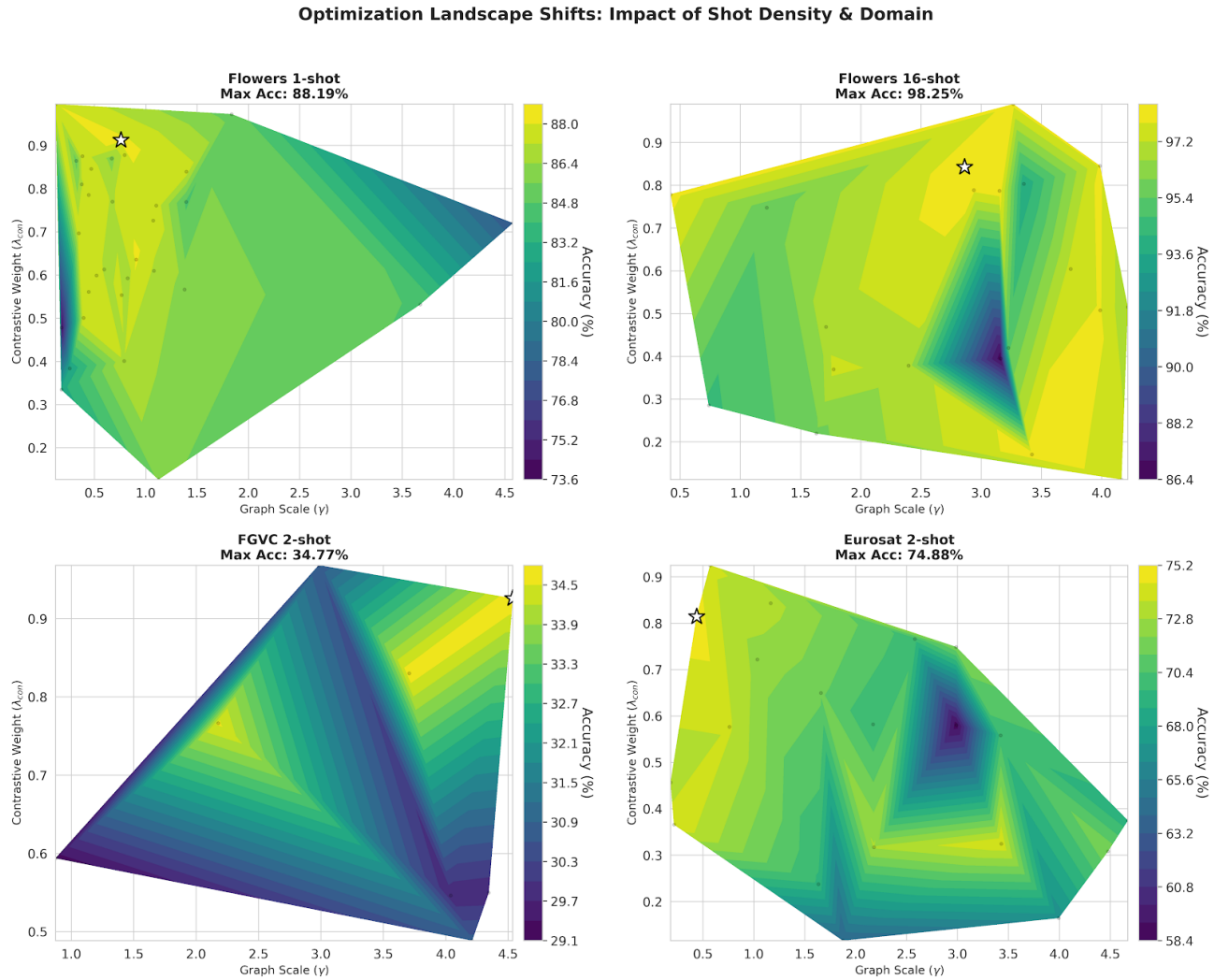
Quantitative comparisons between the Baseline (Tip-Adapter) and our Graph-Refined model on Caltech-101 are presented in Figure S9. Each example includes the ground-truth label, the input image, and the top-5 predicted classes with their corresponding confidence scores. The baseline often exhibits uncertainty or produces incorrect top-1 predictions, while our method consistently yields correct and higher-confidence outputs. In the first and third examples, the baseline misclassifies the input, whereas the graph-refined model correctly identifies the target category with a clear confidence margin. In the second and fourth examples, although the baseline selects the correct class, its probability mass remains distributed across competing categories; our approach suppresses these distractors and concentrates the distribution on the correct label. These results highlight the effectiveness of incorporating inter-class relationships through graph propagation, leading to more stable and discriminative predictions, particularly in challenging or ambiguous cases.

Table S5. Few-shot classification accuracy (in %) for the model across varying numbers of Nodes (rows) and Shots per class (columns) on the Caltech and EuroSAT datasets.

Nodes	1-Shot	2-Shot	4-Shot	8-Shot	16-Shot
Caltech101					
5	94.0	94.7	95.7	95.7	96.0
9	94.0	94.7	95.6	95.7	96.0
14	94.1	94.7	95.6	95.7	96.1
18	**94.3**	**95.0**	**95.7**	**95.8**	**96.3**
21	94.1	94.8	95.6	95.7	96.2
30	93.8	94.9	95.6	95.6	96.0
EuroSAT					
5	64.5	72.6	79.6	83.6	88.4
9	66.2	73.5	79.4	81.8	86.2
14	62.6	74.7	76.9	81.2	86.8
18	**67.4**	74.9	80.3	**84.1**	**89.4**
21	66.8	**75.3**	**81.1**	83.0	88.5
30	65.4	72.6	77.0	83.0	88.6



(a) **Comprehensive Sensitivity Trends.** LOWESS smoothed curves show the impact of hyperparameters across 6 datasets. Note the varying slopes for λ_{con} depending on shot count.



(b) **Landscape Evolution.** We compare the optimization landscape (γ vs. λ_{con}) across different shots and domains. High-shot settings (e.g., Flowers 16-shot) and distinct domains (Eurosat) exhibit different optimality regions compared to low-shot natural images.

Figure S8. **Hyperparameter Sensitivity Analysis.** We visualize the impact of key hyperparameters on test accuracy, highlighting the necessity of our validation-based tuning protocol.

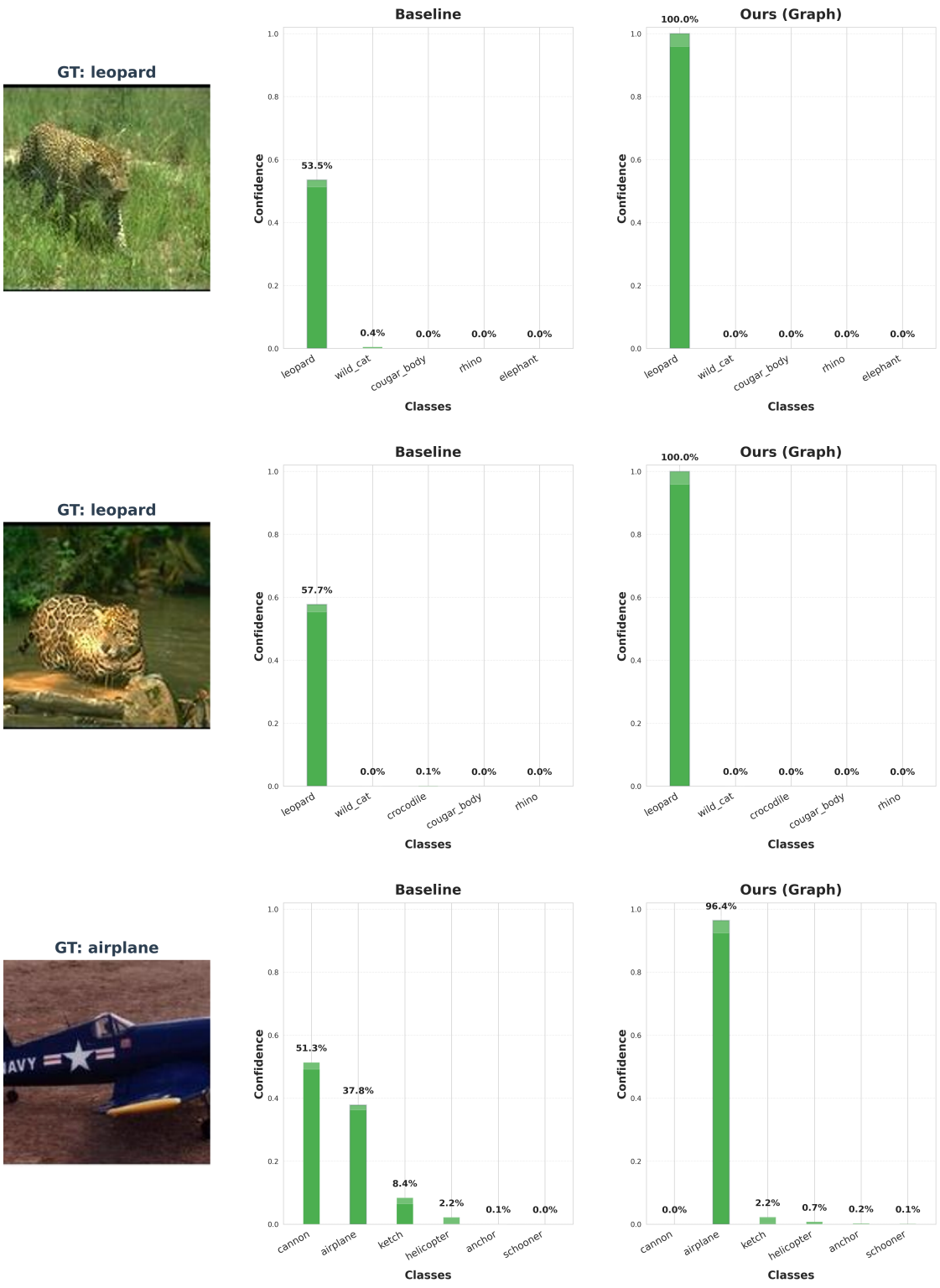


Figure S9. Quantitative comparison between the Baseline (Tip-Adapter) and Ours (Graph Refined) models on four test examples from the Caltech-101 dataset. Each panel displays the ground truth (GT) label, the input image, and the top-5 classifications with their associated probabilities. Green bars denote the correct class, while red bars indicate an incorrect top-1 prediction. The results demonstrate that the graph-based method corrects the baseline's misclassifications (e.g., in the first and third panels) and significantly improves confidence in correct predictions (e.g., in the second panel).

S9. Additional Ablation Studies

Table S6. Impact of Text Nodes in MGT.

Configuration	1 Shot	2 Shot	4 Shot	8 Shot	16 Shot
EuroSAT					
w/o Text	64.6	70.5	75.5	79.8	84.2
w/ Text	**67.4**	**74.9**	**80.3**	**84.1**	**89.4**
Caltech101					
w/o Text	93.6	93.6	95.1	95.2	95.7
w/ Text	**94.3**	**95.0**	**95.7**	**95.8**	**96.3**
Food101					
w/o Text	85.8	86.1	86.3	86.3	87.1
w/ Text	**86.2**	**86.5**	**86.7**	**86.8**	**87.5**

Table S7. Comparison of MGT vs. Transformer architectures across different datasets and shot configurations.

MGT	Trans	1 Shot	2 Shot	4 Shot	8 Shot	16 Shot
EuroSAT						
✗	✓	62.9	72.1	77.4	80.8	87.5
✓	✗	**67.4**	**74.9**	**80.3**	**84.1**	**89.4**
Aircraft						
✗	✓	27.4	31.6	34.9	41.5	45.7
✓	✗	**31.0**	**34.8**	**38.3**	**44.2**	**48.4**
UCF101						
✗	✓	71.8	74.6	78.3	80.2	80.5
✓	✗	**74.5**	**77.9**	**81.7**	**83.9**	**84.9**

Impact of Text Nodes in Graph Topology. We evaluate the contribution of text embeddings as structural nodes in Table S6. A baseline ("w/o Text") constructs a visual-only graph can aggregate patch features but lacks semantic guidance during the reasoning process. This is insufficient for disentangling complex spectral signatures, as seen on EuroSAT. Our proposed method ("w/ Text") explicitly incorporates text nodes, enabling visual features to attend to and align with high-level semantic concepts layer-by-layer. This "steers" the visual representation toward the correct prototype before the final classification stage. The importance of this mechanism is validated by the substantial and widening performance gap at higher shots, indicating that this semantic guidance remains critical even as visual data increases.

Ablation on Architecture: MGT vs. Standard Transformer. As detailed in Table S7, we evaluate the few-shot generalization capabilities of our proposed MGT architecture against a standard Transformer baseline across three diverse visual domains: remote sensing (EuroSAT), fine-grained object classification (Aircraft), and video action recognition (UCF101). The empirical results demonstrate that the MGT consistently outpaces the standard Transformer across all shot configurations ($K \in \{1, 2, 4, 8, 16\}$). Notably, the MGT architecture exhibits pronounced sample efficiency in highly constrained data regimes, yielding substantial absolute performance gains of +4.5% on EuroSAT and +3.6% on FGVC-Aircraft under the extreme 1-shot setting. This consistent performance advantage across distinct modalities and varying data availability highlights the superior feature representation and robustness of the MGT framework when compared to traditional Transformer counterparts.

References

[1] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Prompt learning with optimal transport for vision-language models. In *Proceedings of the ICLR*, 2023. 1, 3

[2] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19113–19122, 2023. 1, 3

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 3

[4] Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, Boyu Diao, and Yongjun Xu. Clip-kd: An empirical study of clip model distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15952–15962, 2024. 1

[5] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6757–6767, 2023. 1, 3

[6] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pages 493–510. Springer, 2022. 1, 5

[7] 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, 3

[8] 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, 3, 5

[9] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15659–15669, 2023. 1, 3