PROMPTING WITH CONNECTIONS: GRAPH-GUIDED ATTRIBUTE SELECTION FOR TEST-TIME CALIBRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

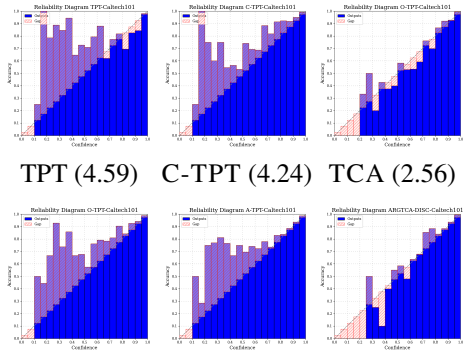
Test-time prompt tuning adapts vision–language models (VLMs) to unseen target domains without labels by minimizing prediction entropy over augmented views. While boosting zero-shot accuracy, this process often induces overconfidence and degrades calibration. Existing approaches initialize prompts with LLM-generated visual attributes under contrastive regularization, but treat attributes independently, ignoring their relational structure. We introduce ARGTCA, a graph-guided framework for calibrated test-time prompt tuning that explicitly models relationships among class-specific visual attributes. By representing class–attribute pairs as nodes and linking within-class siblings and identical attribute across classes in a *Symbolic Attribute Graph*; a GAT trained with a supervised contrastive objective produces relational embeddings and refines links into geometry, making attribute genericity—how many classes share an attribute—legible as distance from the class centroid (Spearman $\rho : -0.130 \rightarrow +0.588$ on Caltech101). The embeddings drive two prompt-initialization strategies: ARGTCA-DIV selects diverse, complementary attributes, while ARGTCA-DISC selects highly discriminative ones; the selected attributes then initialize the test-time prompts. Across ten visual-classification benchmarks on CLIP ViT-B/16, ARGTCA-DIV reduces average ECE to 4.46, a reduction of up to 58.3% relative to existing test-time prompt-tuning methods, while ARGTCA-DISC attains the highest average accuracy (65.02%) at a competitive ECE of 5.17. On ResNet-50, ARGTCA-DIV again attains the lowest average ECE (3.66) and ARGTCA-DISC the second-lowest (3.99), confirming generalization across CNN and Transformer encoders. Relational-Attribute selection runs once, offline, adding no test-time overhead. Code-base will be released upon acceptance.

1 INTRODUCTION

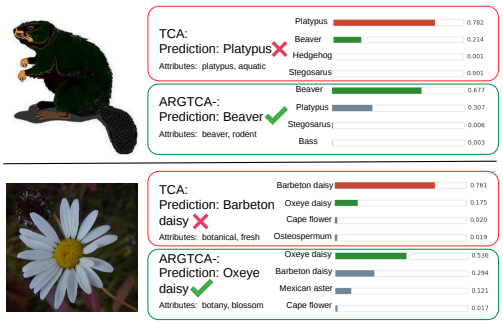
Vision–language models such as CLIP (Radford et al., 2021) classify zero-shot by comparing an image embedding against class-conditioned text embeddings built from a hand-written template, such as "`a photo of a {class}.`" Test-time prompt tuning (TPT) (Shu et al., 2022) removes the need to guess such manual template and adapts soft-prompt tokens at inference time using augmented views of each test sample, without requiring labeled target data. However, its entropy-minimization objective that sharpens the prediction also inflates its confidence, worsening the Expected Calibration Error (ECE) (Guo et al., 2017) even when classification accuracy improves. This accuracy–calibration trade-off limits the deployment of VLMs in safety-critical applications, such as healthcare and autonomous systems, where reliable confidence estimates are essential for risk-aware decision-making. The accuracy–calibration decoupling is not unique to TPT: prompt templates with comparable top-1 accuracy can exhibit different ECE (Yoon et al., 2024), motivating works targeting calibration through the geometry of class-conditioned text representations. C-TPT (Yoon et al., 2024) maximizes Average Text Feature Dispersion among class-conditioned text representations; O-TPT (Sharifdeen et al., 2025) instead imposes orthogonality constraints for angular separation. TCA (Hebbalaguppe et al., 2025) further uses LLM-generated class-level visual attributes for semantically informed prompt initialization, regularized to pull intra-class and push inter-class textual representations.

Independent ranking is suboptimal. Only M' of the multiple LLM-generated candidate attributes reach the prompt, so *attribute selection* determines the text geometry. TCA selects these M' attributes by independently ranking class–attribute cosine similarity, without modeling relationships among attributes. Two failure modes follow from this (Figure 2). *Intra-class redundancy*: near-synonyms

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a)



(b)

Figure 1: (a) **Reliability diagrams:** Caltech101 (ViT-B/16, one seed), ECE (↓ the better) in parentheses. TCA is overconfident; TPT, C-TPT, O-TPT, and A-TPT are underconfident. ARGTCA-DISC (ours) aligns closest to the diagonal. More plots in Figure 8. (b) **Label-corrections:** TCA confidently misclassifies (*Platypus*/*Barbeton daisy*); ARGTCA corrects with better-aligned confidence.

(*rough*, *bumpy*, *grainy*) all score highly against the class name and cluster in CLIP's embedding space, so top-M' selection can pick redundant attributes, limiting representation diversity (Figure 2a). *Inter-class ambiguity*: generic attributes (*mammal*) are similarly related to their class names (*leopard*, *dolphin*, and *beaver*), which can pull unrelated classes together and provide little discriminative signal (Figure 2b). Neither failure mode is a property of a single class–attribute pair: redundancy involves a *pair of attributes within* a class, while ambiguity involves a *pair of classes sharing* an attribute. Neither are visible to TCA's single-pair score $\cos(c_k, a_m)$, which cannot capture relational attribute geometry.

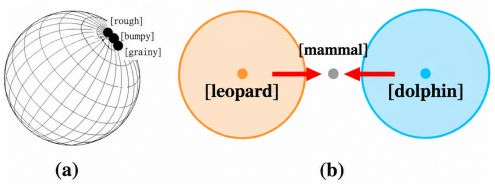


Figure 2: TCA's failure modes. **(a)** Near-synonyms reduce feature diversity. **(b)** Shared attributes reduce inter-class separability.

We propose **ARGTCA** (**A**ttribute **R**elation **G**raph for **T**est-time **C**alibration **A**daptation), which operationalizes this relational view. ARGTCA constructs a *Symbolic Attribute Graph* (SAG), where each $(\text{class}, \text{attribute})$ pair forms a node, with two edge types: within-class edges exposing redundancy and shared-token edges across classes exposing ambiguity. A Graph Attention Network (GAT) is trained without target-domain images. The resulting relational embeddings support two complementary selection strategies: ARGTCA-DIV, which promotes intra-class semantic diversity, and ARGTCA-DISC, which promotes inter-class discriminability. For a new dataset, we recommend ARGTCA-DIV as the more consistent calibration-oriented choice, while ARGTCA-DISC provides an alternative when accuracy or distribution-shift robustness is prioritized. We evaluate ARGTCA on ten benchmarks spanning general, fine-grained, texture, remote-sensing, and action recognition, plus four natural distribution-shift benchmarks. Graph-structured attribute modeling improves calibration, as illustrated by the reliability diagrams in Figure 1(a) and the label-correction examples in Figure 1(b).

Our contributions are: 1. We identify two limitations of attribute-based test-time calibration: *intra-class redundancy* and *inter-class ambiguity*, both pairwise properties that a single-attribute scoring function cannot capture. 2. We introduce ARGTCA, which models class–attribute relationships using a Symbolic Attribute Graph and a contrastively trained GAT, resulting in relation-based attribute selection. 3. We show that graph training makes attribute genericity—how many classes share a token—legible as distance from the class centroid (Spearman $\rho : -0.130 \rightarrow +0.588$ on Caltech101); the resulting inter-class discriminability margin exhibits an $\mathcal{O}(K^{-1/2})$ concentration trend with the number of classes K, while the intra-class diversity margin has no such dependence. 4. Across ten benchmarks and two CLIP backbones, ARGTCA improves calibration without any test-time overhead.

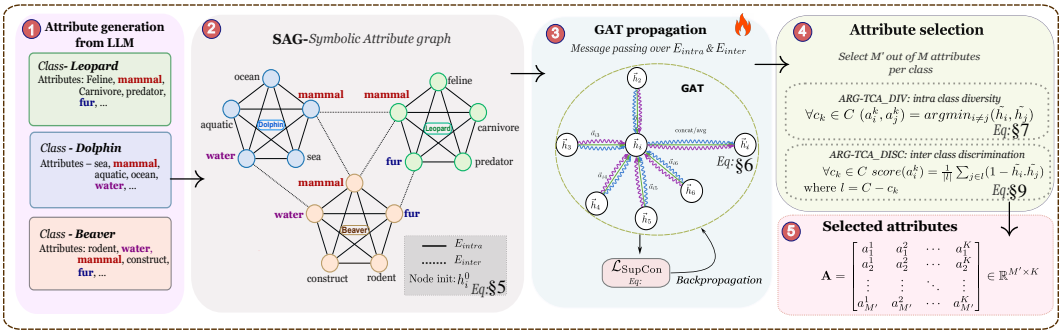


Figure 3: Overview of ARGTCA attribute selection pipeline. ❶ An LLM generates M candidate attributes per class. ❷ These are used to construct the Symbolic Attribute Graph (SAG) with intra-class (E_{intra}) and inter-class (E_{inter}) edges. ❸ A Graph Attention Network (GAT) is trained via supervised contrastive loss ($\mathcal{L}_{\text{SupCon}}$) to propagate structural relationships across the SAG, producing attribute embeddings. ❹ Two selection criteria — ARGTCA-DIV (intra-class diversity) and ARGTCA-DISC (inter-class discrimination) — select the most informative M' attributes per class. ❺ The final selected attributes form the matrix $\mathbf{A} \in \mathbb{R}^{M' \times K}$, used for downstream prompt tuning.

2 RELATED WORK

2.1 TEST-TIME ADAPTATION OF VLMS AND CALIBRATION OF VLMS

Vision–language models such as CLIP and ALIGN (Radford et al., 2021; Jia et al., 2021) enable zero-shot classification by aligning image and text representations. Prompt-learning methods such as CoOp (Zhou et al., 2022b) and CoCoOp (Zhou et al., 2022a) improve performance but require labeled training data. In contrast, TPT (Shu et al., 2022) adapts prompts at inference time by minimizing prediction entropy across augmented views. Subsequent methods, including DiffTPT (Feng et al., 2023) and PromptAlign (Abdul Samadh et al., 2023), improve augmentation and distribution alignment. However, entropy-based adaptation often induces overconfidence and degrades calibration relative to zero-shot CLIP (Yoon et al., 2024).
Classical post-hoc methods such as temperature scaling (Guo et al., 2017) and Platt scaling (Platt, 1999) require labeled validation data. For label-free test-time calibration, C-TPT (Yoon et al., 2024) relates calibration to Average Text Feature Dispersion, while O-TPT (Sharifdeen et al., 2025) and A-TPT (Ahamed et al., 2026) explicitly improve angular separation among text features. TCA (Hebbalaguppe et al., 2025) uses LLM-generated visual attributes with contrastive regularization, but models attributes independently and therefore overlooks their relational structure.

2.2 GRAPH-BASED VLM ADAPTATION

Graph neural networks (Kipf & Welling, 2017; Veličković et al., 2018) provide a natural mechanism for relational representation learning. GraphAdapter (Li et al., 2023), HGCLIP (Xia et al., 2025), and VCGPrompt (Wang et al., 2026) exploit class hierarchies or visual concept graphs for supervised or few-shot VLM adaptation. Unlike these approaches, ARGTCA models relationships among symbolic class attributes in a label-free setting to improve test-time calibration. It extends TCA by replacing independent attribute selection with graph-structured reasoning over text feature geometry.

3 METHODOLOGY

We propose ARGTCA, a graph-guided extension to TCA's (Hebbalaguppe et al., 2025) attribute-selection that leverages pairwise class–attributes relationships for prompt initialization, prior to test-time prompt tuning. Given class names and corresponding LLM-generated visual attributes, ARGTCA selects attributes that are complementary within a class or discriminative across classes. ARGTCA comprises two phases: offline graph-based attribute selection, followed by standard test-time prompt adaptation. In **Phase 1** (offline, once per dataset), we construct a *Symbolic Attribute Graph* (SAG), where each node represents a $(\text{class}, \text{attribute})$ pair. Intra-class edges connect all

nodes of the same class (enabling relational modeling within a class); inter-class edges connect nodes sharing an identical attribute token (cross-class overlap). A Graph Attention Network (GAT), trained using supervised contrastive learning, produces relational attribute embeddings. These embeddings support two attribute selection strategies: ARGTCA-DIV, which promotes intra-class diversity, and ARGTCA-DISC, which promotes inter-class discriminability. In **Phase 2**, selected attributes initialize prompts in the original TCA prompt-tuning pipeline without altering its optimization objective. Thus, the graph serves solely as a structural prior for attribute selection. In ARGTCA, graph construction, GAT training, and attribute selection are performed offline using only class names, hence it requires no target-domain data, and incurs no additional test-time overhead.

3.1 PROBLEM SETUP

Let $\mathcal{C} = \{c_k\}_{k=1}^{K}$ denote the class set and $\mathcal{A}_k = \{a_m^k\}_{m=1}^{M}$ denote the LLM-generated visual attributes for class c_k. Our goal is to select $M' < M$ attributes per class that improve calibration during test-time prompt tuning. Following TCA (Hebbalaguppe et al., 2025), an attribute-conditioned prompt is constructed as

$$t_{k,m} = p \oplus a_m^k \oplus c_k, \tag{1}$$

where p is a learnable soft prompt and $\oplus$ denotes token concatenation. The frozen CLIP text encoder $g(\cdot)$ produces $f_{k,m} = g(t_{k,m})$. Given image feature v, the class probability is

$$P(c_k \mid v) = \frac{\sum_{m=1}^{M'} \exp\big(v^\top f_{k,m}/\tau^*\big)}{\sum_{k'=1}^{K}\sum_{m=1}^{M'} \exp(v^\top f_{k',m}/\tau^*)}, \tag{2}$$

where τ^* is CLIP's learned logit scale. At test-time, TCA optimizes only p of prompt using

$$\mathcal{L}_{\text{TCA}} = \mathcal{L}_{\text{TPT}} - \alpha\mathcal{L}_{\text{inter}} + \beta\mathcal{L}_{\text{intra}}, \tag{3}$$

where $\mathcal{L}_{\text{TPT}}$ minimizes prediction entropy, $\mathcal{L}_{\text{inter}}$ separates class centroids, and $\mathcal{L}_{\text{intra}}$ encourages attribute compactness. All parameters of CLIP's vision and text encoders remain frozen. Figure 9 illustrates the test-time pipeline; Appendix C gives details.

3.2 PHASE 1: SYMBOLIC ATTRIBUTE GRAPH

Graph construction. In *Symbolic Attribute Graph* (SAG), each class–attribute pair (c_k, a_m^k) defines a graph node and connects two nodes when they share either a class or an attribute, as shown in Figure 3 (Step 2). That is:

$$\mathcal{E} = \underbrace{\{(i,j) : c_i = c_j\}}_{E_{\text{intra}}} \cup \underbrace{\{(i,j) : a_i = a_j\}}_{E_{\text{inter}}} \qquad i \neq j. \tag{4}$$

The two edge types carry exactly the two relations independent ranking cannot see. E_{intra} enables relational modeling among attributes of the same class, while E_{inter} captures across-class semantic overlap. We also experiment with semantic-similarity-based edges in Appendix L.6. The feature for each node i in SAG is initialized from frozen CLIP text embedding

$$h_i^{(0)} = \text{norm}\big(g(\text{Prompt}(c_k, a_m^k))\big) \in \mathbb{R}^d, \tag{5}$$

where $\text{Prompt}(\cdot)$ is a fixed template such as "`a {attribute} of {class}`"

Graph representation learning. A multi-head Graph Attention Network aggregates information over both edge types and updates node embeddings as:

$$h_i^{(l)} = \sigma\left(\Big\|_{q=1}^{H} \sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(l,q)} W_q^{(l)} h_j^{(l-1)}\right). \tag{6}$$

where α_{ij} is attention coefficient, W learnable projection matrix, $\mathcal{N}(i)$ neighborhood of node i, and $\|$ is concatenation. The GAT is trained using supervised contrastive loss (SCL) (Khosla et al., 2020), treating same-class nodes as positives. While SCL promotes class-level compactness and inter-class separation, the attention mechanism performs attribute-conditioned message passing, producing graph-refined embeddings without uniform feature diffusion. The resulting embeddings retain sufficient geometric variation for subsequent diversity- and discriminability-based attribute selection. Lower ECE of ARGTCA-DIV's intra-class-only variant (Table 4) than random selection (Table 2) confirms this empirically that the diversity signal (Eq. 7) survives training. The random-selection and no-GAT-training ablations in Appendix L further corroborate this.

3.3 GRAPH-GUIDED ATTRIBUTE SELECTION

We select M' attributes per class using one of two strategies.

ARGTCA-DIV. This strategy selects attributes with minimum within-class cosine similarity, thereby promoting semantic diversity:

$$(a_i^k, a_j^k) = \arg \min_{\substack{i,j \in \Omega_k \\ i \neq j}} \tilde{h}_i^\top \tilde{h}_j, \tag{7}$$

where $\tilde{h}_i$ denotes the normalized graph-refined embedding of node i from last layer of GAT, and Ω_k contains the nodes belonging to class c_k.

For any value of M', Eq. 7 generalizes to the max–min dispersion objective

$$S_k^\star = \arg \max_{\substack{S \subseteq \Omega_k \\ |S|=M'}} \min_{\substack{i,j \in S \\ i \neq j}} \big(1 - \tilde{h}_i^\top \tilde{h}_j\big), \tag{8}$$

which recovers Eq. 7 at M'=2. As exact subset selection is combinatorial, we approximate $S_k^\star$ greedily: initialize S with the pair from Eq. 7, then iteratively add the attribute maximizing its minimum distance to the current set.

ARGTCA-DISC. Each attribute is scored by its mean cosine distance from nodes of all other classes:

$$s(a_i^k) = \frac{1}{|\bar{\Omega}_k|} \sum_{j \in \bar{\Omega}_k} \left(1 - \tilde{h}_i^\top \tilde{h}_j\right), \tag{9}$$

where $\bar{\Omega}_k$ denotes all non-c_k nodes. The top-M' attributes are selected according to this score. Since ARGTCA-DISC ranks candidates within each class, scaling $\mu_{\bar{k}}$ by any positive constant leaves the ranking unchanged; the $O(K^{-1/2})$ scaling of σ_{disc} in Corollary 1 (Section 5), reflects a concentration trend under the stated isotropy assumption. For DIV, the selection objective depends only on the fixed per-class candidate pool M, and therefore has no explicit dependence on K, examined empirically in Appendix K.

3.4 PHASE 2: ATTRIBUTE-CONDITIONED TEST-TIME ADAPTATION

The selected attributes are inserted into the standard TCA prompts using Eq. 1. For each test image, TCA retains the most confident augmented views and performs one optimization step on the soft prompt p using Eq. 3. Prediction is then made on the original image, and p is reset before processing the next sample.

4 EXPERIMENTAL SETUP

4.1 DATASETS AND BASELINES

We evaluate on widely used visual classification benchmark datasets using CLIP ViT-B/16 and ResNet50, following the evaluation protocol of (He et al., 2016; Radford et al., 2021). We compare against: **TPT** (Shu et al., 2022), **C-TPT** (Yoon et al., 2024), **O-TPT** (Sharifdeen et al., 2025), **A-TPT** (Ahamed et al., 2026) and **TCA** (Hebbalaguppe et al., 2025). Baseline results reported here may differ from those in their original papers, as the original works tune hyperparameters per dataset. Instead, we re-evaluate each baseline using its own set of hyperparameters, tuned on the Caltech101 validation set and applied across all datasets, following TCA. Appendix A describes details of baselines and dataset statistics.

4.2 IMPLEMENTATION DETAILS

Phase 1 — Offline Graph Training. Attribute graph is built from M =10 LLM-generated attributes per class (from (Hebbalaguppe et al., 2025)), giving $K \times 10$ nodes. The Attribute GAT (2 layers, 4 heads, 512-d features) is trained for 100 epochs with supervised contrastive loss (τ=0.07) using Adam (lr=10^{-3}); training time is < 5 mins per dataset (Fig 6), except large ImageNet (~ 25 min).

Phase 2 — Test-Time Prompt Tuning. We follow the standard TCA protocol: soft context/prompt initialized to *"a photo of a"* (n_{ctx}=4), 64 augmented views per image, retaining the 10% with lowest

Table 1: Top-1 accuracy (Acc., ↑) and Expected Calibration Error (ECE, ↓) on 10 visual classification benchmarks using CLIP ViT-B/16 and ResNet-50 backbones. Results are averaged over three random seeds. **Red** and orange denote the best and second-best ECE values.

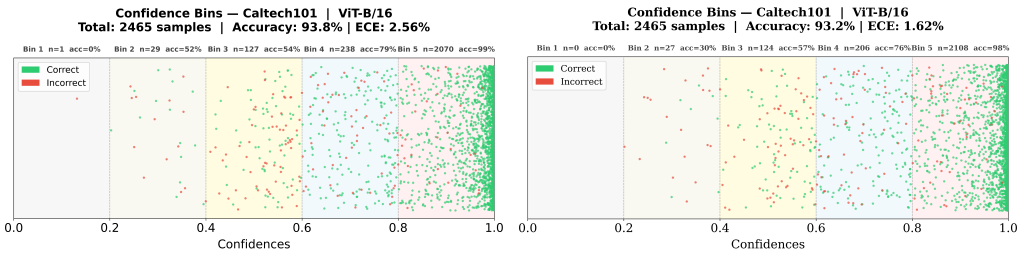
Method	Metric	ImageNet	Caltech101	OxfordPets	StanfordCars	Flower	Food101	Aircraft	DTD	EuroSAT	UCF101	**Average**
Backbone: CLIP ViT-B/16 (512-d)												
TPT	Acc.	68.70	93.64	87.06	65.89	69.18	83.88	23.79	46.48	42.28	64.22	64.51
NeurIPS'22	ECE	11.40	4.61	5.85	5.47	13.47	4.12	16.74	21.27	21.29	2.66	10.69
C-TPT	Acc.	68.5	93.58	88.26	65.49	69.68	83.28	24.02	45.57	43.16	65.26	64.68
ICLR'24	ECE	3.15	4.47	1.93	1.69	5.16	3.78	4.37	12.69	13.28	2.36	5.29
TCA	Acc.	67.37	92.90	90.54	65.84	69.09	81.06	25.18	44.70	45.80	66.86	64.93
ECML'25	ECE	2.27	3.01	6.30	7.85	3.76	5.28	3.60	5.48	8.37	2.82	4.87
O-TPT	Acc.	58.91	93.66	88.51	65.77	69.96	83.52	24.02	46.97	43.24	66.85	64.14
CVPR'25	ECE	3.10	3.58	1.85	1.60	7.04	3.08	7.51	19.43	20.47	3.52	7.12
A-TPT	Acc.	66.73	93.70	88.43	66.41	69.60	83.59	23.51	46.89	23.49	67.57	62.99
ICLR'26	ECE	2.15	2.93	2.37	2.45	10.58	2.01	13.46	20.03	13.48	6.12	7.56
ARGTCA-DIV	Acc.	66.87	92.95	89.47	65.02	70.56	83.88	24.64	47.79	32.91	69.50	**64.36**
(Ours)	ECE	2.02	2.70	3.55	2.83	3.63	1.85	5.10	4.22	15.80	2.85	**4.46**
ARGTCA-DISC	Acc.	67.40	93.10	88.91	66.44	70.56	84.05	24.87	46.73	39.53	68.61	**65.02**
(Ours)	ECE	2.27	1.94	3.16	4.04	3.33	2.16	4.45	6.62	19.84	3.85	**5.17**
Backbone: CLIP ResNet-50 (1024-d)												
TPT	Acc.	68.80	87.00	84.50	58.00	62.50	74.90	17.00	41.50	28.30	59.50	58.20
NeurIPS'22	ECE	10.72	5.04	3.65	3.76	13.40	5.25	16.10	25.70	22.50	12.40	11.85
C-TPT	Acc.	60.20	86.90	84.10	56.50	65.20	74.70	17.00	42.20	27.80	59.70	57.43
ICLR'24	ECE	3.01	2.07	2.77	1.94	4.14	1.86	10.70	19.80	15.10	3.83	6.52
TCA	Acc.	68.10	86.69	86.21	55.95	64.47	75.38	17.04	39.59	31.32	61.04	58.58
ECML'25	ECE	1.88	1.79	5.43	3.35	3.70	2.45	4.48	8.16	5.50	4.33	4.11
O-TPT	Acc.	58.97	86.86	83.40	56.44	65.61	74.22	16.77	41.90	28.35	58.84	57.14
CVPR'25	ECE	3.10	2.75	3.50	1.69	2.50	1.20	8.18	16.53	14.71	2.60	5.68
A-TPT	Acc.	60.50	86.69	84.49	57.51	62.93	74.74	17.40	41.61	29.91	59.61	57.54
ICLR'26	ECE	9.14	3.52	1.59	3.54	10.88	3.63	15.55	25.90	10.90	9.29	9.39
ARGTCA-DIV	Acc.	58.30	87.51	86.32	55.98	66.34	74.51	17.43	41.43	29.17	59.98	57.70
Ours	ECE	1.55	2.47	5.29	2.87	4.76	2.51	4.71	5.27	4.07	3.14	**3.66**
ARGTCA-DISC	Acc.	58.96	86.86	84.63	55.60	64.72	75.61	16.98	42.08	24.46	58.74	56.86
Ours	ECE	1.74	2.12	4.58	2.73	3.78	2.16	4.56	7.59	8.43	2.25	3.99

entropy, and one gradient step per image (batch size 64). Regularization weights: α=10, β=35. M'=2 attributes per class are selected from the graph-refined embeddings via either the diversity (ARGTCA-DIV) or discriminability (ARGTCA-DISC) criterion.

Metrics. Our primary metric is Expected Calibration Error (ECE, %, ↓; 20 equal-width bins); we report top-1 accuracy (Acc, %, ↑) for completeness, following (Hebbalaguppe et al., 2025). We also report MCE and Brier score in Appendix E. All experiments use single A100 80GB GPU.

5 RESULTS

Calibration on Visual Classification Benchmarks. Table 1 reports results averaged over three random seeds. On the CLIP ViT-B/16 backbone, ARGTCA-DIV achieves the best average ECE of 4.46, reducing miscalibration by 15.7%, 8.4%, 37.4%, 41.0% and 58.3% relative to C-TPT, TCA, O-TPT, A-TPT, and TPT respectively. ARGTCA-DISC attains the highest average accuracy of 65.02% with an ECE of 5.17, outperforming O-TPT and A-TPT in both accuracy and calibration, and remaining within 0.30 of TCA's ECE (4.87) while improving accuracy. Similarly, on the CLIP ResNet-50 backbone, ARGTCA-DIV achieves the lowest average ECE of 3.66, a reduction of 10.9%, 35.6%, 43.9%, 61.0%, and 69.1% relative to TCA, O-TPT, C-TPT, A-TPT, and TPT respectively, while ARGTCA-DISC attains the second-lowest ECE at 3.99. The large improvements fall on several datasets where TCA is poorly calibrated, e.g., StanfordCars 7.85 → 2.83, Food101 5.28 → 1.85, OxfordPets 6.30 → 3.55, DTD 5.48 → 4.22. The exception is EuroSAT, whose small label space and sparse cross-class attribute sharing leaves little relational structure for the graph to exploit (Appendix E). Evaluations on natural distribution shift datasets are reported in Appendix D;



(a) TCA (b) ARGTCA-DISC (Ours)

Figure 4: Confidence-bin plots on Caltech101 (ViT-B/16), ARGTCA-DISC vs. TCA (closest competitor), one representative seed. Green/red indicate correct/incorrect predictions. ARGTCA-DISC shifts most errors into low-confidence bins while preserving confident correct predictions, achieving the lowest ECE. Additional plots are in Appendix Figures 10, 11, 12.

ARGTCA-DISC attains the lowest ECE on ImageNet-A and ImageNet-S, while ARGTCA-DIV attains the lowest ECE on ImageNet-V2.

Confidence–Error Separation and Reliability

Figures 4, 1(a), and 1(b) together characterize the alignment at three levels: aggregate error, per-bin calibration, and per-instance behavior (on one representative seed). ARGTCA-DISC reduces incorrect predictions at high confidence while retaining most correct ones (Figure 4), yielding the lowest ECE of 1.62 against 2.56, 3.62, and 2.67 for TCA, O-TPT, and A-TPT. This is consistent with the reliability diagrams (Figure 1(a)): TCA is overconfident (bars below the diagonal), TPT, C-TPT, O-TPT, and A-TPT overcorrect into underconfidence (bars above), while ARGTCA-DISC tracks the diagonal most closely. At the instance level (Figure 1b), TCA's overconfidence manifests concretely as a confident misclassification (*platypus*), which ARGTCA-DISC corrects (*beaver*) with better-aligned confidence. Additional confidence plots (Figure 10), reliability diagrams (Figure 8), and qualitative examples in Appendix.

How Graph Training Reshapes Attribute Geometry

Figure 5 visualizes ARGTCA-DISC node embeddings on Caltech101 before ($h^{(0)}$) and after ($h^{(L)}$) GAT training, colored by attribute genericity (number of classes sharing that attribute token) with selected attributes marked in green. Genericity is essentially *weakly negatively correlated* with radial position in $h^{(0)}$ ($\rho = -0.130$) but *strongly positively correlated* in $h^{(L)}$ ($\rho = +0.588$): training pushes generic attributes outward and contracts class-specific ones toward the centroid, and ARGTCA-DISC's selections concentrate in the resulting low-genericity core. GAT training does not introduce semantic content absent from CLIP's embeddings, but reorganizes distance-from-centroid into a legible proxy for genericity — the signal the discriminability score exploits when ranking attributes by projection away from $\mu_{\bar{k}}$.

Effect of Attribute Selection Strategy The two strategies target complementary embedding geometries. Proposition 1 and its corollary formalize this geometry (proofs in Appendix J). The empirical validation (Appendix K) connect it to the observed calibration behavior.

Proposition 1 (Selection geometry)**.** *Let $\tilde{h}_i \in \mathbb{R}^d$ denote normalized graph-refined embeddings of candidate attributes from class c_k, with Ω_k, $\bar{\Omega}_k$ the sets of same-class and other-class nodes.*
(a) Intra-class diversity. *For $M' = 2$ and $\mu_{ij} = (\tilde{h}_i + \tilde{h}_j)/2$: $1 - \tilde{h}_i^\top \tilde{h}_j = 2(1 - \|\mu_{ij}\|_2^2)$.* ARGTCA-DIV*'s angular-dispersion objective is equivalent to minimizing the pair-centroid norm.*
(b) Inter-class discriminability. *With $\mu_{\bar{k}} = \frac{1}{|\bar{\Omega}_k|} \sum_{j \in \bar{\Omega}_k} \tilde{h}_j$ and $s_i = \frac{1}{|\bar{\Omega}_k|} \sum_{j \in \bar{\Omega}_k} (1 - \tilde{h}_i^\top \tilde{h}_j)$, score $s_i = 1 - \tilde{h}_i^\top \mu_{\bar{k}}$:* ARGTCA-DISC *ranks attributes by projection away from competing-class centroid.*

Corollary 1 (DISC: Label-space)**.** *Under approximately isotropic competing-class embeddings and fixed M, $\|\mu_{\bar{k}}\|_2 = O((KM)^{-1/2})$, and $\sigma_{\text{disc}} = O(K^{-1/2})$. The ranking depends only on the normalized competing-class centroid $\hat{\mu}_{\bar{k}}$.*

Figure 18 confirms Corollary 1 empirically: σ_{disc} decays as $K^{-0.46}$ ($R^2 = 0.86$) across ten benchmarks (K=10 to 1000), while σ_{div} shows no specific trend (Appendix K).

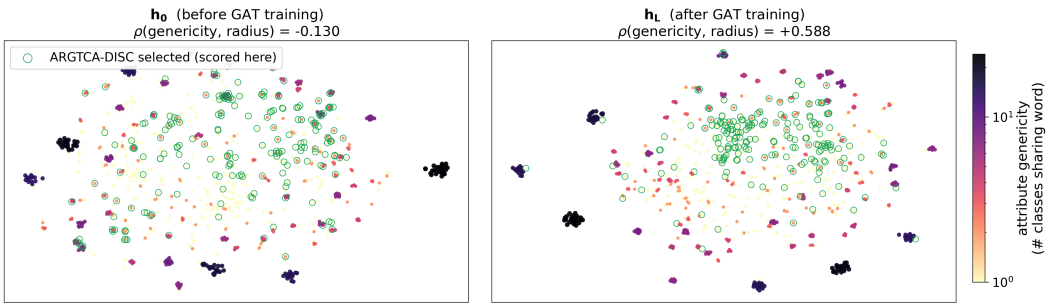


Figure 5: Attribute geometry for ARGTCA-DISC on Caltech101 before ($h^{(0)}$) and after ($h^{(L)}$) GAT training. Nodes are colored by genericity and positioned by class-centroid radius; green rings mark selected attributes. Genericity–radius correlation shifts from $\rho = -0.130$ to $+0.588$, with selections concentrating at low genericity after training.

Table 2: Component ablation of ARGTCA. Acc. (%,↑) and ECE (%,↓).

Variant	Sel.	Caltech101		OxfordPets		DTD	
		Acc	ECE	Acc	ECE	Acc	ECE
ARGTCA	Div	92.82	2.52	89.48	3.44	47.34	**3.63**
	Disc	93.19	**1.62**	88.91	**3.37**	46.63	6.36
ARGTCA w/o edges	Div	93.18	2.36	89.32	3.84	45.04	4.91
	Disc	93.51	1.74	90.00	3.50	45.74	5.31
ARGTCA w/o GAT	Div	92.21	2.73	88.74	3.87	46.81	5.65
	Disc	93.39	2.60	88.99	3.71	45.98	8.95
Non-GNN (Random)	Div	93.10	2.53	88.63	3.44	45.92	5.16
	Disc	92.25	2.43	88.50	3.55	44.27	6.33

Table 3: Backbone ablation. Mean ± std. over three seeds.

Sel.	Backbone	Acc. ↑	ECE ↓
Div	MLP	75.74 ± 0.49	4.15 ± 0.98
	GAT	76.74 ± 0.10	3.49 ± 0.22
	GCN	76.31 ± 0.34	3.55 ± 0.16
	GraphSAGE	75.39 ± 0.78	4.17 ± 0.32
Disc	MLP	76.19 ± 0.46	3.53 ± 0.83
	GAT	76.25 ± 0.12	3.91 ± 0.25
	GCN	76.28 ± 0.41	4.44 ± 0.33
	GraphSAGE	75.71 ± 1.02	3.69 ± 0.76

6 ABLATION STUDIES

All ablations use ViT-B/16 on Caltech101, OxfordPets, and DTD. Table 2, isolates three components of ARGTCA: (i) *w/o edges* ($\mathcal{E} = \emptyset$, no message passing, contrastive training and DIV/DISC selection retained); (ii) *w/o GAT* (DIV/DISC applied to raw CLIP embeddings $\mathbf{h}^{(0)}$, Phase 1 skipped); (iii) *random selection* (Phase 1 skipped entirely). ARGTCA-DIV outperforms all three ablations in $8/9$ comparisons (exception: zero-edge Caltech101, 2.36 vs. 2.52), suggesting a consistent benefit from graph-based training. For ARGTCA-DISC, performance is more dataset-dependent: on DTD, it underperforms both zero-edge (5.31) and random selection (6.33), while substantially outperforming no-GAT (8.95) and achieving the highest accuracy. For DISC, the edge-free and no-GAT variants yield slightly higher accuracy on Caltech101 and OxfordPets, suggesting that the benefit of graph propagation is more evident in calibration than accuracy on these datasets.

Training backbone ablation. Table 3 (right) additionally compares GAT (inductive) against GCN (transductive), GraphSAGE (inductive), and MLP: GAT has the lowest standard deviation under both criteria ($\pm 0.10/\pm 0.12$; vs. ± 0.34–1.02 GCN/GraphSAGE, ± 0.46–0.49 MLP) and lowest ECE variance under ARGTCA-DISC (± 0.25), trailing only GCN under ARGTCA-DIV (± 0.22 vs. ± 0.16). MLP's marginally lower mean ECE comes at 3–4× GAT's seed variance; given our focus on calibration reliability, we retain GAT (full configurations in Appendix H).

Edge-type contribution. Table 4 isolates intra- vs. inter-class edges. For ARGTCA-DIV, combining both outperforms either alone on OxfordPets/DTD, with inter-only marginally ahead on Caltech101 (2.07 vs. 2.52) — edge types are largely complementary. For ARGTCA-DISC, combining helps only on Caltech101; intra-only dominates OxfordPets and DTD (3.67 vs. 6.36), consistent with inter-class edges degrading representations in attribute-overlapping classes given ARGTCA-DISC's inter-class angular score (Eq. 9) — left to future work to verify. Overall, graph-based training is beneficial for ARGTCA-DIV, while the benefit for ARGTCA-DISC varies across datasets and edge configurations.

Qualitative attribute selection. Table 5 (right) shows TCA's cosine top-2 frequently often picks

Table 4: Effect of intra- and inter-class edges on ARGTCA (ECE, ↓). Single seed

Variant	Sel.	Caltech101		OxfordPets		DTD	
		Acc	ECE	Acc	ECE	Acc	ECE
TCA (no graph)		93.82	2.56	90.51	6.30	44.73	5.82
(i) Intra-class	Div	90.87	2.11	88.69	4.03	45.80	5.34
	Disc	92.90	2.34	88.12	**3.04**	45.92	3.67
(ii) Inter-class	Div	92.74	2.07	89.42	3.71	44.50	4.38
	Disc	92.98	1.85	89.45	4.12	44.56	6.66
ARGTCA (Both)	Div	92.82	2.52	89.48	3.44	47.34	**3.63**
	Disc	93.19	**1.62**	88.91	3.37	46.63	6.36

Table 5: Graph-based selection yields more informative attributes than TCA. Red: synonym/redundant; Green: complementary/discriminative.

Dataset	Class	TCA (top-2)	ARGTCA-DIV	ARGTCA-DISC
Caltech101	kangaroo	furry, animal	nature, baby	wild, marsupial
	cellphone	technology, device	portable, talk	cellphone, numeric
OxfordPets	Abyssinian	Abyssinian, cats	restful, observant	cats, feline
	Boxer	Dog, Boxer	Athletic, Friendly	Boxer, Pet

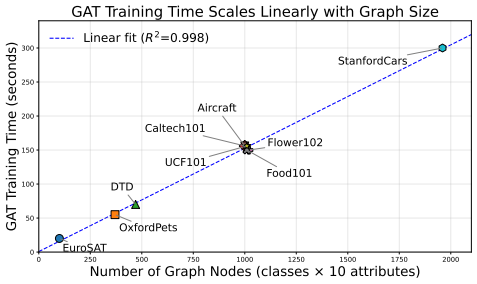


Figure 6: GAT training time vs. graph size.

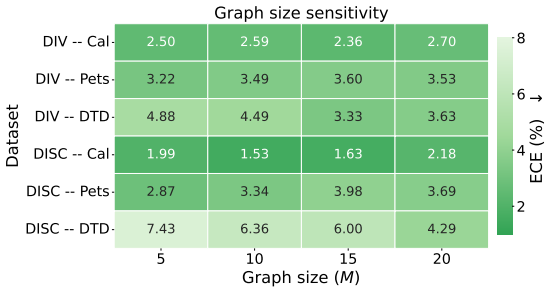


Figure 7: ECE sensitivity to graph size M.

near-synonyms of the class label ("dog" for boxer), the clustering failure of Fig. 2a, whereas ARGTCA-DIV selects semantically orthogonal pairs (*portable, talk* for cellphone) and ARGTCA-DISC selects category-identifying tokens (*wild, marsupial* for kangaroo) maximizing inter-class angular distance.

Sensitivity to M and computational overhead. Figure 7 sweeps $M \in \{5, 10, 15, 20\}$ with $M' = 2$. ECE varies by less than 0.7 points in 3/6 criterion–dataset pairs, with larger but non-monotonic changes otherwise; the exception is ARGTCA-DISC on DTD, which improves monotonically ($7.43 \rightarrow 6.36 \rightarrow 6.00 \rightarrow 4.29$). We retain $M = 10$ as the default because larger M adds computational overhead without consistent ECE gains, and fix $M' = 2$ for comparability with TCA. Full sweeps are in Appendix G.

This robustness to graph-size extends to compute cost: as attribute generation, graph construction, GAT training, and selection are performed once offline, ARGTCA adds no test-time overhead beyond TCA regardless of M. GAT training scales near-linearly with graph size ($R^2 = 0.998$): ∼22 s (EuroSAT) to ∼25 min (ImageNet, 10,000 nodes) (Figure 6); per-dataset run-times in Appendix F.

7 CONCLUSION

We introduced ARGTCA, a graph-guided framework for improving the test-time calibration of CLIP. Unlike TCA, which ranks attributes independently, ARGTCA models their intra- and inter-class relationships through a *Symbolic Attribute Graph* and a contrastively trained GAT. The resulting relational embeddings support two complementary selection strategies: ARGTCA-DIV for semantic intra-class diversity and ARGTCA-DISC for inter-class discriminability. Across diverse visual recognition benchmarks and vision backbones, ARGTCA-DIV achieves the lowest average ECE of 4.46 on ViT-B/16 and 3.66 on ResNet50, with relative reductions in miscalibration of up to 58.3% and 69.1% respectively. Ablations further establish the individual contributions of graph topology, contrastive learning, and geometric attribute selection. As all graph operations are performed offline, ARGTCA improves confidence reliability without modifying the underlying CLIP model or adding test-time overhead. Future work will investigate joint diversity–discriminability objectives and transferable attribute graphs that generalize across datasets.

AI USE STATEMENT

Generative AI tools were used for paraphrasing, grammar checking, readability edits, and proof-writing assistance, but not for research ideation, literature review, method or experiment design, or figure generation. We additionally used the Stanford Agentic Reviewer as an author-side review tool for obtaining preliminary feedback and identifying areas requiring further clarification and improvement. All AI-assisted content was reviewed by the authors, and we take full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

All hyperparameters are listed in Section 4 and Appendix A.1; the GAT architecture is specified in Appendix B, and the Phase 2 objective, inherited unchanged from TCA (Hebbalaguppe et al., 2025), in Appendix C. We use the public test splits of Zhou et al. (2022b) and public attribute lists from Hebbalaguppe et al. (2025); no target-domain data, labels, or images are used in the offline stage. The main results in Table 1 report means over three seeds due to space limitations, while other multi-seed results report mean$\pm$std. Ablations and figures reports single-seed results. Code and the selected attribute lists for all ten benchmarks will be released upon acceptance.

REFERENCES

Jameel Abdul Samadh, Mohammad Hanan Gani, Noor Hussein, Muhammad Uzair Khattak, Muhammad Muzammal Naseer, Fahad Shahbaz Khan, and Salman Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 80396–80413. Curran Associates, Inc., 2023. doi: 10.52202/075280-3525.

Shihab Ahamed, Udaya Sampath K Perera Miriya Thanthrige, Ranga Rodrigo, and Muhammad Haris Khan. A-tpt: Angular diversity calibration properties for test-time prompt tuning of vision-language models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 69175–69199, 2026.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *Proceedings of the 13th European Conference on Computer Vision*, ECCV, pp. 446–461, 2014.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014.

J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255, 2009.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples. In *Proceedings of the CVPR 2004 Workshop on Generative-Model Based Vision*, 2004.

Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zhong. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, ICCV, pp. 2704–2714, 2023.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, ICML, pp. 1321–1330, 2017.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, NeurIPS, pp. 1024–1034, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Ramya Hebbalaguppe, Tamoghno Kandar, Abhinav Nagpal, and Chetan Arora. Prompting without panic: Attribute-aware, zero-shot, test-time calibration. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2025, Porto, Portugal, September 15–19, 2025, Proceedings, Part VI*, pp. 289–305, Berlin, Heidelberg, 2025. Springer-Verlag. ISBN 978-3-032-06105-8. doi: 10.1007/978-3-032-06106-5_17.

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.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8340–8349, October 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021b.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, ICML, 2021.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, ICLR, 2017.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *ICCV Workshop on 3D Representation and Recognition*, 2013.

Xin Li, Dongze Lian, Zhihe Lu, Jiawang Bai, Zhibo Chen, and Xinchao Wang. Graphadapter: Tuning vision-language models with dual knowledge graph. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 13448–13466. Curran Associates, Inc., 2023. doi: 10.52202/075280-0591.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *CoRR*, abs/1306.5151, 2013.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, 2012.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, 1999.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, ICML, pp. 8748–8763. PMLR, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, pp. 5389–5400, 2019.

Ashshak Sharifdeen, Muhammad Akhtar Munir, Sanoojan Baliah, Salman Khan, and Muhammad Haris Khan. O-tpt: Orthogonality constraints for calibrating test-time prompt tuning in vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 19942–19951, June 2025.

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. In *Advances in Neural Information Processing Systems*, volume 35 of *NeurIPS*, pp. 14274–14289, 2022.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, ICLR, 2018.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019.

Mengjia Wang, Fang Liu, Licheng Jiao, Shuo Li, Lingling Li, Puhua Chen, Xu Liu, and Wenping Ma. VCGPrompt: Visual Concept Graph-Aware Prompt Learning for Vision-Language Models. *Pattern Recognition*, 170:112012, February 2026.

Peng Xia, Xingtong Yu, Ming Hu, Lie Ju, Zhiyong Wang, Peibo Duan, and Zongyuan Ge. HGCLIP: Exploring vision-language models with graph representations for hierarchical understanding. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 269–280, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark Hasegawa-Johnson, Yingzhen Li, and Chang D. Yoo. C-tpt: Calibrated test-time prompt tuning for vision-language models via text feature dispersion, 2024.

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*, CVPR, pp. 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130:2337–2348, 2022b.

Appendix Contents

Section	Page
A Datasets and Baselines	15
Implementation Details of ARGTCA	15
Offline Graph Training	15
Test-Time Prompt Tuning	16
Metrics	16
B GAT Architecture Details	16
Attention and Aggregation	17
Residual and Normalization	18
C Test-Time Prompt Tuning (Phase 2)	18
D Robustness to Natural Distribution Shifts	19
E MCE and Brier Score	20
F Computational Overhead of the SAG and GAT Training Phase	22
G M and M' Sensitivity	23
Number of Selected Attributes M'	24
Graph Size M	24
H Graph Neural Network Ablation	25
H.1 Backbone Configurations	25
H.2 Backbone Model Analysis	26
I Representation Geometry Before and After Graph Training	26
J Theoretical Analysis	28
Proposition 1 (a) (Intra-Class Diversity)	28
Proposition 1 (b) (Inter-Class Discriminability)	29
Proposition 2 (Confidence lower bound)	29
Corollary 1 (DISC: Label-space)	30
K Empirical Attribute-Selection Analysis	30
Structure of the Symbolic Attribute Graph	30
Aggregation Robustness	31
Accuracy–Calibration Relationship	31
Gain over Random Selection	31
Why the Discriminability Margin Concentrates as $K^{-1/2}$	31
Predicting the Better Variant	32
L Ablation Studies	32
L.1 Attribute Comparison: TCA *vs.* ARGTCA-DIV *vs.* ARGTCA-DISC	32
L.2 Zero Graph Edges	33
L.3 No GAT *Training (Raw* CLIP *EOS Embeddings)*	33
L.4 Random Attribute Selection	33
L.5 MLP Attribute Transform	33

continued on next page

Appendix Contents (continued)

Section (continued)	Page
L.6 Exact versus Semantic Attribute Matching	33
L.7 Hyperparameter Sensitivity (α, β)	34
M Qualitative Label Corrections Across Datasets	34
N Broader Impact and Implications	34

Key Tables and Figures

Summary of Evaluation Datasets . . . 15
Full Positional Breakdown of a Prompt in ARGTCA . . . 16
Notation Reference . . . 18
One-Time Offline GAT Training Time per Dataset . . . 18
Reliability Diagrams Across All Datasets (ViT-B/16) . . . 17
Test-Time Prompt Calibration Pipeline . . . 20
Natural Distribution Shift Results . . . 20
Confidence-Bin Comparison Against All Baselines (Caltech101) . . . 21
Confidence-Bin Comparison Against All Baselines (DTD) . . . 21
Confidence-Bin Comparison Against All Baselines (UCF101) . . . 22
MCE and Brier Score Results . . . 22
Graph-Size M Sensitivity . . . 23
Prompt-Width M' Sensitivity . . . 24
Backbone Configurations . . . 25
Backbone Ablation on the Attribute Graph . . . 26
Seed-Paired Difference Against the GAT . . . 26
Reproducibility Across Seeds . . . 25
Per-Dataset Backbone Results (ARGTCA-DIV) . . . 27
Per-Dataset Backbone Results (ARGTCA-DISC) . . . 27
Hub Structure Before and After Graph Training . . . 28
Genericity Before and After Graph Training . . . 29
Symbolic Attribute Graph Statistics . . . 31
Verification of the Discriminability Decomposition . . . 32
Selection Margin vs. Label-Space Size . . . 32
Exact versus Semantic Matching for Inter-Class Edges . . . 34
Connectivity and ECE under Exact and Semantic Matching . . . 34
ECE vs. α (Inter-Class Regularization) . . . 35
ECE vs. β (Intra-Class Regularization) . . . 35
Label Corrections: ARGTCA-DISC vs. TCA (Aircraft, Caltech101) . . . 36
Label Corrections: ARGTCA-DISC / ARGTCA-DIV vs. TCA (DTD) . . . 37
Label Corrections: ARGTCA-DIV vs. TCA (Caltech101, Stanford Cars) . . . 38

A Datasets and Baselines

We evaluate on ten visual classification benchmarks spanning general object recognition, fine-grained recognition, and specialized domains: ImageNet (Deng et al., 2009); Caltech101 (Fei-Fei et al., 2004) (general objects); StanfordCars (Krause et al., 2013), FGVC-Aircraft (Maji et al., 2013), OxfordPets (Parkhi et al., 2012), and Flowers102 (Nilsback & Zisserman, 2008) (fine-grained recognition); Food101 (Bossard et al., 2014) (food categories); DTD (Cimpoi et al., 2014) (perceptual textures); UCF101 (Soomro et al., 2012) (action recognition); and EuroSAT (Helber et al., 2019) (satellite land-use).

We additionally evaluate on four natural distribution shift benchmarks derived from ImageNet: ImageNet-A (Hendrycks et al., 2021b), a test set of natural adversarial examples; ImageNet-V2 (Recht et al., 2019), an independently collected complementary test set; ImageNet-R (Hendrycks et al., 2021a), containing artistic renditions of ImageNet classes; and ImageNet-S (Wang et al., 2019), containing black-and-white sketches of ImageNet classes. We follow the standard test splits of (Zhou et al., 2022b), as used in TCA (Hebbalaguppe et al., 2025), evaluating on the test set only with no labeled data.

Table 7: Summary of evaluation datasets.

Dataset	**# Classes**	**Test set size**
ImageNet	1,000	50,000
Caltech101	100	2,465
OxfordPets	37	3,669
StanfordCars	196	8,041
Flowers	102	2,463
Food101	101	30,300
FGVCAircraft	100	3,333
DTD	47	1,692
EuroSAT	10	8,100
UCF101	101	3,783
ImageNet-A	200	7,500
ImageNet-V2	1000	10,000
ImageNet-R	200	30,000
ImageNet-S	1000	50,889

We compare against:

- **TPT** (Shu et al., 2022): the first test-time prompt tuning approach, using entropy minimization to tune prompts at inference time without requiring labeled data.
- **C-TPT** (Yoon et al., 2024): adds an average text feature dispersion loss (ATFD) to TPT (Shu et al., 2022) for calibration.
- **TCA** (Hebbalaguppe et al., 2025): initializes prompts with LLM-generated visual attributes and applies contrastive intra-/inter-class regularization; our method builds on and refines its attribute selection.
- **O-TPT** (Sharifdeen et al., 2025): replaces ATFD with pairwise angular separation between class centroids.
- **A-TPT** (Ahamed et al., 2026): corrects O-TPT's degeneracy when the number of classes exceeds the embedding dimension.

Following TCA (Hebbalaguppe et al., 2025), we re-evaluate each baseline using its own set of hyperparameters, tuned on the Caltech101 validation set and applied consistently across all datasets.

A.1 Implementation Details of ArgTca

Offline Graph Training. For each dataset, we construct a Symbolic Attribute Graph with 10 LLM-generated attributes per class, yielding $10K$ nodes. A two-layer, four-head GAT with 512-

Table 8: Full positional breakdown of a prompt in ARGTCA. $l = 1$ is number of token in attribute.

Position	Content	Notes
0	SOS token embedding	Frozen
$1, \ldots, n_{\text{ctx}}$	Soft prompt $p \in \mathbb{R}^{n_{\text{ctx}} \times d}$	**Learnable**
$n_{\text{ctx}} + 1$	Selected attribute a_m^k	**Graph-selected**
$n_{\text{ctx}} + 2, \ldots, l - 1$	Class tokens + EOS + padding	Frozen

dimensional node features is trained for 100 epochs using supervised contrastive loss ($\tau = 0.07$) and Adam ($\text{lr} = 10^{-3}$). Training takes at maximum five minutes per dataset except large ImageNet dataset (~ 25 minutes).

Test-Time Prompt Tuning. Following TCA, textual prompt is initialized with four soft context tokens *"a photo of a"*. For each image, 64 augmented views are generated, the lowest-entropy 10% of these views are retained, and one prompt-update step is performed. We set $\alpha = 10$, $\beta = 35$, and select $M' = 2$ attributes per class using ARGTCA-DIV or ARGTCA-DISC. Table 8 details the positional structure of the ARGTCA prompt. The SOS token embedding and class tokens, together with EOS and padding, remain frozen. Only the soft prompt parameters $p \in \mathbb{R}^{n_{\text{ctx}} \times d}$ are learnable, while the graph-selected attribute a_m^k fills the attribute-token position. Thus, ARGTCA modifies the prompt initialization through offline attribute selection while retaining the standard test-time optimization of the soft prompt. The CLIP backbone remains frozen.

All experiments use a single NVIDIA A100 80GB GPU.

A.2 METRICS

Our primary metric is Expected Calibration Error (ECE, %, ↓); we additionally report top-1 accuracy (Acc, %, ↑) for completeness, following the protocol of (Hebbalaguppe et al., 2025), along with Maximum Calibration Error (MCE, %, ↓) and Brier score (↓) in Appendix E.

Confidence is partitioned into B_e=20 equal-width bins over $[0, 1]$; for bin b containing n_b of the n test samples, let $\text{acc}(b)$ and $\text{conf}(b)$ denote the empirical accuracy and mean predicted confidence within that bin.

ECE is the frequency-weighted average calibration gap across bins:

$$\text{ECE} = \sum_{b=1}^{B_e} \frac{n_b}{n} \big|\text{acc}(b) - \text{conf}(b)\big|. \tag{10}$$

MCE reports the worst-case gap instead of the weighted average, and is therefore sensitive to sparsely populated bins:

$$\text{MCE} = \max_{b \in \{1,\ldots,B_e\}} \big|\text{acc}(b) - \text{conf}(b)\big|. \tag{11}$$

Brier score is the mean squared error between the predicted class-probability vector $P(c_k \mid v)$ (Eq. 18) and the one-hot ground-truth label, evaluated over all K classes and all n samples:

$$\text{Brier} = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \big(P(c_k \mid v_i) - \mathbb{1}[y_i = k]\big)^2. \tag{12}$$

Unlike ECE and MCE, Brier score is a strictly proper scoring rule that rewards both calibration and sharpness, so a sharper but less calibrated predictor can obtain a lower Brier score.

B GAT ARCHITECTURE DETAILS

Our graph-based attribute selection follows the Graph Attention Network (GAT) framework of (Veličković et al., 2018), which assigns learned importance weights to neighborhood edges during message passing, enabling the model to selectively aggregate information from the most relevant neighbors. Our SAG is an *undirected* graph; each undirected edge $\{i, j\} \in \mathcal{E}$ is represented as two directed edges $(i \to j)$ and $(j \to i)$ during message passing, following standard practice in GAT implementations.

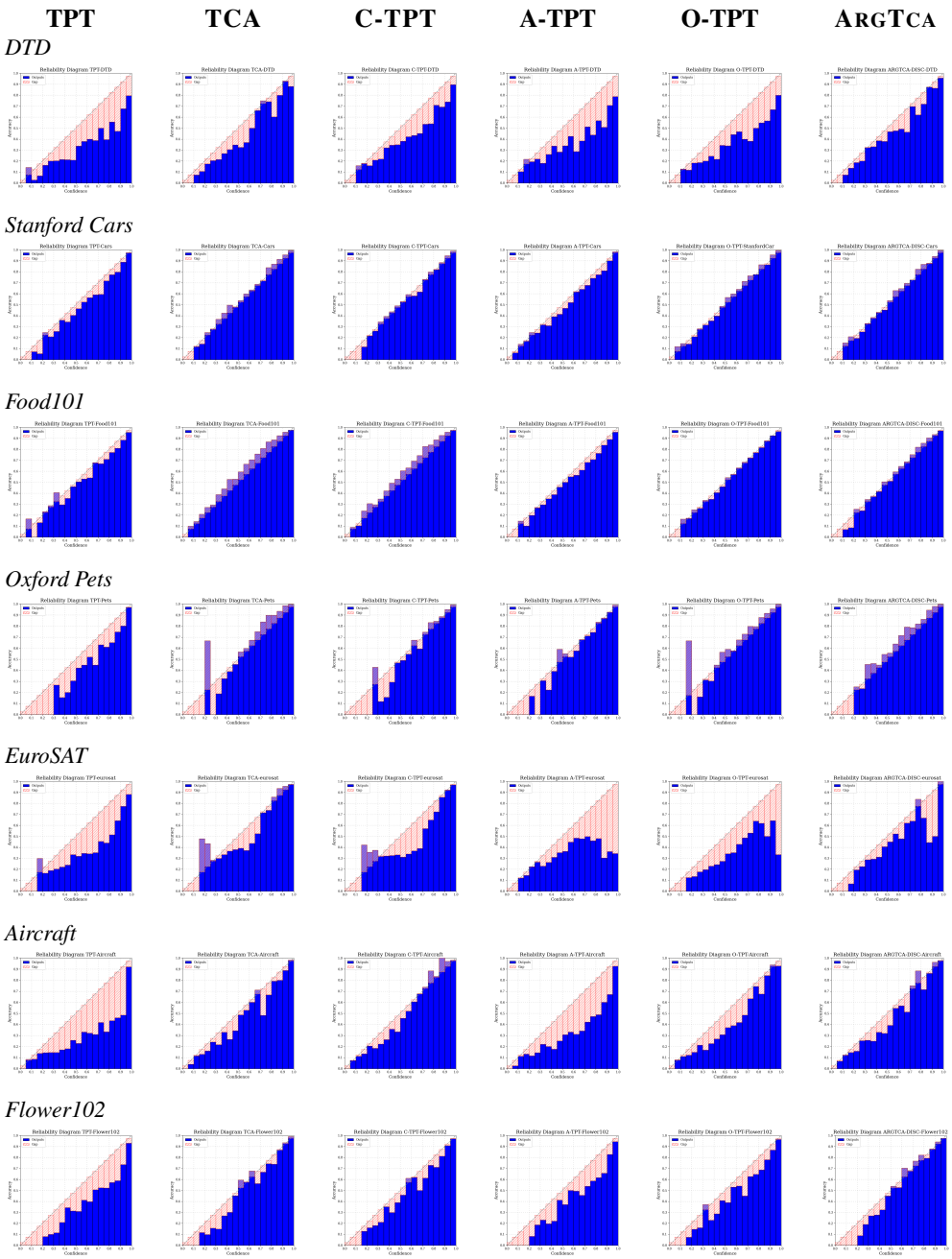


Figure 8: Reliability diagrams across all datasets using ViT-B/16. Each row is a dataset; each column is a method. Bars above/below the diagonal (pink) indicate under/overconfidence. ARGTCA consistently shows better alignment with the diagonal across all datasets.

Attention and aggregation. For each edge $\{i, j\}$, layer l, and head q, let $\tilde{h}_i^{(l),q} = W_q^{(l)} h_i^{(l-1)}$ denote the head-q linear projection of node i's embedding. The normalized attention coefficient is:

$$\alpha_{ij}^{(l,q)} = \frac{\exp\Big(\mathrm{LeakyReLU}\big(a_q^{(l)\top}[\tilde{h}_i^{(l),q} \| \tilde{h}_j^{(l),q}]\big)\Big)}{\sum_{j' \in \mathcal{N}(i)} \exp\Big(\mathrm{LeakyReLU}\big(a_q^{(l)\top}[\tilde{h}_i^{(l),q} \| \tilde{h}_{j'}^{(l),q}]\big)\Big)}. \tag{13}$$

Table 9: Notation used throughout the paper.

Symbol	Shape / Type	Meaning
K	scalar	Number of classes
M	scalar	Attributes per class
M'	scalar	Selected attributes for prompt injection
$N = K \times M$	scalar	Total graph nodes
$d = 512$	scalar	CLIP token embedding dimension
c_k	string	Class name of class k
a_j	string	j-th attribute string
(c_k, a_j)	tuple	Graph node
$\mathbf{h}_i^{(0)}$	$\mathbb{R}^d$	Initial node embedding
$\mathbf{h}_i^{(L)}$	$\mathbb{R}^d$	Final GAT node embedding
$\mathbf{p}$	$\mathbb{R}^{n_{\text{ctx}} \times d}$	Learnable soft prompt
$g(\cdot)$	$\mathbb{R}^d$	CLIP text encoder (EOS feature)
$f(\cdot)$	$\mathbb{R}^{d_{\text{out}}}$	CLIP projection head
τ	scalar	SCL temperature
τ^*	scalar	CLIP's learned logit scale
α, β	scalars	Inter-/intra-class loss weights

Table 10: One-time offline GAT training time for each dataset. Each graph contains 10 attribute nodes per class. The trained graph is reused during test-time prompt tuning.

Dataset	Classes	Graph Nodes	GAT Training
EuroSAT	10	100	∼20 s
OxfordPets	37	370	∼55 s
DTD	47	470	∼70 s
FGVCAircraft	100	1,000	∼2.6 min
Caltech101	100	1,000	∼2.6 min
Flowers102	102	1,020	∼2.6 min
Food101	101	1,010	∼2.5 min
UCF101	101	1,010	∼2.6 min
StanfordCars	196	1,960	∼5.0 min
ImageNet	1000	10,000	∼25.0 min

Node representations are updated as:

$$h_i^{(l)} = \text{ELU}\left(\text{concat}_{q=1}^{H}\left[\sum_{j \in \mathcal{N}(i)} \text{drop}\big(\alpha_{ij}^{(l,q)}\big) \cdot \tilde{h}_j^{(l),q}\right]\right). \tag{14}$$

Residual and Normalization After L layers, a residual connection (with W_{res} initialized to $\mathbf{I}$) mitigates over-smoothing, followed by ℓ_2 normalization:

$$h_i^{(L)} = \text{norm}\Big(h_i^{(L)} + W_{\text{res}}\, h_i^{(0)}\Big). \tag{15}$$

C Test-Time Prompt Tuning (Phase 2)

Phase 2 follows the attribute-conditioned test-time prompt tuning objective of TCA (Hebbalaguppe et al., 2025). The graph stage is used only to select the attribute tokens supplied to the downstream prompt. The learned GAT node embeddings are not inserted into the prompt and are not used during test-time optimization.

For class c_k and a graph-selected attribute a_m^k, the prompt is assembled as

$$t_{k,m} = \mathbf{p} \oplus \mathbf{a}_m^k \oplus \mathbf{c}_k, \tag{16}$$

where $\mathbf{p}$ is the soft context and $\oplus$ denotes token concatenation. Only $\mathbf{p}$ is learnable in Phase 2; the attribute tokens, class-name tokens, CLIP image encoder, and CLIP text encoder remain frozen. Let

$$f_{k,m} = \frac{g(t_{k,m})}{\|g(t_{k,m})\|_2} \tag{17}$$

denote the normalized CLIP text feature for the m-th selected attribute of class c_k, and let v denote the normalized image feature. The bag-of-attributes class probability sums exponentiated image-text similarities over the M' selected attributes:

$$P(c_k \mid v) = \frac{\sum_{m=1}^{M'} \exp\big(v^\top f_{k,m}/\tau^*\big)}{\sum_{k'=1}^{K}\sum_{m=1}^{M'} \exp(v^\top f_{k',m}/\tau^*)}, \tag{18}$$

where τ^* denotes the CLIP logit logit scale, distinct from the Phase 1 supervised-contrastive temperature.

As shown in Figure 9, for each test image, B augmented views are generated. We retain the $\lfloor Bp_{\text{sel}} \rfloor$ lowest-entropy views, average their predictions to obtain $\bar{P}$, and minimize the entropy term

$$\mathcal{L}_{\text{TPT}} = -\sum_{k=1}^{K} \bar{P}_k \log \bar{P}_k. \tag{19}$$

The TCA regularizers operate on the attribute-conditioned text features. We define the class and global centroids as

$$\mu'_k = \frac{1}{M'}\sum_{m=1}^{M'} f_{k,m}, \qquad \mu = \frac{1}{K}\sum_{k=1}^{K} \bar{\mu}_k. \tag{20}$$

The intra-class and inter-class terms are

$$\mathcal{L}_{\text{intra}} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{M'}\sum_{m=1}^{M'} \|f_{k,m} - \mu'_k\|_2 , \tag{21}$$

$$\mathcal{L}_{\text{inter}} = \frac{1}{K}\sum_{k=1}^{K} \|\mu'_k - \mu\|_2 . \tag{22}$$

The Phase 2 objective is

$$\mathcal{L}_{\text{TCA}} = \mathcal{L}_{\text{TPT}} - \alpha\mathcal{L}_{\text{inter}} + \beta\mathcal{L}_{\text{intra}}. \tag{23}$$

Minimizing $\mathcal{L}_{\text{intra}}$ contracts same-class attribute features around their class centroid, while the negative sign on $\mathcal{L}_{\text{inter}}$ maximizes dispersion among class centroids.

At inference time, one AdamW step is taken on $\mathbf{p}$ for each test image. The final prediction is then computed on the unaugmented image. The prompt $\mathbf{p}$ is reset before processing the next image, so test samples do not share adapted prompt state. Our experiments use $B = 64$, $p_{\text{sel}} = 0.10$, $M' = 2$, $\alpha = 10$, $\beta = 35$, and $n_{\text{ctx}} = 4$.

D Robustness to Natural Distribution Shifts

Following (Radford et al., 2021), we evaluate robustness under natural distribution shift on four ImageNet variants: ImageNet-A (Hendrycks et al., 2021b), ImageNet-V2 (Recht et al., 2019), ImageNet-R (Hendrycks et al., 2021a), and ImageNet-S (Wang et al., 2019). Results are in Table 11, reported per fixed variant as in Table 1. Focusing on ECE as the primary metric, ArgTca-Disc attains the lowest ECE on ImageNet-A (3.35 vs. 7.22 for the next-best, O-TPT) and ImageNet-S (6.61 vs. 6.87), while ArgTca-Div achieves the lowest ECE on ImageNet-V2 (3.94 vs. 3.97 for O-TPT). On ImageNet-R, however, O-TPT remains substantially better (1.46 vs. 7.76 for ArgTca-Disc), indicating that the calibration benefit of ARGTCA is shift-dependent. Averaged over the four shifts, neither variant matches O-TPT's ECE (4.88; ArgTca-Disc 5.42, ArgTca-Div 5.92), although ArgTca-Disc provides the stronger average calibration of the two ARGTCA variants. Overall, taking the better-performing ArgTca variant per benchmark, ArgTca attains the lowest ECE on 3 of 4 datasets.

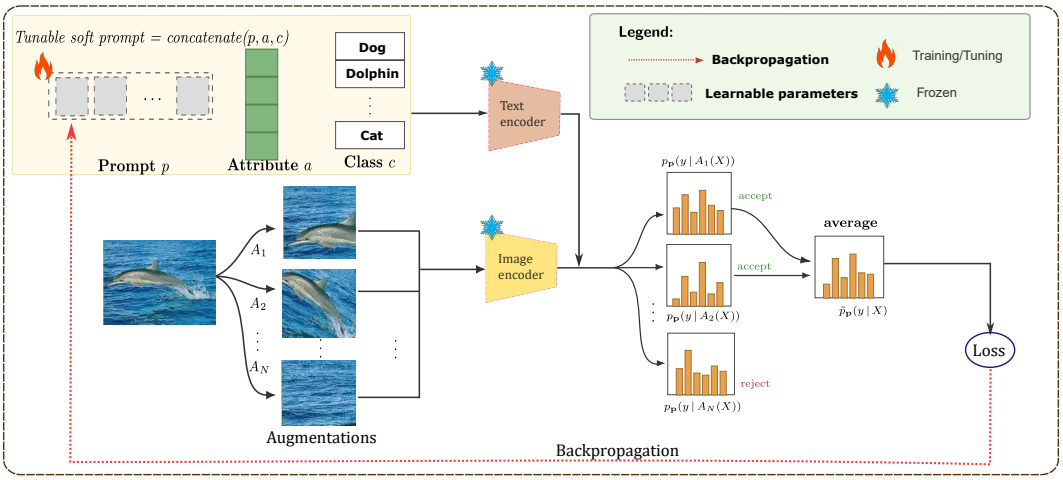


Figure 9: Test-time prompt calibration.

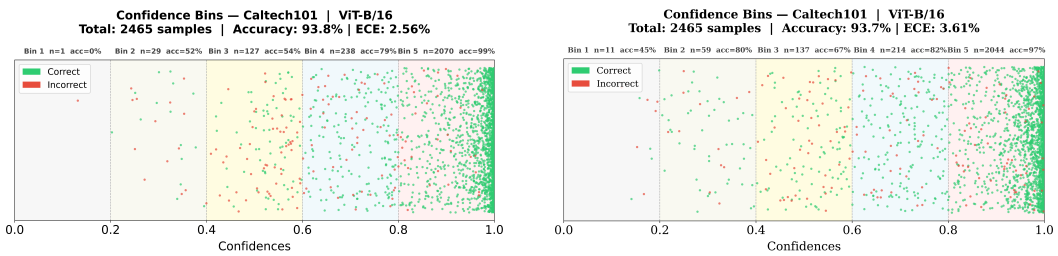
Table 11: Top-1 accuracy (Acc., ↑) and Expected Calibration Error (ECE, ↓) on natural distribution shift datasets using the CLIP ViT-B/16 backbone. Taking the better-performing ARGTCA variant per benchmark, ARGTCA attains the lowest ECE on 3 of 4 shifts (ImageNet-A, -V2, -S), with O-TPT retaining the lowest ECE on ImageNet-R.

Method	Metric	ImageNet-A	ImageNet-V2	ImageNet-R	ImageNet-S	Average
Backbone: CLIP ViT-B/16 (512-d)						
TPT NeurIPS'22	Acc.	52.6	63.0	76.7	47.5	59.9
	ECE	16.4	11.1	4.36	16.1	12.0
C-TPT ICLR'24	Acc.	51.6	62.7	76.0	47.9	59.6
	ECE	8.16	6.23	1.54	7.35	5.82
TCA ECML'25	Acc.	46.95	59.94	72.78	45.1	56.19
	ECE	8.59	4.95	5.1	8.62	6.81
O-TPT CVPR'25	Acc.	49.87	61.65	72.55	47.12	57.80
	ECE	7.22	3.97	1.46	6.87	4.88
A-TPT ICLR'26	Acc.	51.31	62.76	76.20	46.90	59.29
	ECE	9.99	7.78	2.04	8.47	7.07
ARGTCA-DIV (Ours)	Acc.	46.05	59.19	71.78	44.89	55.48
	ECE	10.06	3.94	2.69	6.98	**5.92**
ARGTCA-DISC (Ours)	Acc.	47.96	60.35	72.58	46.16	56.76
	ECE	3.35	3.96	7.76	6.61	**5.42**

E MCE AND BRIER SCORE

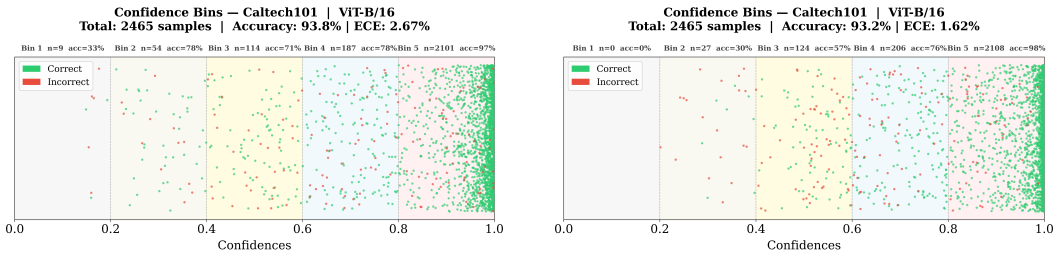
Table 12 reports Maximum Calibration Error (MCE) and Brier score on nine benchmarks, compared against TCA (the closest competitor). ARGTCA-DIV and ARGTCA-DISC outperform TCA on 5 of 9 and 6 of 9 datasets, respectively, under MCE, and on 4 of 9 and 5 of 9 datasets under Brier score, indicating gains that are directionally consistent with, but less pronounced than, those observed under ECE.

The MCE value of 91.07 for ARGTCA-DIV on Caltech101 reflects a sparse-bin artifact rather than a genuine calibration failure: MCE is a maximum over bins and is therefore dominated by any single sparsely populated bin. Excluding Caltech101 from every method for a like-for-like comparison, ARGTCA-DIV's average MCE falls from 24.61 to 16.30, ahead of TCA's 17.89 on the same eight datasets. ARGTCA-DISC, which is unaffected by this artifact, attains the best average MCE across all nine datasets (19.28 versus TCA's 19.45) without any exclusion.



(a) TCA

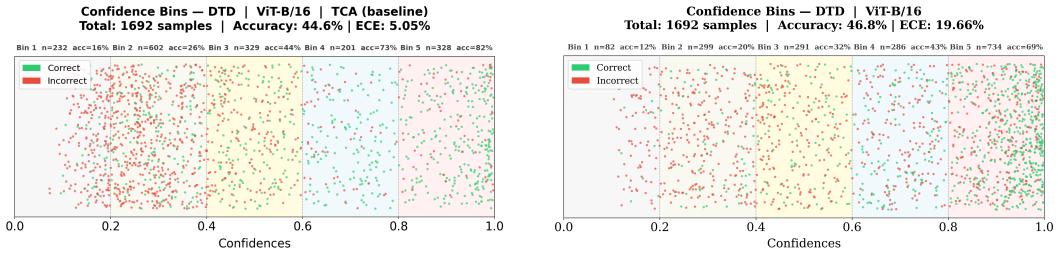
(b) O-TPT



(c) A-TPT

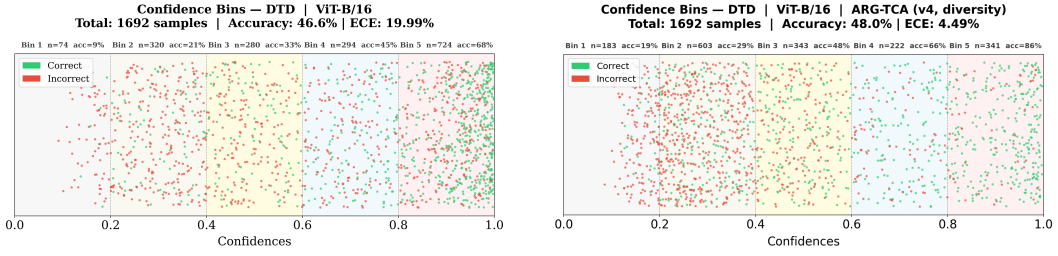
(d) ARGTCA-DISC (Ours)

Figure 10: Confidence-bin plots on Caltech101 (ViT-B/16). Green/red indicate correct/incorrect predictions. ARGTCA-DISC achieves the lowest ECE, with improved separation between correct and incorrect predictions across confidence bins.



(a) TCA

(b) O-TPT



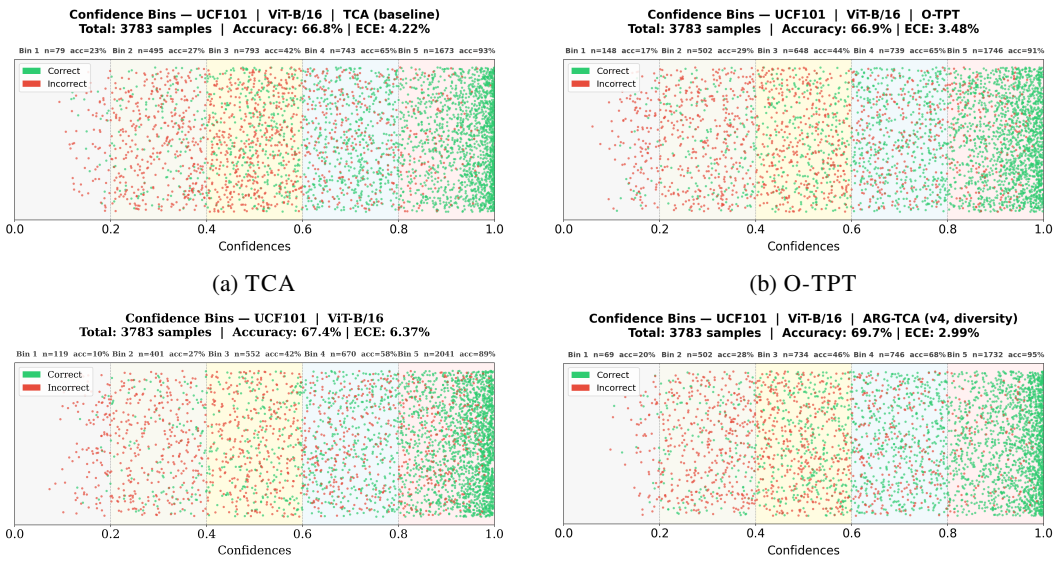
(c) A-TPT

(d) ARGTCA-DIV (Ours)

Figure 11: Confidence-bin plots on DTD (ViT-B/16). Green/red indicate correct/incorrect predictions. ARGTCA-DIV achieves the lowest ECE, with improved separation between correct and incorrect predictions across confidence bins.

Under the Brier score, the ordering differs: TCA attains the best average (0.4479 versus 0.4536 for ARGTCA-DISC and 0.4576 for ARGTCA-DIV), even though ARGTCA-DISC wins on 5 of 9 individual datasets. Brier score couples calibration with accuracy through its squared-error term, so it can favor sharper predictions even when they are less well calibrated.

Both variants show a consistent weakness on EuroSAT across all three calibration metrics considered in this work (ECE, MCE, Brier). We attribute this to EuroSAT's small label space (10 coarse land-cover classes) and correspondingly limited attribute vocabulary, which yields the sparsest cross-class



(a) TCA (b) O-TPT

(c) A-TPT (d) ARGTCA-DIV (Ours)

Figure 12: Confidence-bin plots on UCF101 (ViT-B/16). Green/red indicate correct/incorrect predictions. ARGTCA-DIV achieves the lowest ECE, with improved separation between correct and incorrect predictions across confidence bins.

Table 12: Maximum Calibration Error (MCE, ↓) and Brier Score (↓) on visual classification benchmarks using CLIP ViT-B/16. Lower values indicate better calibration; **bold** marks the best value. ‡The Caltech101 MCE of 91.07 reflects a sparse-bin artifact rather than a genuine calibration failure; excluding it, ARGTCA-DIV's average MCE is 16.30 against 17.89 of TCA on same eight datasets.

Method	Metric	Caltech	Pets	Cars	Flower	Aircraft	DTD	EuroSAT	UCF	Food101	Avg.
Backbone: CLIP ViT-B/16 (512-d)											
TCA	MCE	31.88	42.44	9.70	16.64	18.41	13.00	19.64	15.06	8.26	19.45
ECML'25	Brier	0.1013	0.1443	0.4479	0.4113	0.8458	0.7073	0.7062	0.4333	0.2341	**0.4479**
ARGTCA-DIV	MCE	91.07‡	22.67	8.26	18.75	15.14	18.44	28.28	10.97	7.91	24.61‡
(Ours)	Brier	0.1075	0.1510	0.4482	0.4075	0.8441	0.6826	0.8287	0.4075	0.2417	0.4576
ARGTCA-DISC	MCE	27.56	23.69	8.36	17.72	17.74	21.71	35.62	13.51	7.59	**19.28**
(Ours)	Brier	0.1008	0.1599	0.4402	0.3977	0.8471	0.6808	0.8150	0.4060	0.2348	0.4536

attribute sharing among our benchmarks and therefore provides little relational structure for the graph to exploit. ARGTCA-DISC additionally underperforms TCA on DTD under MCE (21.71 vs. 13.00), indicating that calibration improvements are metric-dependent and should be interpreted alongside the specific evaluation criterion used.

F COMPUTATIONAL OVERHEAD OF THE SAG AND GAT TRAINING PHASE

To empirically validate the computational cost of the offline attribute-graph preparation phase, we measured GAT training time across benchmark datasets on a single NVIDIA A100 80GB GPU. Regressing training time against the number of graph nodes (number of classes $\times$ 10 attributes per class) reveals a near-perfectly linear relationship ($R^2 = 0.998$; slope ≈ 0.15 s/node, intercept ≈ 1.1 s), as shown in figure 6; per-dataset runtimes are listed in Table 10. This fit is corroborated across datasets spanning nearly a $20\times$ range in class count — EuroSAT (10 classes, 100 nodes, $\approx$20 s) and StanfordCars (196 classes, 1,960 nodes, $\approx$5.0 min) both lie almost exactly on the same linear trend, as do the intermediate-sized datasets (OxfordPets, DTD, Aircraft, Caltech101, Flower102, Food101, UCF101), whose training times of 55 s–2.6 min match the fitted line. This confirms that GAT training scales linearly with the number of sampled (class, attribute) nodes rather than exhibiting the high cost typical of larger, densely connected graphs. This cost is incurred exactly once per dataset



Figure 13: **ECE sensitivity to graph size M (left) and number of selected attributes M' (right)** for ARGTCA-DIV (DIV) and ARGTCA-DISC (DISC) on Caltech101 (Cal), OxfordPets (Pets), and DTD. Left: M' fixed at 2; right: M fixed at 10.

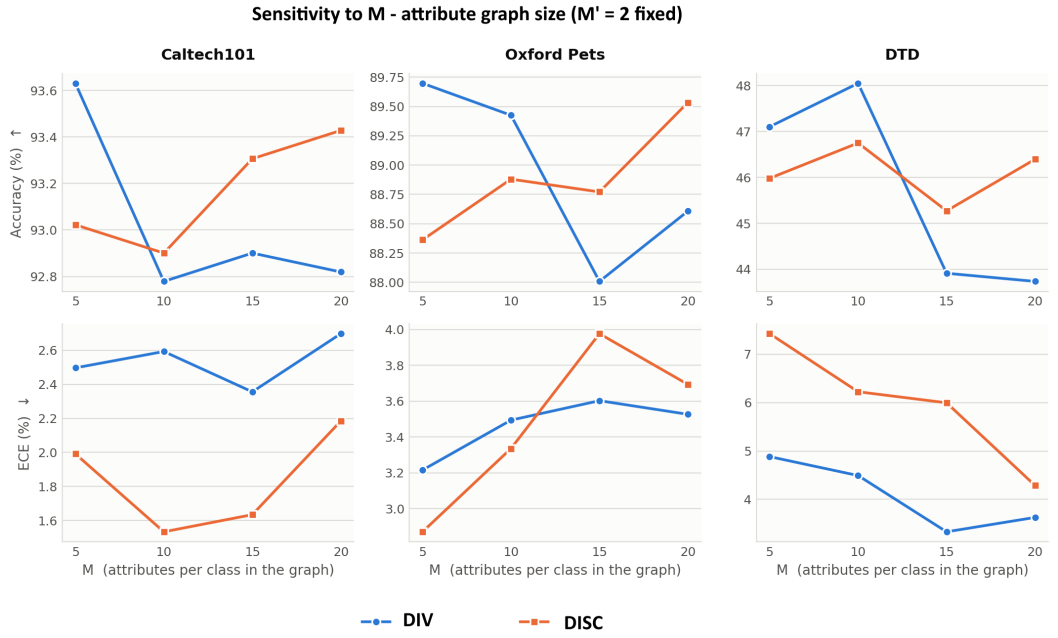


Figure 14: M sensitivity

— the resulting graph-refined attributes are reused across the subsequent test-time prompt-tuning runs. Combined with the negligible one-time cost of LLM attribute generation (a single prompt call), this analysis confirms that the ARGTCA does not incur substantial computational cost: the entire offline preparation phase is a small, linear, one-time process, and does not compromise the practicality of deploying ARGTCA at scale.

G SENSITIVITY TO M AND M'

ARGTCA has two attribute counts. M is the number of attributes per class admitted to the graph, so Phase 1 operates on $K \times M$ nodes and M fixes the candidate pool the selection criteria rank. M' is the number of attributes the criteria actually select per class; each selected attribute forms its own prompt, so $K \times M'$ prompts are encoded at every test-time tuning step. Phase-1 cost scales with M, test-time cost with M'; the two are independent and we sweep them separately: $M' \in \{2, 3, 4, 5, 6\}$ at the default $M = 10$, and $M \in \{5, 10, 15, 20\}$ at the default $M' = 2$. All runs use Caltech101, OxfordPets and DTD with ViT-B/16, the two-layer four-head GAT under $\mathcal{L}_{\text{SupCon}}$ ($\tau = 0.07$, 100

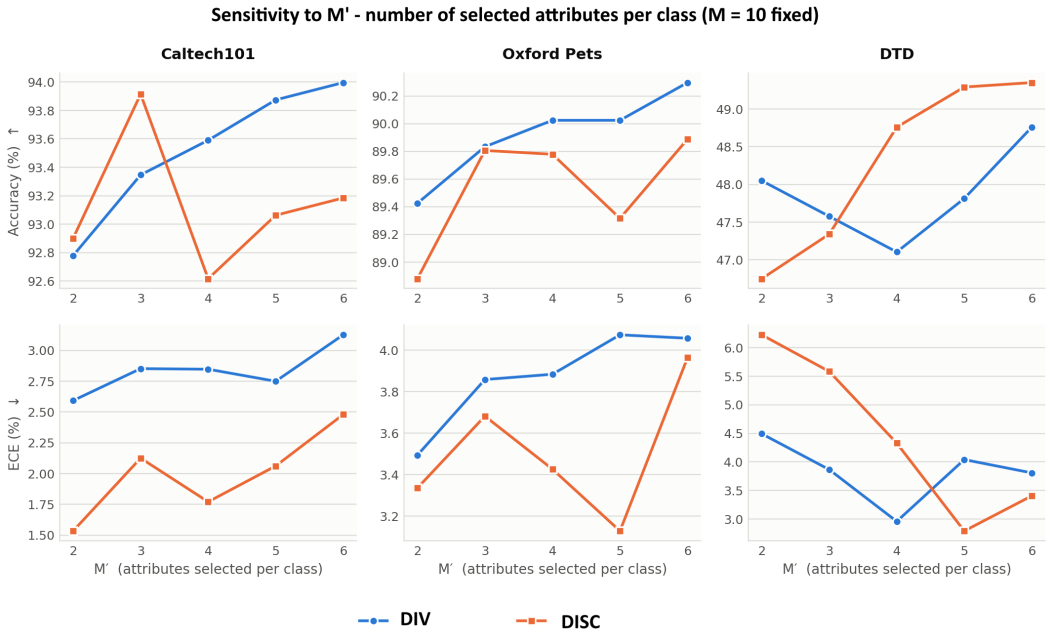


Figure 15: M' sensitivity

epochs), unmodified Phase 2 TCA ($\alpha = 10$, $\beta = 35$, one step, selection fraction 0.1); every other quantity is held fixed.

Number of selected attributes M' (Figure 15). The two criteria trade off in opposite directions. ArgTCA-Disc converts additional selected attributes into calibration gains where the task is hardest: on DTD, ECE falls monotonically ($6.36 \to 5.58 \to 4.33 \to 2.79$ for $M' = 2, \ldots, 5$) while accuracy simultaneously rises ($46.75 \to 49.29$); the best mean configuration is $M' = 5$ (mean ECE $3.70 \to 2.66$, mean accuracy $76.18 \to 77.22$). ArgTCA-Div, in contrast, trades calibration for accuracy: on Caltech101, accuracy rises $92.78 \to 94.00$ while ECE worsens $2.59 \to 3.13$, and on OxfordPets accuracy rises $89.42 \to 90.30$ while ECE worsens $3.49 \to 4.06$, leaving mean ECE effectively flat across M'. Brier improves monotonically in M' for both criteria on all three datasets — the cleanest single trend in the sweep. Any gain is paid for linearly at inference.

Graph size M (Figure 14). Graph size is effectively immaterial: mean ECE moves only ~ 0.4–0.7 points across $M = 5, \ldots, 20$ for either criterion, non-monotonically and with no consistent per-dataset direction. The one large effect is a DTD accuracy collapse under ArgTCA-Div ($48.05 \to 43.91 \to 43.74$ at $M = 10, 15, 20$); the accompanying lower ECE is the familiar lower-accuracy–lower-confidence artefact rather than a genuine calibration gain. This is structurally expected under the `same-class OR same-attribute` edge rule: each class forms an internally complete clique, so within-class edges grow as $K\,M(M-1)/2$, while cross-class edges—the only channel carrying attribute sharing *between* classes—grow only linearly in M. On Caltech101, the within-to-cross ratio moves from $1.5:1$ at $M = 5$ to $6.6:1$ at $M = 20$ (2,000 nodes: 19,000 within-class, 2,899 cross-class edges). Enlarging the candidate pool thus dilutes the cross-class signal rather than enriching it.

Takeaway. Performance is relatively less sensitive to the candidate pool size M, while M' provides a useful calibration–accuracy trade-off. We therefore use $M = 10$ as the default, as larger values show limited impact while increasing computational cost. We set $M' = 2$ to match TCA and enable direct comparison. Both sweeps use a single seed.

Table 13: **Backbone configurations.**

Backbone	Layers	Hidden	Heads	Params	Inductive
GAT	2	512	4×128	526,336	Yes
GCN	2	512	—	524,288	No
GraphSAGE	2	512	—	1,048,576	Yes

Table 14: **Reproducibility across seeds.** Standard deviation of dataset-mean accuracy over three seeds. GAT consistently shows the lowest variance across both selection criteria.

	Acc. std. across seeds		Relative to GAT	
Backbone	ARGTCA-DIV	ARGTCA-DISC	ARGTCA-DIV	ARGTCA-DISC
GAT	0.099	0.117	1.0×	1.0×
GCN	0.338	0.415	3.4×	3.5×
GraphSAGE	0.780	1.018	7.9×	8.7×

H GRAPH NEURAL NETWORK ABLATION

H.1 BACKBONE CONFIGURATIONS

All three backbones share the wrapper $h^{(0)} \to [\text{layer, ELU}]^{\times L} \to h^{(L)} + h^{(0)} \to \ell_2$-norm, with $L = 2$ layers of hidden width 512 (the CLIP token dimension), dropout 0.1, Xavier-uniform initialization and no biases. They differ only in how a node aggregates its neighborhood:

- **GAT** (Veličković et al., 2018) computes $e_{ij}^h = \text{LeakyReLU}_{0.2}\big(\mathbf{a}_h^\top [\mathbf{W}h_j \,\|\, \mathbf{W}h_i]\big)$ per head h and normalizes it with a softmax over the incoming neighborhood $\mathcal{N}(i)$, giving $h_i' = \big\|_h \sum_{j \in \mathcal{N}(i)} \alpha_{ij}^h \mathbf{W} h_j$. We use 4 heads of 128 dimensions each ($4 \times 128 = 512$). Per layer this is one 512×512 matrix plus $4 \times 256 = 1{,}024$ attention parameters.
- **GCN** (Kipf & Welling, 2017) replaces learned attention with the fixed symmetric renormalization $H' = \tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}HW$, $\tilde{A} = A + I$. Edge coefficients are $1/\sqrt{\deg(i)\deg(j)}$: a function of degree alone, with no dependence on node content. This is the precise ablation of attention, removing the learned, content-dependent weighting while leaving the neighborhood, the linear map and the normalization structure intact.
- **GraphSAGE** (Hamilton et al., 2017) uses the mean aggregator $h_i' = \mathbf{W}_{\text{self}} h_i + \mathbf{W}_{\text{neigh}} \,\text{mean}_{j \in \mathcal{N}(i)} h_j$. No self-loop is added to the neighborhood—the node's own features enter only through $\mathbf{W}_{\text{self}}$—so the model can trade self-information against neighborhood information, but still cannot weight individual neighbors against one another. We omit the per-layer ℓ_2 normalization of the original formulation so that the wrapper matches the other two arms exactly.

Inductive versus transductive. GAT and GraphSAGE are *inductive*: a node's update is a shared function of node features and a local neighborhood, so the learned parameters transfer to nodes—or whole graphs—unseen during training (Veličković et al., 2018; Hamilton et al., 2017). The GCN as originally formulated is *transductive*: its propagation operator $\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}$ is defined on one fixed graph and its coefficients change with node degree, so embeddings are undefined for new nodes without recomputation (Kipf & Welling, 2017). This distinction has *no* operational consequence for the results below: Phase 1 runs offline on a fixed graph of $K \times M$ nodes and the network is discarded before test time, so the pipeline is transductive by construction under any backbone. It matters for extensibility instead—adding a class introduces new nodes, which an inductive backbone could in principle embed without retraining Phase 1, whereas the GCN would require a rebuild. We did not test inductive generalization, so we report this as a property of the model family, not an experimental finding.

Table 15: **Backbone ablation on the attribute graph.** GAT is replaced by GCN and GraphSAGE with all configurations held fixed; only the aggregation rule changes. Means ± standard deviation over three seeds, averaged across Caltech101, OxfordPets, and DTD; best per column in bold. GAT is better calibrated under ARGTCA-DISC and markedly more stable across seeds.

Selection	Backbone	Acc. ↑	ECE ↓	Brier ↓
ARGTCA-DIV (diversity)	GAT	$\mathbf{76.74 \pm 0.10}$	$\mathbf{3.49 \pm 0.22}$	$\mathbf{0.3135 \pm 0.0017}$
	GCN	76.31 ± 0.34	3.55 ± 0.16	0.3161 ± 0.0048
	GraphSAGE	75.39 ± 0.78	4.17 ± 0.32	0.3275 ± 0.0104
ARGTCA-DISC (discrim.)	GAT	76.25 ± 0.12	3.91 ± 0.25	$\mathbf{0.3128 \pm 0.0008}$
	GCN	$\mathbf{76.28 \pm 0.41}$	4.44 ± 0.33	0.3167 ± 0.0065
	GraphSAGE	75.71 ± 1.02	$\mathbf{3.69 \pm 0.76}$	0.3175 ± 0.0089

Table 16: **Seed-paired differences against GAT.** Differences are computed seed-by-seed and averaged; positive values favor the alternative backbone, with ± indicating standard deviation.

Selection	Backbone	ΔAcc.	ΔECE	ΔBrier
ARGTCA-DIV	GCN	-0.43 ± 0.43	-0.06 ± 0.07	-0.0026 ± 0.0065
	GraphSAGE	-1.35 ± 0.86	-0.68 ± 0.43	-0.0140 ± 0.0118
ARGTCA-DISC	GCN	$+0.03 \pm 0.45$	-0.53 ± 0.25	-0.0039 ± 0.0061
	GraphSAGE	-0.53 ± 1.10	$+0.21 \pm 0.52$	-0.0047 ± 0.0097

H.2 BACKBONE MODEL ANALYSIS

No backbone wins consistently on accuracy: Table 17 and Table 18 show GAT, GCN, and GraphSAGE each take the best accuracy or ECE on at least one dataset. Paired seed-by-seed (Table 16), GCN's accuracy difference from GAT is smaller than its own standard deviation under both selection criteria.

We select GAT for its stability across seeds. Table 14 reports the standard deviation across three seeds, averaged over the three datasets. GAT has the lowest variance under both selection criteria (0.10 ARGTCA-DIV, 0.12 ARGTCA-DISC), compared to 0.34/0.41 for GCN and 0.78/1.02 for GraphSAGE—a 3.4–$3.5\times$ and 7.9–$8.7\times$ reduction, respectively. This holds for both selection criteria, so it is not specific to one dataset or one criterion. We choose GAT on this basis: it gives the most consistent results across random seeds, which matters more than winning individual cells when the backbone cannot be re-tuned per deployment.

GAT is also better calibrated under ARGTCA-DISC: its ECE gap with GCN is the only difference in Table 16 exceeding two standard deviations (-0.53 ± 0.25), and GAT is never worse than either alternative on Brier score. GraphSAGE is the weakest backbone on accuracy under both criteria despite having twice the parameters of GAT or GCN (Table 13), so its lower stability is not explained by under-parameterization.

I REPRESENTATION GEOMETRY BEFORE AND AFTER GRAPH TRAINING

In Figure 16, each dot is a $(\text{class}, \text{attribute})$ node. Radial position is the node's distance from its own class centroid, and term this distance the node's *radius*. Color encodes *genericity* $g(a) = |\{c_k : a \in \mathcal{A}_k\}|$, the number of classes whose attribute set contains word a. Green rings denote nodes selected by the named criterion. h_0 and h_L differ only in whether attention aggregation occurred. Reported ρ is the Spearman correlation between genericity and radius. Figure 17 plots this relation directly—mean radius versus binned genericity.

Figure 16 places h_0 (before GAT training) beside h_L (after) for each dataset and criterion. Points are a t-SNE embedding of the unit residuals (cosine metric, perplexity 10, PCA initialization), coloured by genericity on a log scale—pale for a word used by a single class, dark for one shared by many—with green rings marking the attributes the named criterion selects *when scoring in that panel's space*.

Table 17: **Per-dataset results, selection ARGTCA-DIV (diversity).** Mean ± standard deviation over three seeds; best per column within each dataset in bold.

Dataset	Backbone	Acc. ↑	ECE ↓	Brier ↓
Caltech101	GAT	**92.95 ± 0.24**	**2.70 ± 0.43**	**0.1061 ± 0.0022**
	GCN	92.71 ± 0.75	3.07 ± 0.20	0.1096 ± 0.0099
	GraphSAGE	92.91 ± 0.20	3.13 ± 0.05	0.1107 ± 0.0022
OxfordPets	GAT	**89.47 ± 0.10**	3.55 ± 0.09	**0.1508 ± 0.0002**
	GCN	89.46 ± 0.22	**3.40 ± 0.09**	0.1513 ± 0.0020
	GraphSAGE	88.76 ± 0.59	4.00 ± 0.09	0.1626 ± 0.0097
DTD	GAT	**47.79 ± 0.30**	4.22 ± 0.39	**0.6837 ± 0.0033**
	GCN	46.77 ± 0.21	**4.17 ± 0.46**	0.6873 ± 0.0037
	GraphSAGE	44.48 ± 1.61	5.38 ± 1.05	0.7093 ± 0.0224

Table 18: **Per-dataset results, selection ARGTCA-DISC (discriminability).** Mean ± standard deviation over three seeds; best per column within each dataset in bold.

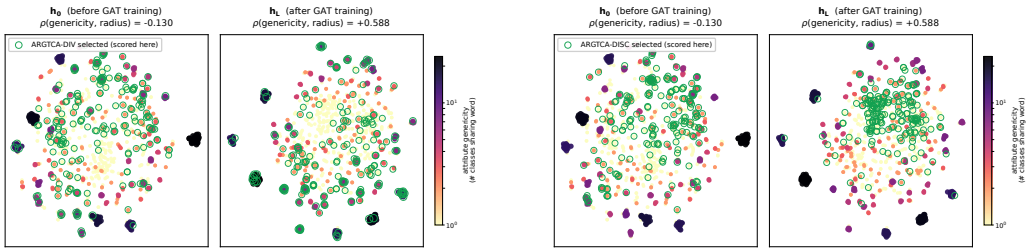
Dataset	Backbone	Acc. ↑	ECE ↓	Brier ↓
Caltech101	GAT	93.10 ± 0.28	1.94 ± 0.50	0.0969 ± 0.0048
	GCN	92.55 ± 0.71	1.91 ± 0.30	0.1061 ± 0.0072
	GraphSAGE	**93.37 ± 0.23**	**1.71 ± 0.19**	**0.0939 ± 0.0041**
OxfordPets	GAT	**88.91 ± 0.26**	3.16 ± 0.33	**0.1603 ± 0.0021**
	GCN	88.88 ± 0.59	**2.75 ± 0.69**	0.1622 ± 0.0091
	GraphSAGE	88.32 ± 0.34	3.22 ± 0.28	0.1653 ± 0.0062
DTD	GAT	46.73 ± 0.27	6.62 ± 0.56	**0.6812 ± 0.0016**
	GCN	**47.40 ± 0.20**	8.64 ± 0.76	0.6819 ± 0.0032
	GraphSAGE	45.45 ± 3.04	**6.15 ± 2.66**	0.6933 ± 0.0267

Each panel is annotated with ρ, the Spearman correlation between genericity and radius. The h_0 arm is a genuine counterfactual rather than a picture of the raw input: Phase 1 is re-run with training skipped so that $h_L := h_0$, and the same selection functions are then applied to it.

Figure 17 reports the same relation without t-SNE: mean radius, z-scored within each arm so the two curves are comparable, against genericity binned as 1, 2, 3–4, 5–8, 9+. It is criterion-independent—it describes the space, not the selection.

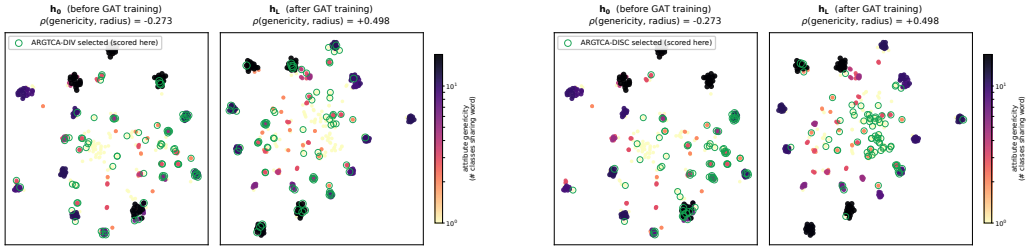
Genericity concentrates as radius after training. Genericity and radius are weakly negatively correlated in h_0 ($\rho = -0.130$ Caltech101, -0.102 DTD, -0.273 OxfordPets), and strongly positively correlated in h_L ($\rho = +0.588$, $+0.551$, $+0.498$; $p < 5 \times 10^{-94}$): training inverts the relationship between how many classes share an attribute and how far that attribute sits from its own class centroid. In h_L, single-class attributes sit $\sim 0.6\sigma$ inside their centroid while attributes shared by ≥ 9 classes sit 0.4–1.0σ outside; in h_0, the ordering is reversed. Message passing does not create the underlying similarity—cross-class cosine similarity for shared attributes is essentially unchanged ($0.65 \rightarrow 0.54$)—it converts this pre-existing directional signal into a radial one, making genericity legible as distance from centroid.

Once this radius exists, the two selection criteria read it in opposite directions. Under ARGTCA-DISC, selected attributes (green rings in Figure 16) shift inward from the generic periphery toward the class-specific interior—mean genericity falls from h_0 to h_L ($2.15 \rightarrow 1.33$ Caltech101, $3.09 \rightarrow 1.45$ DTD, $4.95 \rightarrow 2.38$ OxfordPets)—consistent with a criterion that selects for discriminability once genericity is expressed as distance from centroid. ARGTCA-DIV moves the opposite way, toward the periphery ($2.80 \rightarrow 5.25$, $4.22 \rightarrow 5.52$, $4.85 \rightarrow 6.01$).



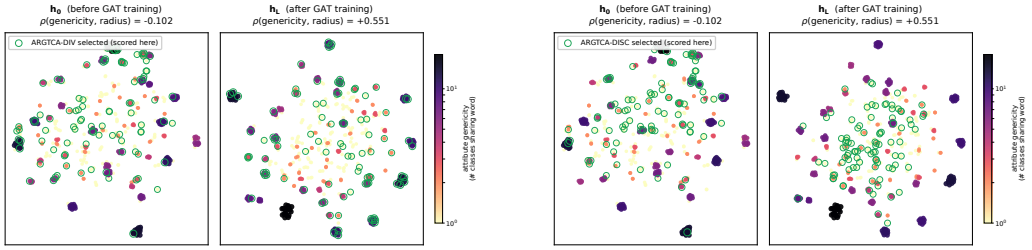
(a) ARGTCA-DIV Caltech101

(b) ARGTCA-DISC Caltech101



(c) ARGTCA-DIV OxfordPets

(d) ARGTCA-DISC OxfordPets



(e) ARGTCA-DIV DTD

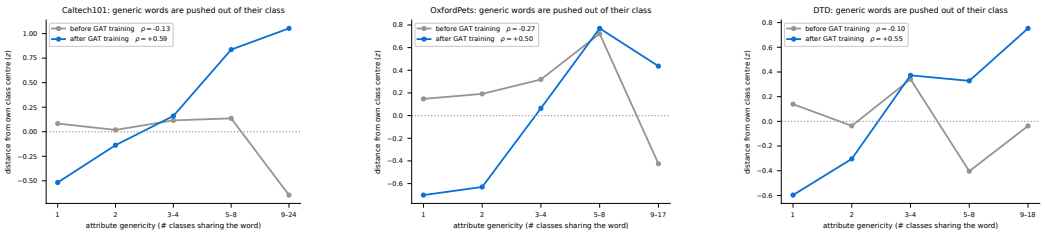
(f) ARGTCA-DISC DTD

Figure 16: **Attribute genericity becomes a radius after graph training.** t-SNE of the class-centered residual embeddings (cosine metric, perplexity 10) for each dataset, h_0 (before GAT training) beside h_L (after). One point per (class, attribute) node, coloured by genericity—the number of classes sharing that attribute word, log scale, pale = class-specific, dark = widely shared. Green rings mark the attributes the named criterion selects when scoring in that space. ρ is the Spearman correlation between genericity and distance from the node's own class center, computed in the full 512-d space. In h_L, ARGTCA-DISC concentrates on the pale interior while ARGTCA-DIV spreads toward the dark periphery; in h_0 neither is ordered. Seed 0.

J THEORETICAL ANALYSIS

This appendix provides proofs for the geometric characterizations of ARGTCA-DIV and ARGTCA-DISC stated as Proposition 1 and Corollary 1 in Section 5, together with a separate confidence bound (Proposition 2 below) relating class-level margin to prediction confidence on correctly classified examples. Proposition 1 and its corollary describe the geometry of the two selection criteria, and Proposition 2 bounds confidence conditional on correctness. The calibration behavior reported in Section 5 is established empirically.

Proof of Proposition 1(a) (Intra-class diversity). For $M' = 2$, let $\tilde{h}_i, \tilde{h}_j \in \mathbb{R}^d$ be normalized graph-refined embeddings of two candidate attributes from class c_k, and let $\mu_{ij} = (\tilde{h}_i + \tilde{h}_j)/2$. Since



(a) Caltech101 (b) OxfordPets (c) DTD

Figure 17: **Generic words are pushed out of their class.** Mean distance from a node's own class center (z-scored within each arm) against attribute genericity, binned as 1, 2, 3–4, 5–8, 9+. Before GAT training the relation is flat or negative and turns sharply down at the most generic bin; after training it is positive and near-monotone on all three datasets. Criterion-independent: this describes the embedding space, not the selection. Seed 0.

$\|\tilde{h}_i\|_2 = \|\tilde{h}_j\|_2 = 1,$

$$\|\mu_{ij}\|_2^2 = \frac{1}{4}\left\|\tilde{h}_i + \tilde{h}_j\right\|_2^2 \tag{24}$$

$$= \frac{1}{4}\left(\|\tilde{h}_i\|_2^2 + \|\tilde{h}_j\|_2^2 + 2\tilde{h}_i^\top \tilde{h}_j\right) \tag{25}$$

$$= \frac{1 + \tilde{h}_i^\top \tilde{h}_j}{2}. \tag{26}$$

Rearranging equation 26 gives

$$1 - \tilde{h}_i^\top \tilde{h}_j = 2(1 - \|\mu_{ij}\|_2^2)$$

.

Proof of Proposition 1(b) (Inter-class discriminability). For a candidate attribute $i \in \Omega_k$, let $\bar{\Omega}_k$ denote the nodes belonging to classes other than c_k, and define

$$s_i = \frac{1}{|\bar{\Omega}_k|}\sum_{j\in\bar{\Omega}_k}\left(1 - \tilde{h}_i^\top \tilde{h}_j\right), \qquad \mu_{\bar{k}} = \frac{1}{|\bar{\Omega}_k|}\sum_{j\in\bar{\Omega}_k}\tilde{h}_j.$$

Linearity of the inner product gives

$$s_i = 1 - \tilde{h}_i^\top \left(\frac{1}{|\bar{\Omega}_k|}\sum_{j\in\bar{\Omega}_k}\tilde{h}_j\right) \tag{27}$$

$$= 1 - \tilde{h}_i^\top \mu_{\bar{k}}, \tag{28}$$

which establishes the stated identity. Taking the difference between two candidate scores,

$$s_i - s_j = \left(1 - \tilde{h}_i^\top \mu_{\bar{k}}\right) - \left(1 - \tilde{h}_j^\top \mu_{\bar{k}}\right) = -(\tilde{h}_i - \tilde{h}_j)^\top \mu_{\bar{k}},$$

so ARGTCA-DISC ranks attributes by their projection away from the competing-class centroid $\mu_{\bar{k}}$.

Proposition 2 (Confidence lower bound on correctly classified examples)**.** *Let the multi-attribute class evidence for image feature v be*

$$Z_k(v) = \sum_{m=1}^{M'} \exp\left(\frac{v^\top f_{k,m}}{\tau^*}\right),$$

and define the corresponding class logit

$$\ell_k(v) = \tau^* \log Z_k(v)$$

For a sample whose true class y already attains the highest logit ($\ell_y(v) > \ell_k(v)$ for all $k \neq y$), define the class-level margin

$$\Delta_y(v) = \ell_y(v) - \max_{k\neq y} \ell_k(v) > 0.$$

Then its predicted probability satisfies

$$P(y \mid v) \geq \frac{1}{1 + (K-1)\exp(-\Delta_y(v)/\tau^*)}.$$

For fixed K and τ^*, a larger positive margin on an already-correct prediction yields a strictly larger lower bound on that prediction's confidence.

Proof. From the definition of $\ell_k(v)$,

$$P(y \mid v) = \frac{\exp(\ell_y(v)/\tau^*)}{\sum_{k=1}^{K}\exp(\ell_k(v)/\tau^*)} = \frac{1}{1 + \sum_{k\neq y}\exp\left(-\frac{\ell_y(v)-\ell_k(v)}{\tau^*}\right)}.$$

By definition of $\Delta_y(v)$,

$$\ell_y(v) - \ell_k(v) \geq \Delta_y(v), \qquad \forall k \neq y.$$

Therefore,

$$\sum_{k\neq y}\exp\left(-\frac{\ell_y(v)-\ell_k(v)}{\tau^*}\right) \leq (K-1)\exp\left(-\frac{\Delta_y(v)}{\tau^*}\right),$$

which gives the stated bound. □ □

Scope. Proposition 2 bounds $P(y \mid v)$ for samples where y is already the top-scoring class ($\Delta_y(v) > 0$), giving the margin-to-confidence relationship on correct predictions. The complementary behavior on incorrect predictions is reported empirically via the confidence-bin analysis (Figure 4) and reliability diagrams (Figure 1) in the main text.

Proof of Corollary 1 (Label-space dependence). Under approximately isotropic competing-class embeddings, $\mu_{\bar{k}}$ is the mean of $|\bar{\Omega}_k| = O(KM)$ approximately isotropically distributed unit vectors, so its norm concentrates at the standard rate for an average of n isotropic unit vectors

$$\|\mu_{\bar{k}}\|_2 = O(n^{-1/2}) = O((KM)^{-1/2})$$

Since the ARGTCA-DISC score is the affine functional $s_i = 1 - \tilde{h}_i^\top \mu_{\bar{k}}$ (Proposition 1b), its within-class dispersion decomposes as

$$\sigma_{\text{disc}}(k) = \text{std}_{i\in\Omega_k}\left(\tilde{h}_i^\top \mu_{\bar{k}}\right) = \|\mu_{\bar{k}}\|_2 \cdot \text{std}_{i\in\Omega_k}\left(\tilde{h}_i^\top \hat{\mu}_{\bar{k}}\right),$$

where $\hat{\mu}_{\bar{k}} = \mu_{\bar{k}}/\|\mu_{\bar{k}}\|_2$. The second factor is the spread of the M within-class embeddings along a single fixed direction and carries no dependence on K; all K-dependence therefore enters through $\|\mu_{\bar{k}}\|_2$, giving $\sigma_{\text{disc}} = O(K^{-1/2})$ for fixed M. In contrast, ARGTCA-DIV's dispersion margin (Proposition 1a) is computed over the fixed $\binom{M}{2}$ within-class candidate pairs, whose sample size is fixed by M and does not grow with K; no analogous concentration applies.

K EMPIRICAL ATTRIBUTE-SELECTION ANALYSIS

Structure of the Symbolic Attribute Graph. The SAG is determined entirely by the LLM-generated attribute vocabulary. Intra-class edges always form K disjoint cliques of size M=10, so their count $|E_{\text{intra}}| = K\binom{M}{2}$ is fixed by the label-space size and is identical in shape for every dataset. All cross-dataset structural variation is therefore carried by the inter-class edges, which exist only where two classes share an identical attribute token. Table 19 reports this decomposition.

The average number of inter-class neighbours per node spans a 23× range, from 1.26 on EuroSAT to 28.92 on StanfordCars. Counter-intuitively, the *fine-grained* datasets have the *densest* inter-class connectivity: their attribute vocabularies are dominated by category-generic tokens — `Automobile` appears in 116 of the 196 StanfordCars classes, and `aircraft` in 65 of the 100 FGVCAircraft classes. Accordingly, the fraction of nodes whose attribute token is shared with at least one other class is 85.5% for StanfordCars and 87.1% for FGVCAircraft, against 49.0% for EuroSAT. This is precisely the redundancy that ARGTCA-DISC is designed to suppress.

EuroSAT combines the smallest label space (K=10) with the sparsest inter-class sharing and the smallest attribute vocabulary ($|V|$=68), leaving the graph with little relational structure to exploit — consistent with EuroSAT being the benchmark on which both variants are weakest across all three calibration metrics.

Table 19: Symbolic Attribute Graph statistics, ordered by inter-class connectivity. $|V|$ is the number of distinct attribute tokens; E_{intra} and E_{inter} are undirected edge counts; deg_{inter} is the mean number of inter-class neighbours per node; *generic* is the fraction of nodes whose attribute token is shared with at least one other class.

Dataset	K	$\lvert V\rvert$	E_{intra}	E_{inter}	$\mathbf{deg}_{\text{inter}}$	**generic**
EuroSAT	10	68	450	63	1.26	49.0%
Caltech101	100	605	4,500	1,599	3.20	53.8%
DTD	47	239	2,115	904	3.85	62.5%
UCF101	101	468	4,545	2,725	5.40	67.4%
OxfordPets	37	157	1,665	1,060	5.73	70.5%
Food101	101	320	4,545	5,171	10.24	81.5%
FGVCAircraft	100	226	4,500	9,591	19.18	87.1%
Flowers102	102	264	4,590	11,701	22.94	82.9%
ImageNet	1,000	2,965	45,000	123,568	24.76	81.9%
StanfordCars	196	490	8,820	28,340	28.92	85.5%

Aggregation robustness. Over the ten benchmarks ARGTCA-DIV attains mean ECE 4.46, median 3.20 and mean rank 2.70 of six methods; ARGTCA-DISC attains 5.17/3.59/3.35. O-TPT attains a competitive median (3.55) but the worst mean rank (4.10), indicating markedly less consistent behavior.

Accuracy–calibration relationship. The correlation between ΔAcc and ΔECE (ARGTCA-DISC − ARGTCA-DIV) is Pearson r=0.67; excluding EuroSAT it reverses to $r = -0.25$, and the rank correlation is ρ=0.15 — a single high-leverage point rather than a systematic trade-off.

Gain over random selection. ARGTCA-DIV: Caltech101 2.52 vs. 2.53, OxfordPets 3.44 vs. 3.44, DTD 3.63 vs. 5.16. ARGTCA-DISC: Caltech101 1.62 vs. 2.43, OxfordPets 3.37 vs. 3.55, DTD 6.36 vs. 6.33.

Why the discriminability margin concentrates as $K^{-1/2}$. The discriminability score admits an exact closed form. Let

$$\bar{\mu}_k \;=\; \frac{1}{|\bar{\Omega}_k|} \sum_{j \in \bar{\Omega}_k} \tilde{h}_j \tag{29}$$

denote the centroid of all other-class embeddings. Because the score is an average of inner products with a *fixed* node i, the sum passes through:

$$\begin{aligned} s(a_i) &= \frac{1}{|\bar{\Omega}_k|} \sum_{j \in \bar{\Omega}_k} \big(1 - \tilde{h}_i^\top \tilde{h}_j\big) \\ &= 1 - \tilde{h}_i^\top \bar{\mu}_k. \end{aligned} \tag{30}$$

The score is therefore an *affine functional* of $\tilde{h}_i$: one minus its projection onto the other-class centroid. ARGTCA-DISC ranks attributes by Eq. 30, so the quantity that determines whether that ranking is informative is its within-class dispersion, The raw within-class dispersion characterizes the scale of the DISC scores, but does not by itself determine the induced ranking, since positive rescaling of $\bar{\mu}_k$ preserves the ordering.

$$\begin{aligned} \sigma_{\text{disc}}(k) &= \operatorname*{std}_{i \in \Omega_k} \big(\tilde{h}_i^\top \bar{\mu}_k\big) \\ &= \|\bar{\mu}_k\| \;\cdot \operatorname*{std}_{i \in \Omega_k} \big(\tilde{h}_i^\top \hat{\mu}_k\big), \end{aligned} \tag{31}$$

where $\hat{\mu}_k = \bar{\mu}_k / \|\bar{\mu}_k\|$ is the unit vector along the centroid. The first factor depends on K; the second does not. Indeed, the second factor is the spread of the M within-class embeddings along a single direction, and with M fixed at 10 it carries no dependence on the label-space size. All K-dependence therefore enters through $\|\bar{\mu}_k\|$, the norm of a mean of $n = |\bar{\Omega}_k| \approx (K-1)M$ unit vectors.

Table 20: Verification of Eq. 31. The discriminability margin σ_{disc} and other-class centroid norm $\|\bar{\mu}_k\|$ decay with K, while their ratio remains approximately constant. Values are averaged over classes.

Dataset	K	σ_{disc}	$\|\bar{\mu}_k\|$	$\sigma_{\text{disc}}/\|\bar{\mu}_k\|$
EuroSAT	10	0.0215	0.572	0.038
OxfordPets	37	0.0097	0.308	0.031
DTD	47	0.0135	0.393	0.034
Caltech101	100	0.0095	0.330	0.029
FGVCAircraft	100	0.0122	0.251	0.049
Food101	101	0.0104	0.342	0.030
UCF101	101	0.0108	0.313	0.035
Flowers102	102	0.0088	0.269	0.033
StanfordCars	196	0.0061	0.133	0.046
ImageNet	1,000	0.0023	0.080	0.029
log–log slope vs. K		−0.46	−0.43	−0.03
R^2		0.86	0.84	0.03

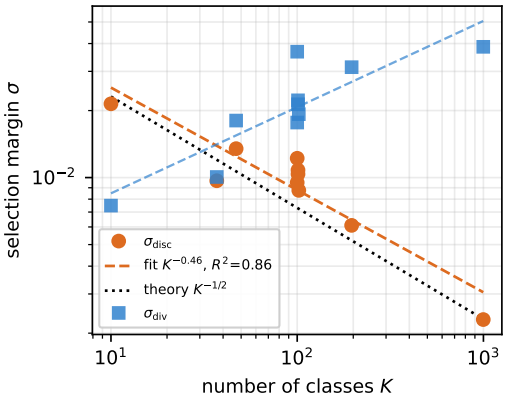


Figure 18: Selection margin vs. label-space size. The discriminability margin σ_{disc} concentrates as $K^{-0.46}$ (R^2=0.86), matching the predicted $K^{-1/2}$ rate, whereas the diversity margin σ_{div} does not.

Whenever those directions are not systematically aligned, this norm obeys the usual concentration rate $\|\bar{\mu}_k\| = \mathcal{O}(n^{-1/2}) = \mathcal{O}\big((KM)^{-1/2}\big)$, and hence

$$\sigma_{\text{disc}} \;\propto\; K^{-1/2}. \tag{32}$$

Intuitively, increasing label-space size concentrates the raw DISC scores through the shrinking competing-class centroid; the remaining selection signal is determined by the stability of its direction. We verify each link of this argument on the ten benchmarks (Table 20). First, Eq. 30 holds to machine precision (maximum deviation 1.2×10^{-7} in `float32`). Second, a log–log regression gives $\|\bar{\mu}_k\| \propto K^{-0.43\pm0.07}$ (R^2=0.84, p=2.2×10^{-4}), consistent with the predicted $-1/2$. Third — and most directly — once the centroid norm is divided out, the residual $\sigma_{\text{disc}}/\|\bar{\mu}_k\|$ shows *no* remaining trend in K (slope -0.03 ± 0.05, R^2=0.03, p=0.64), confirming that the observed concentration is entirely attributable to $\|\bar{\mu}_k\|$ rather than to any incidental property of the datasets. The measured margin itself scales as $\sigma_{\text{disc}} \propto K^{-0.46\pm0.07}$ (R^2=0.86, $p<10^{-4}$, 95% CI $[-0.59, -0.33]$), again containing $-1/2$.

By contrast, σ_{div} is computed from the $\binom{M}{2}$ *within*-class pairs alone. Its sample size is fixed by M, not K, so no analogous concentration applies; empirically it does not decrease with K. This asymmetry — one criterion averaging over a set that grows with the label space, the other over a set that does not — is the reason the two margins diverge as K increases.

Predicting the better variant. We tested fifteen candidate predictors (σ_{div}, σ_{disc}, their ratio, selection gaps, K, and eight graph statistics including cross-class edge density, isolated-node fraction and vocabulary genericity). None is a robust predictor of ΔECE (best rank-correlation p=0.28).

L Ablation Studies

L.1 Attribute comparison: TCA vs. ArgTca-Div vs. ArgTca-Disc.

The attribute-comparison table in the main paper lists the two attributes selected per method for a representative subset of classes from Caltech101, OxfordPets, and DTD. TCA ranks attributes by cosine similarity between each attribute's frozen CLIP embedding and the class name embedding, selecting the top-M'; ArgTca-Div selects the most complementary pair by minimum pairwise cosine similarity in $\mathbf{h}^{(L)}$ space; ArgTca-Disc selects the pair with highest mean cosine distance to all other-class embeddings in $\mathbf{h}^{(L)}$ space.

L.2 ZERO GRAPH EDGES.

To isolate the contribution of the SAG edge structure, we train the GAT with $\mathcal{E} = \emptyset$: all nodes are present and trained via supervised contrastive loss, but no message passing occurs. Without edges, each GAT layer reduces to an independent per-node linear transform; no cross-class or within-class structural context propagates. Attribute selection uses the same Div/Disc criteria on these edge-free embeddings. The calibration drop on DTD relative to the full model (see the ablation table in the main paper) confirms that the SAG's edge topology — not the contrastive objective alone — drives selection quality.

L.3 NO GAT TRAINING (RAW CLIP EOS EMBEDDINGS).

We skip Phase 1 entirely and apply the Div/Disc selection criteria directly to the initial embeddings $\mathbf{h}^{(0)}$ (frozen CLIP EOS hidden states), without any contrastive training. This tests whether GAT training adds value beyond the semantic geometry already present in CLIP's encoder. The DTD ECE degrades substantially (Disc: 8.95 vs. 6.36 for the full model), showing that contrastive relational training is necessary for reliable attribute discrimination on texture-heavy datasets.

L.4 RANDOM ATTRIBUTE SELECTION.

Replace Div/Disc selection with a random draw of $M' = 2$ attributes per class. This isolates the selection criterion: random selection degrades ECE on DTD (Div: 5.16 vs. 3.63 for the full model) while leaving Caltech accuracy largely unchanged, confirming that the geometric Div/Disc criteria are the operative mechanism for calibration improvement, not merely the graph training.

L.5 MLP ATTRIBUTE TRANSFORM.

To test whether graph topology is necessary beyond a learned per-node transformation, we replace the SAG with a two-layer MLP applied independently to each class-attribute embedding. The MLP maps each frozen CLIP EOS node embedding through a linear layer, GELU activation, and second linear layer, followed by L2 normalization, with no residual connection and no message passing between nodes. It is trained with the same supervised contrastive objective as the full graph model, and the resulting transformed embeddings are used with the same ArgTCA-Div/Disc selection criteria. On Caltech101, OxfordPets, and DTD, the results are reported in the ablation table in the main paper. The MLP variant improves over raw/no-training selection in some cases but remains below the full graph model, indicating that a learned nonlinear re-embedding alone is insufficient; relational graph propagation provides the additional structure needed for reliable attribute selection.

L.6 EXACT VERSUS SEMANTIC ATTRIBUTE MATCHING.

In the SAG, two nodes from different classes are connected only when they carry the same attribute token. Related tokens therefore remain unconnected. For example, the Caltech101 tokens `vintage` (class *camera*) and `antique` (class *gramophone*) share no edge. To test whether this exact matching rule is too restrictive, we compare three rules for the inter-class edges and keep the intra-class edges unchanged: (i) *Exact*, the rule used in ARGTCA; (ii) *Semantic*, which encodes each distinct token with the CLIP text encoder, mean-centers the token embeddings over the vocabulary of the dataset, and links each token to its most similar other token if their cosine similarity is at least τ=0.35. Two nodes from different classes are connected when their tokens are linked. A token is never linked to itself, so this rule contains no exact-match edges; (iii) *Hybrid*, the union of the exact and semantic edges. In (ii) and (iii), tokens are lowercased before matching. Mean-centering is needed because raw CLIP token embeddings are highly similar to one another (mean pairwise cosine similarity between 0.76 and 0.83 on these datasets). The threshold τ was fixed by inspecting token pairs before any evaluation. All other settings are the same as for the full model in Table 2.

Figure 19(a) shows that both alternative rules connect more nodes to other classes. On Caltech101, for example, the fraction of nodes with at least one inter-class neighbor rises from 53.8% with exact matching to 84.0% with the hybrid rule. This additional connectivity does not improve calibration (Table 21). On Caltech101 and OxfordPets, the three rules give similar accuracy and ECE. On DTD, both alternative rules increase ECE and reduce accuracy for both selection strategies. For example,

Table 21: Exact versus semantic matching for inter-class edges. Acc. (%,↑) and ECE (%,↓). The exact rule is the full ARGTCA model of Table 2. Avg. is the mean over the three datasets.

Rule	Sel.	Caltech101		OxfordPets		DTD		Avg.	
		Acc	ECE	Acc	ECE	Acc	ECE	Acc	ECE
Exact (ours)	Div	92.82	2.52	89.48	3.44	47.34	**3.63**	76.55	**3.20**
	Disc	93.19	**1.62**	88.91	**3.37**	46.63	6.36	76.24	3.78
Semantic	Div	92.66	2.39	89.02	3.58	45.63	6.92	75.77	4.30
	Disc	93.43	1.99	89.40	3.57	45.27	8.34	76.03	4.63
Hybrid	Div	92.94	2.49	88.88	3.44	45.98	5.39	75.93	3.77
	Disc	93.39	1.82	89.40	3.44	44.44	9.66	75.74	4.97

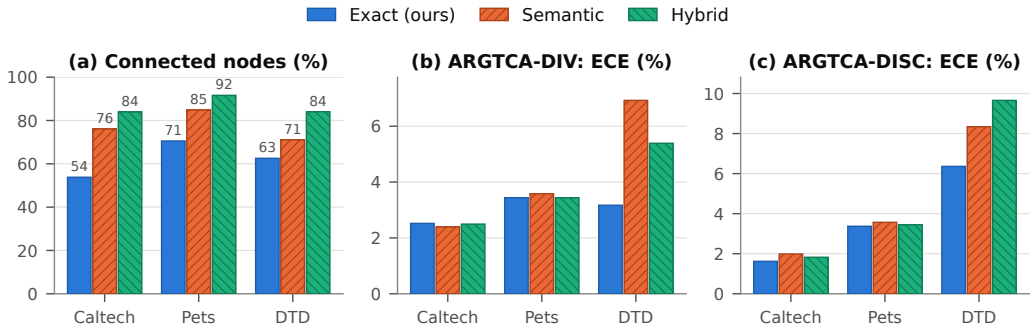


Figure 19: Exact versus semantic matching for inter-class edges. (a) Fraction of nodes with at least one inter-class neighbor. (b, c) ECE (%, lower is better) of ARGTCA-DIV and ARGTCA-DISC. Both alternative rules connect more nodes but give a higher average ECE.

the ECE of ARGTCA-DIV rises from 3.63 to 6.92 with the semantic rule. Averaged over the three datasets, exact matching gives the lowest ECE and the highest accuracy for both strategies. We therefore keep exact matching in ARGTCA. It is also simpler, since it needs no token embeddings and no similarity threshold.

L.7 HYPERPARAMETER SENSITIVITY (α, β).

Figures 20 and 21 show ECE as α and β are swept independently ($\alpha \in \{5, 10, 20, 40\}$ with β=35 fixed; $\beta \in \{15, 25, 35, 50\}$ with α=10 fixed), evaluated on Caltech101, OxfordPets, and DTD for both ARGTCA-DIV and ARGTCA-DISC. Both metrics are stable across the swept range: ECE varies by at most $\sim$1.5% on Caltech101 and $\sim$3.0% on DTD for α, and by at most $\sim$1.8% on DTD for β. The optimal region is $\alpha \in [10, 20]$ and $\beta \in [25, 35]$; the ViT-B/16 defaults (α=10, β=35) lie within this plateau, confirming robustness to precise hyperparameter choice.

M QUALITATIVE LABEL CORRECTIONS ACROSS DATASETS

Figures 22–24 show the selected annotated label-flip cases where graph-structured attribute selection changes the top prediction relative to TCA. The examples are grouped by selection strategy.

N BROADER IMPACT AND IMPLICATIONS

ARGTCA improves confidence calibration, not predictive accuracy; the two are distinct, and nothing here supports removing human review from a decision pipeline. The relational stage runs entirely offline from class names — no labeled validation set, no target-domain data, no inference-time cost — making it applicable in exactly the settings where post-hoc calibrators are unavailable. Its output

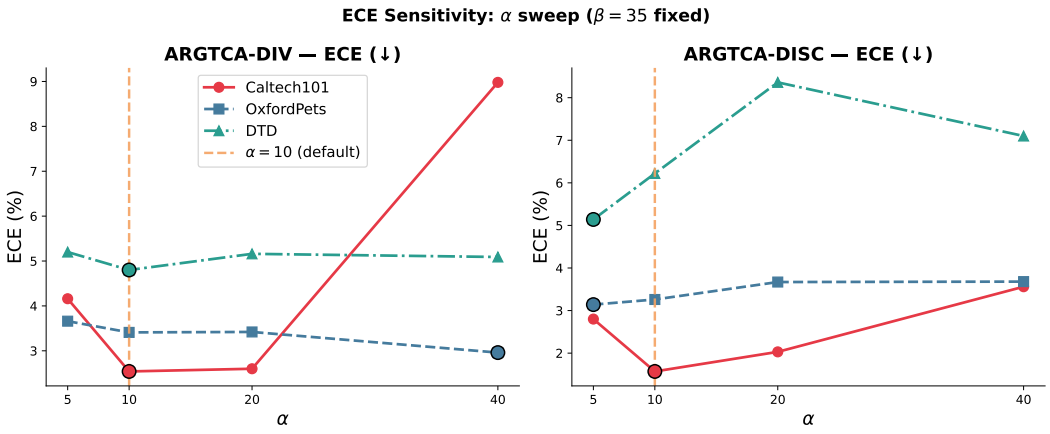


Figure 20: ECE vs. α (inter-class regularization weight) on Caltech101, OxfordPets, and DTD (β=35 fixed).

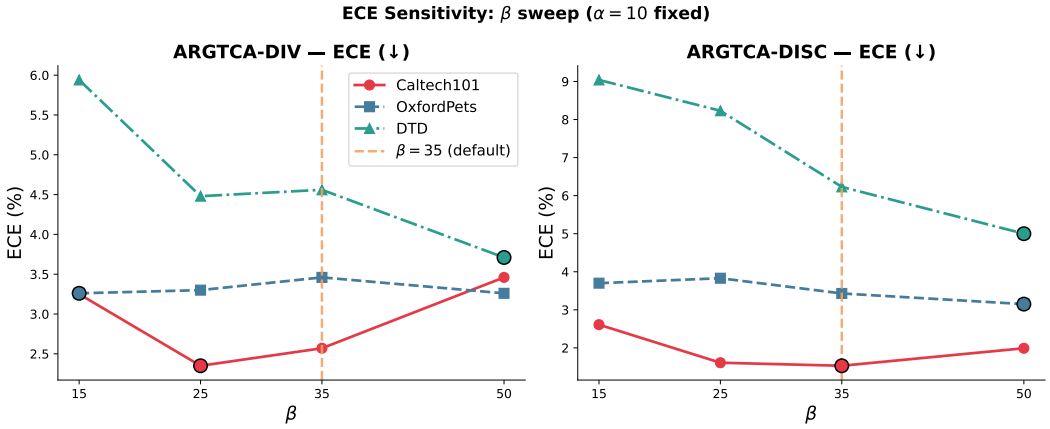


Figure 21: ECE vs. β (intra-class regularization weight) on Caltech101, OxfordPets, and DTD (α=10 fixed).

is a short, human-readable list of M' attribute tokens per class: directly inspectable, editable, and swappable without retraining.

This inspectability also bounds the risks. Gains are non-uniform: EuroSAT's calibration degrades under both variants, so deployment should validate per-dataset against SAG connectivity statistics rather than assume aggregate results transfer. The attribute vocabulary is LLM-generated, inherited *unchanged* from (Hebbalaguppe et al., 2025); our selection criteria are purely geometric (Eqs. 7, 9) and carry no fairness or content term, so ARGTCA reorganizes the LLM's proposals but does not filter them. On person-related or culturally variable categories, a discriminability criterion could surface a class-separating token that encodes a stereotype, and a diversity criterion could surface an unrepresentative one. Auditing selected attribute lists — inexpensive, since they are tokens rather than weights — is therefore a prerequisite for use beyond the object-centric benchmarks studied here.

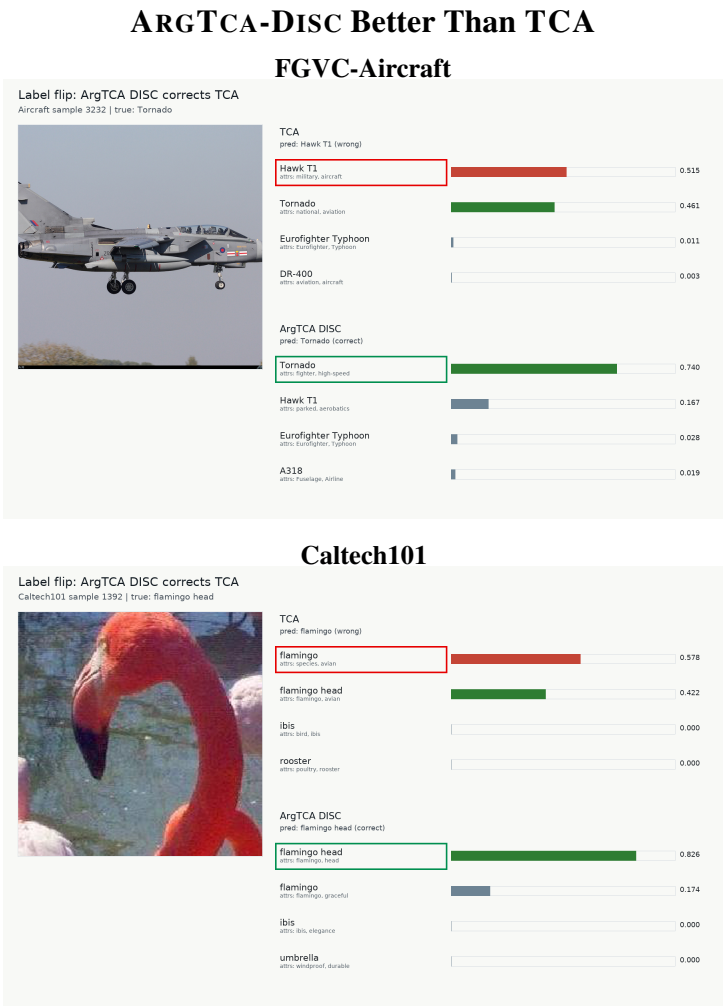
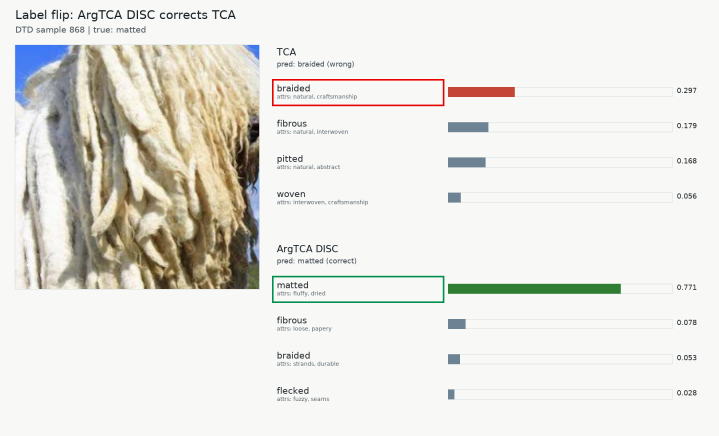


Figure 22: Selected ARGTCA-DISC label-flip examples where ARGTCA-DISC corrects the top prediction made by TCA: FGVC-Aircraft sample 3232 and Caltech101 sample 1392.

ArgTCA-Disc Better Than TCA (Continued)

DTD



ArgTCA-Div Better Than TCA

DTD

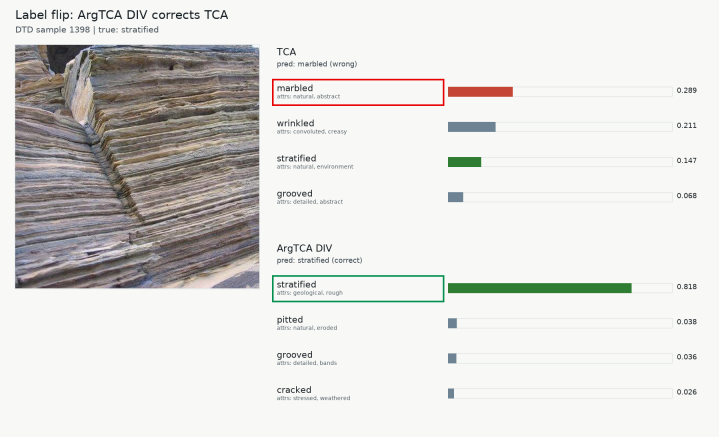


Figure 23: Selected label-flip examples: ArgTCA-Disc corrects TCA on DTD sample 868, and ArgTCA-Div corrects TCA on DTD sample 1398.

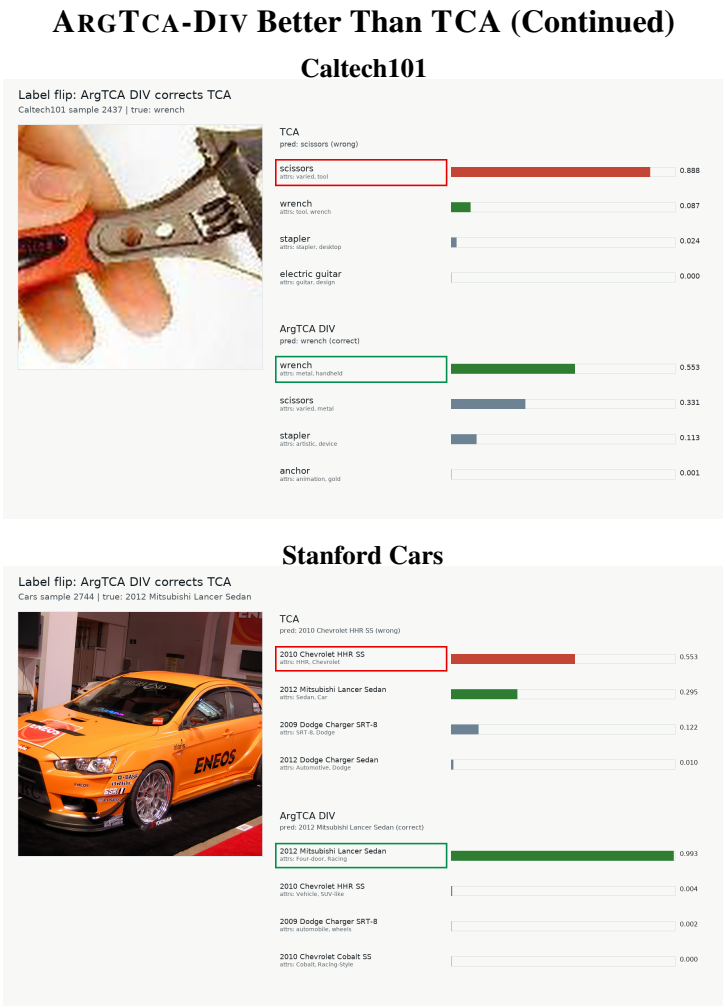


Figure 24: Selected ARGTCA-DIV label-flip examples where ARGTCA-DIV corrects the top prediction made by TCA: Caltech101 sample 2437 and Stanford Cars sample 2744.