Beyond Heuristic Prompting: A Concept-Guided Bayesian Framework for Zero-Shot Image Recognition

Supplementary Material

1. Experimental Settings

Table 1. Statistics of fifteen datasets used in our work.

Dataset	Number of Classes	Test Set Size
SUN397 [**?**]	397	19,850
Aircraft [**?**]	100	3,333
EuroSAT [**?**]	10	8,100
StanfordCars [**?**]	196	8,041
Food101 [**?**]	101	30,300
OxfordPets [**?**]	37	3,669
Flower102 [**?**]	102	2,463
Caltech101 [**?**]	100	2,465
DTD [**?**]	47	1,692
UCF101 [**?**]	101	3,783
ImageNet [**?**]	1,000	50,000
ImageNet-A [**?**]	200	7,500
ImageNet-V2 [**?**]	1,000	10,000
ImageNet-R [**?**]	200	30,000
ImageNet-Sketch [**?**]	1,000	50,889

1.1. Datasets

We primarily assess our framework on eleven classification datasets, including automobiles (Cars [**?**]), textures (DTD [**?**]), human actions (UCF101 [**?**]), aircraft types (Aircraft [**?**]), satellite imagery (EuroSAT [**?**]), pet breeds (Pets [**?**]), flowers (Flower102 [**?**]), food items (Food101 [**?**]), scenes (SUN397 [**?**]), and general objects (Caltech101 [**?**] and ImageNet [**?**]). We additionally evaluate the robustness of our framework to natural distribution shifts on four ImageNet variants, involving ImageNet-A [**?**], ImageNet-V2 [**?**], ImageNet-R [**?**], ImageNet-Sketch [**?**] as out-of-distribution (OOD) data for ImageNet [**?**]. The dataset statistics are shown in Table 1.

1.2. Details of baselines

CLIP and CLIP + E are two versions adopted by **?**] while the former uses a default prompt "A photo of a {class}." and the latter uses the ensemble of 80 hand-crafted prompts. TPT and MTA generate multiple augmented views of test images to enhance prediction robustness. TPT minimizes the entropy of the prediction distributions among all views via prompt tuning. MTA, on the other hand, optimizes an importance score for each view to find the mode of the density of all views and enforces similar importance scores for views with similar prediction distributions. COLA enhances robustness by projecting adversarial image features into a text-induced subspace and refining alignment through optimal transport. CuPL uses five fixed questions (e.g., "What does a {class} look like?") to prompt an LLM and treats the responses as textual prompts. C-TPT [**?**] jointly optimizes prompts during test time to improve calibration by maximizing the Average Text Feature Dispersion, motivated by the observation that prompts inducing higher text-feature dispersion yield lower calibration error. DiffTPT [**?**] leverages a pre-trained diffusion model to synthesize diverse augmented views for each test image and applies a cosine-similarity filtration to remove spurious generations, thereby balancing augmentation diversity and prediction fidelity for test-time prompt tuning. ZERO [**?**] is a training-free test-time adaptation baseline that augments the test image, keeps the most confident views, then sets the Softmax temperature to zero so that marginalization effectively becomes voting over confident predictions, requiring only a single batched forward pass through the vision encoder and no backpropagation.

1.3. Implementation details

For concept synthesis, we use GPT-4.1 Turbo to generate 10 concepts per API call. For each class Y_i, we generate 50 atomic concepts ($M_A = 50$) and randomly sample 500 composite combinations ($|\hat{\mathcal{C}}_i| = 500$) from the atomic set with 3 atomic concepts per combination. The final compositional concept set $\mathcal{C}_i$ is selected using DPP algorithm with a size of either 16 or 50 ($M_i = 16$ or 50) in our main experiments. The prompt used by the LLM for the discriminative concept generation utilized in CGBC is provided in Table 2, while the prompt for the descriptive baseline, as discussed in Sec. **??**, is detailed in Table 3.

In the Adaptive Soft-Trim Likelihood, we set the outlier threshold as $\lambda = 2.5$ and sigmoid slope as the logit scale of CLIP $k = e^{4.6}$. All baselines are evaluated using their default hyperparameters as reported in their respective papers. Unless otherwise specified, we use CLIP with a ViT-B/16 backbone as the primary VLM. To mitigate the effect of randomness, we report the average top-1 accuracy over three random seeds for all main experiments.

1.4. Concept Generation Cost Analysis

In our experiments, we utilize three variants of the GPT-4.1 model, involving GPT-4.1-Turbo, GPT-4.1-Mini, and GPT-4.1-Nano. The token pricing for each model is as follows:

- **GPT-4.1-Turbo**: \$2.00 per 1M input tokens, \$8.00 per 1M output tokens.
- **GPT-4.1-Mini**: \$0.40 per 1M input tokens, \$1.60 per 1M output tokens.
- **GPT-4.1-Nano**: \$0.10 per 1M input tokens, \$0.40 per 1M

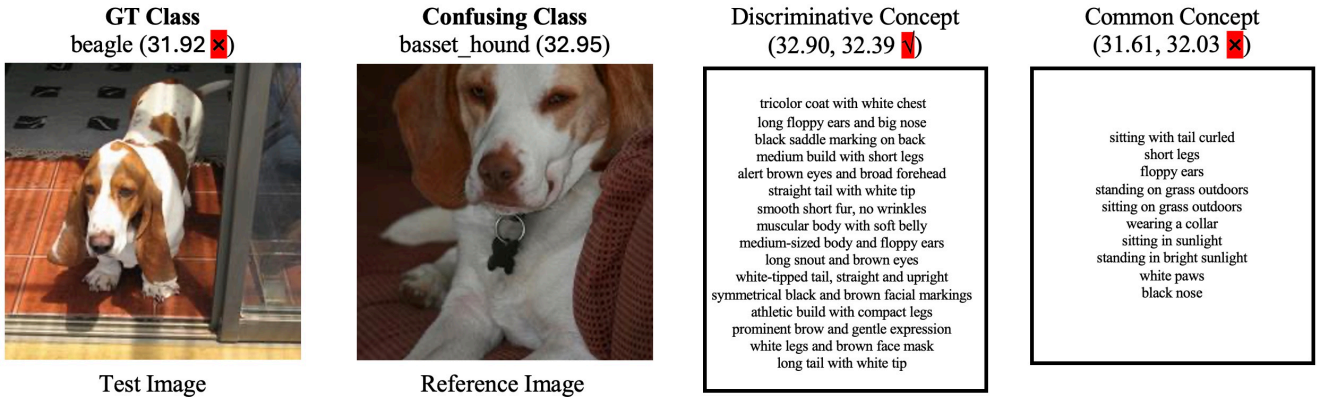

Figure 1. Illustrative examples of Discriminability in constructing the Concept Proposal Distribution. We multiply the raw similarity values by CLIP's scaling factor.

output tokens.

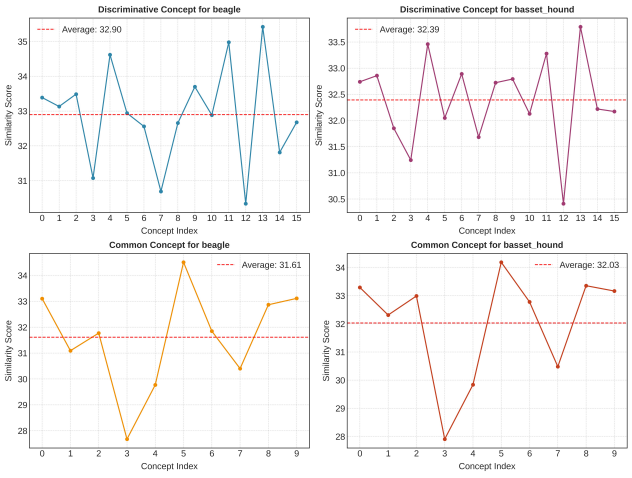

Figure 2. Visualization of similarity between prompts enhanced by discriminative and common concepts for the beagle and basset hound classes. We multiply the raw similarity values by CLIP's scaling factor.

For each API call, we generate 10 candidate concepts. To obtain 50 unique concepts per class, considering possible duplication, we typically require approximately 6 API calls per class. Assuming a total of 1,000 classes, the approximate cost for each model variant is as follows:

- **GPT-4.1-Turbo**: $10.00
- **GPT-4.1-Mini**: $2.00
- **GPT-4.1-Nano**: $0.50

2. Supplementary Experiments

2.1. Robustness Analysis

To evaluate the robustness of our proposed framework under natural distribution shifts, we compare its performance against other zero-shot methods, as presented in Table 4. The results indicate that CGBC consistently outperforms CGBC Prior, highlighting the necessity of likelihood-based refinement for the concept prior distribution. Although both CGBC and CGBC Prior maintain a clear performance margin over CLIP, they fall behind view-based methods under significant domain shifts. This observation suggests that multi-view information is essential in scenarios with substantial distributional changes or noise. To address this, we propose an enhanced variant, CGBC + View. Given an input image X and its N_{aug} augmented views, we apply the CGBC framework to each view to obtain robust similarity scores. We then compute the entropy of the predicted distributions for each view, select the top 10% of views with the lowest entropy (i.e., most confident predictions), following [**?**] and average their results. Empirical results show that CGBC + View outperforms view-augmentation-based baselines by an average of 2 percents across four OOD datasets when the number of prompts is set to 50. Notably, it achieves particularly strong performance on ImageNet-A. These phenomenon further validates the effectiveness of concept-based prompt augmentation in enhancing generalization under distribution shifts.

2.2. Running Time Analysis

All experiments were conducted using four NVIDIA 3090Ti GPUs. To ensure a fair comparison of computational efficiency, we report the inference-time adaptation cost for TPT, MTA, CGBC, and CGBC + View on the ImageNet dataset under identical hardware conditions. Specifically, we exclude model loading time and assume that text prompt embeddings are precomputed. The total adaptation times are approximately 11 hours 26 minutes for TPT, 3 hours 42 minutes for MTA, 2 minutes 43 seconds for CGBC, and 2 hours 33 minutes for CGBC + View. These results demonstrate that our proposed CGBC frame-

Table 2. Prompt for discriminative concept generation. The red text represents variables that are dynamically replaced based on the specific datasets and image classes being tested. "Core class" and "Other Classes" correspond to Y_i and $L_{i,j}$, respectively.

SYSTEM PROMPT:

You are a visual concept proposer tasked with enhancing text descriptions for zero-shot image classification on the test dataset using CLIP.

Given:

- A core class from the test dataset
- The set of other classes in the dataset

Task:

Propose concise, visually discriminative concepts to append to the text description (i.e., "A photo of {core class} with {your concept}") that help CLIP better distinguish the core class from the other classes.

Guidelines:

- Analyze the unique visual characteristics of the core class compared to other classes
- Propose concepts that capture these discriminative visual features.
- Ensure concepts are concrete, easily understandable by CLIP, and specific to the test dataset.
- Each concept should enable CLIP to more accurately classify images of the core class while minimizing confusion with other classes.

IMPORTANT: Your response must follow this exact format:

```
<concepts begin>
concept1
concept2
concept3
</concepts end>
```

Rules:

- Start with `<concepts begin>` and end with `</concepts end>`
- Each concept should be on a new line
- Each concept MUST start with `"The final concept is:  "`
- Ensure concepts are clear, specific, and relevant to the core class
- Avoid generic or ambiguous concepts
- Each concept should be unique and distinct from others
- Keep each concept brief (ideally $\leq$6 words), specific, and easy for CLIP to parse.

USER PROMPT:

Core class: [Core class]. Other classes: [Other Classes]. Please generate 10 unique and visually discriminative concepts. Follow the required format and rules.

work requires significantly less adaptation time compared to view-augmentation-based zero-shot approaches, while still achieving superior performance. Furthermore, even when incorporating multi-view inputs, CGBC + View remains more efficient than standard view-augmented zero-shot methods, owing to its optimization-free nature.

2.3. Qualitative Analysis

In this subsection, we use illustrative examples to demonstrate the roles of discriminability and compositionality in constructing the concept proposal distribution. **Discriminability** Compared to descriptive prompts, our proposed LLM-driven Concept Synthesis Pipeline is more likely to generate discriminative concepts through the use of discriminative prompts. This suggests that while descriptive prompts can also produce effective discriminative prompts, the number and quality are generally lower than ours. The effectiveness of our approach is particularly evident when comparing discriminative concepts with shared concepts that are likely to be generated by descriptive prompts and are common to both classes. As shown in Fig. 1, for the class beagle, when using CLIP with the original prompt format "a photo of class" instantiated with beagle and basset hound, the model misclassifies the image. Even when using a shared concept, a concept frequently generated by descriptive prompts and common to both classes, the reduction in the decision margin between the two classes is insufficient to correct the misclassification. In contrast, when using a discriminative concept, we can effectively adjust the margin and correct the classification. Fig. 2 further illustrates how prompts enhanced with different types of concepts (common vs. discriminative) affect the similarity between the test image and both class prompts. **Compositionality**: Among the atomic concepts generated by the LLM,

Table 3. Prompt for descriptive concept generation. The red text represents variables that will be replaced based on the specific datasets and image classes being tested. "Core class" will be instantiated by Y_i.

SYSTEM PROMPT:
You are a visual concept proposer tasked with enhancing text descriptions for zero-shot image classification on the test dataset using CLIP.
Given:
- A class from the test dataset
Task:
Propose descriptive concepts to append to the text description (i.e., "A photo of {core class} with {your concept}") that help CLIP better understand and recognize the core class.
Guidelines:
- Focus on the visual characteristics and attributes of the core class itself.
- Generate descriptive concepts that capture various aspects, appearances, or contexts of the core class.
- Ensure concepts are concrete, easily understandable by CLIP, and specific to the test dataset.
- Think about different visual perspectives, settings, or attributes that describe the core class.
IMPORTANT: Your response must follow this exact format:
`<concepts begin>`
concept1
concept2
concept3
`</concepts end>`
Rules:
- Start with `<concepts begin>` and end with `</concepts end>`
- Each concept should be on a new line
- Each concept MUST start with `"The final concept is:  "`
- Ensure concepts are clear, specific, and relevant to the given class
- Avoid generic or ambiguous concepts
- Each concept should be unique and distinct from others
- Keep each concept brief (ideally $\leq$6 words), specific, and easy for CLIP to parse.

USER PROMPT:
Core class: [Core class]. Please generate 10 unique and descriptive concepts that capture different visual aspects of this class.

there often exist noisy or outlier concepts as analyzed in the main body. In addition to the likelihood function that helps mitigate the impact of such outliers, as illustrated in Fig. 3, compositional concepts can also alleviate the effect of poor atomic concepts by combining them with more informative ones. However, when noisy concepts constitute a significant portion of the atomic concept set, compositionality alone becomes insufficient. In such cases, a likelihood-based refinement of the concept prior remains essential to effectively suppress the influence of outlier concepts. **Diversity**: For a more comprehensive assessment of diversity, we refer readers to the quantitative analyses reported in the main text and to the final concept set released in our open-source repository.

2.4. Supplementary Ablation studies

Table 5. The rows marked '(avg)' indicate results using concept prompt averaging to replace "or". We use 50 prompts).

Method	SUN.	Air.	Eur.	Car.	Food.	Pets	Flow.	Cal.	DTD	UCF	Ima.	Avg
CGBC Prior	68.5	26.1	59.3	66.6	83.8	88.5	72.4	94.0	55.1	71.8	68.3	68.6
Prior (avg)	68.5	26.1	55.6	66.9	84.2	87.9	71.7	94.2	52.6	70.2	68.9	67.9
CGBC	68.8	26.0	60.3	66.6	83.9	90.7	73.7	94.4	55.7	72.3	69.4	**69.3**
CGBC (avg)	68.7	26.1	56.6	67.0	84.3	88.5	72.3	94.2	52.9	70.6	69.1	68.2

Reasonability of using "or" for concept composition. We compose atomic concepts using natural-language coordination with "or" (e.g., "`A or B`"). Although CLIP does not implement symbolic logic, it is trained on large-scale image–text data and can leverage the distributional semantics of common coordination patterns; thus "or" serves as a cue for *alternatives*, encouraging a text representation compatible with multiple visual realizations. We additionally evaluate a non-compositional baseline that averages the

Table 4. Performance of zero-shot methods on different variants of ImageNet datasets. We report the average performance (Avg.) and average out-of-distribution performance (OOD Avg.). Best and second-best results are **bolded** and underlined, respectively. The *Auxiliary* column shows the number of views and prompts.

Method	A	R	K	V	I	Avg.	OOD Avg.	Auxiliary
CLIP	47.80	73.99	46.15	60.84	66.73	59.10	57.20	(1, 1)
CLIP + E.	49.95	**77.71**	48.26	61.91	68.38	61.24	59.46	(1, 80)
TPT	54.63	77.04	47.97	63.41	68.94	62.40	60.76	(64, 1)
MTA	57.41	76.92	48.58	63.61	69.29	63.16	61.63	(64, 1)
Cupl	47.52	74.14	46.43	59.6	66.75	58.89	56.92	(1, 16)
Cupl	48.12	74.66	46.96	60.43	66.47	59.33	57.54	(1, 50)
CGBC Prior	49.44	75.30	48.06	61.56	68.21	60.51	58.59	(1, 16)
CGBC	49.54	75.35	48.16	62.11	69.22	60.88	58.79	(1, 16)
CGBC + View	61.68	77.65	49.24	64.04	70.5	64.62	63.15	(64, 16)
CGBC Prior	49.91	75.4	48.14	61.61	68.32	60.68	58.77	(1, 50)
CGBC	49.79	75.49	48.36	62.15	69.43	61.04	58.95	(1, 50)
CGBC + View	**62.19**	77.41	**49.45**	**64.37**	**70.64**	**64.81**	**63.36**	(64, 50)

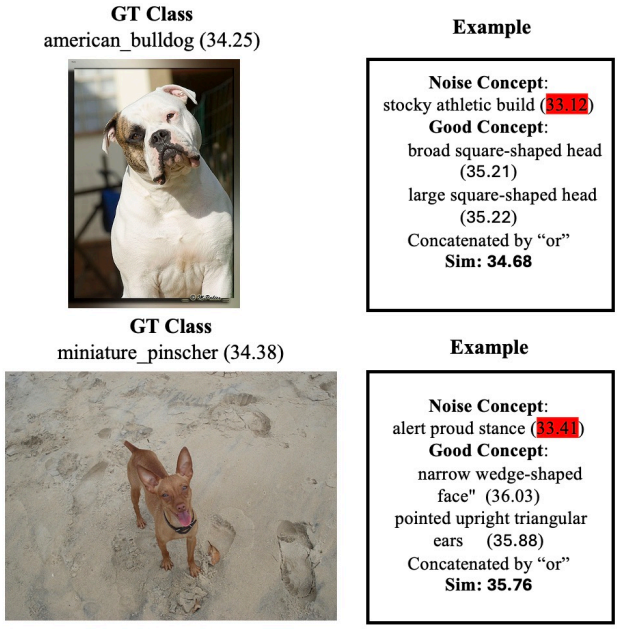

Figure 3. Illustrative examples of Compositionality in constructing the Concept Proposal Distribution. We multiply the raw similarity values by CLIP's scaling factor.

CLIP text embeddings of the atomic prompts within a concept (Table 5, *Prior (avg)*). Empirically, "or" outperforms prompt averaging in our setting. We attribute this to the fact that textual composition preserves syntactic structure and a learned coordination pattern, whereas embedding averaging is a purely mixture that can dilute distinctive semantics.

Versatility across concept generators. Beyond the GPT-family concept generators validated in the main text, we further evaluate CGBC using the Gemini series as the concept generator. Table 6 shows that CGBC maintains consistent performance across model groups: replacing GPT-4.1 turbo with Gemini 2.5 Flash, Gemini 2.5 Flash Lite, or Gemini 2.0 Flash results in only marginal changes in accuracy on all datasets, with the overall average remaining within a narrow band (61.0 vs. 59.9–60.4). These results indicate that CGBC is not tied to a specific LLM family and is robust to the choice of concept generator, supporting its practicality for deployment under different cost/latency constraints.

Table 6. Performance on various model group.

Avg	$\lvert L_i \rvert$	EuroSAT	DTD	Aircraft	Pets	UCF101	Average	Avg/EuroSAT
GPT-4.1 turbo	10	60.3	55.7	26.0	90.7	72.3	61.0	61.2
Gemini 2.5 Flash	10	59.5	54.3	26.0	90.6	71.3	60.4	60.6
Gemini 2.5 Flash Lite	10	58.9	53.2	26.4	89.8	71.4	59.9	60.1
Gemini 2.0 Flash	10	58.8	53.0	26.1	90.6	71.5	60.0	60.2

3. Theoretical Analysis of Adaptive Soft-Trim based likelihood

3.1. Proof of Robust Guarantee

We restate Theorem **??** below for convenience.

Theorem (Robust Guarantee)**.** *If the size of concept set $\mathcal{C}_i$ for each class Y_i satisfies $M_i \geq \frac{C_0 \log(1/\delta)}{\rho_i^2}$ for a universal constant C_0, then with probability at least $1-\delta$, the estimation error of $\hat{\mu}_i$ is bounded by:*

$$|\hat{\mu}_i - \mu_i| \leq C_1 \sigma_i \rho_i + C_2 \sigma_i \sqrt{\frac{\log(1/\delta)}{M_i}} + \frac{C_3 \sigma_i}{k} \qquad (1)$$

where C_1, C_2, and C_3 are universal constants and k is sigmoid slope parameter in Eq. (**??**).

Proof. The proof proceeds by decomposing the total estimation error into three components: (1) a bias term arising from the adversarial contamination, (2) a statistical error term due to finite sampling, and (3) an approximation error term from the soft-trimming mechanism.

Let the set of samples be $\mathcal{S}_i = \{S_{i,1}, \ldots, S_{i,M_i}\}$. By the Huber contamination model, we can conceptually partition $\mathcal{S}_i$ into a set of inliers $\mathcal{G}_i$ drawn from the 1-sub-Gaussian distribution $\mathcal{N}(\mu_i, \sigma_i^2)$ and a set of outliers $\mathcal{O}_i$ from an arbitrary adversarial distribution. We have $|\mathcal{G}_i| \geq (1-\rho_i)M_i$ and $|\mathcal{O}_i| \leq \rho_i M_i$.

The Adaptive Soft-Trim estimator is defined as $\hat{\mu}_i = \frac{\sum_{j=1}^{M_i} w_{i,j} S_{i,j}}{\sum_{j=1}^{M_i} w_{i,j}}$, where the weights $w_{i,j}$ are a decreasing function of the normalized deviation from the sample median, i.e., $w_{i,j} = f(-|S_{i,j} - m_i|/\text{MAD}_i)$ for some decreasing function f (e.g., a sigmoid we utlize in our implementation).

Robustness of Median and MAD A foundational result in robust statistics is that the median and Median Absolute Deviation (MAD) are robust estimators of location and scale. Under the given sample complexity condition $M_i \geq \frac{C_0 \log(1/\delta)}{\rho_i^2}$, it can be shown that with probability at least $1-\delta$ [**?**]:

$$|m_i - \mu_i| \leq O(\sigma_i \rho_i) \tag{2}$$

$$|\text{MAD}_i - c\sigma_i| \leq O(\sigma_i \rho_i) \tag{3}$$

where $m_i = \text{median}(\mathcal{S}_i)$ and c is the constant factor relating MAD to the standard deviation for the underlying clean distribution (e.g., $c \approx 0.6745$ for a Gaussian). This ensures that the center and scale used for trimming are themselves close to the true parameters, up to a factor of the contamination rate.

Error Decomposition We analyze the error $|\hat{\mu}_i - \mu_i|$ by triangle inequality. Let $W = \sum_{j=1}^{M_i} w_{i,j}$.

$$\begin{aligned} |\hat{\mu}_i - \mu_i| &= \frac{1}{W}\left|\sum_{j=1}^{M_i} w_{i,j}(S_{i,j} - \mu_i)\right| \\ &\leq \frac{1}{W}\left|\sum_{j\in\mathcal{G}_i} w_{i,j}(S_{i,j} - \mu_i)\right| + \frac{1}{W}\left|\sum_{j\in\mathcal{O}_i} w_{i,j}(S_{i,j} - \mu_i)\right| \end{aligned} \tag{4}$$

Bias from Contamination: The second term, $\frac{1}{W}|\sum_{j\in\mathcal{O}_i} w_{i,j}(S_{i,j} - \mu_i)|$, captures the influence of outliers. Our proposed adaptive soft-trim based likelihood is designed to down-weight points with large deviations $|S_{i,j} - m_i|$. While outliers can be arbitrary, their influence is controlled. The theory of robust mean estimation [**?** **?**] establishes that for any estimator in this class, the bias induced by a ρ_i-fraction of contamination is at least $\Omega(\sigma_i\rho_i)$. Our proposed likelihood, being a variant of a trimmed mean, achieves a near-optimal rate. The bias from both the residual influence of down-weighted outliers and the error in the trimming center ($|m_i - \mu_i|$) is bounded by $C_1\sigma_i\rho_i$.

Statistical Error: The first term, $\frac{1}{W}|\sum_{j\in\mathcal{G}_i} w_{i,j}(S_{i,j} - \mu_i)|$, represents the statistical error from estimating a mean from a finite number of the weighted inlier concepts. It shrinks as the sample size M increases. For $j \in \mathcal{G}_i$, $S_{i,j} - \mu_i$ are i.i.d. 1-sub-Gaussian random variables with mean 0. The sum is a weighted sum of such variables. By a weighted version of Hoeffding's inequality, for any fixed weights, this sum is bounded by $O(\sigma_i\sqrt{\sum w_{i,j}^2 \log(1/\delta)})$. Since $w_{i,j} \in [0,1]$ and $W \approx (1-\rho_i)M_i$, we have $\sum w_{i,j}^2 \leq W$. The normalized sum is thus bounded by $O(\sigma_i\sqrt{\log(1/\delta)/W}) \approx O(\sigma_i\sqrt{\log(1/\delta)/M_i})$. This gives the term $C_2\sigma_i\sqrt{\frac{\log(1/\delta)}{M_i}}$.

Approximation Error The use of a "soft" sigmoid function with slope k instead of a "hard" threshold introduces a third error term. The weight function $w_{i,j}$ transitions from 1 to 0 over a region whose width is proportional to $1/k$ in the normalized deviation space. Inliers that fall into this transition region are unduly down-weighted, and outliers that fall into it are not fully removed. This imperfect trimming introduces a bias. The magnitude of this error is proportional to the width of the transition region, leading to an error term of the form $C_3\sigma_i/k$. As $k \to \infty$, the soft-trim approaches a hard-trim, and this term vanishes.

Combining the three error bounds via the triangle inequality, we obtain the final result. With probability at least $1-\delta$, the total error is bounded by the sum of the bias, statistical, and approximation errors:

$$|\hat{\mu}_i - \mu_i| \leq C_1\sigma_i\rho_i + C_2\sigma_i\sqrt{\frac{\log(1/\delta)}{M_i}} + \frac{C_3\sigma_i}{k} \tag{5}$$

This completes the proof. □

3.2. Proof of Multi-class excess risk

We restate Corollary **??** below for convenience.

Corollary (Multi-class excess risk)**.** *Under the condition in Theorem* **??***, the classifier for image X in Eq.* (**??**) *satisfies:*

$$\mathcal{R} - \mathcal{R}^* \leq \Pr\left[\textit{margin}_\mu(X) \leq 2 \cdot \max_i |\hat{\mu}_i(X) - \mu_i(X)|\right] \tag{6}$$

where $\mathcal{R}$ is the risk of our classifier, $\mathcal{R}^*$ is the optimal Bayes risk, and $\textit{margin}_\mu(X)$ is the margin of the true mean $\mu_i(X)$.

Consequently, With probability at least $1-\delta$, the excess risk is bounded by:

$$\mathcal{R} - \mathcal{R}^* \leq \Pr\Big[\textit{margin}_\mu(X) \leq 2\big(C_1\sigma_{\max}\rho_{\max} + C_2\sigma_{\max}\sqrt{\tfrac{\log(K/\delta)}{M_{\min}}} + \tfrac{C_3\sigma_{\max}}{k}\big)\Big] \quad (7)$$

where K is the number of classes, $\sigma_{\max} = \max_i \sigma_i$, $\rho_{\max} = \max_i \rho_i$, and $M_{\min} = \min_i M_i$.

Proof. The proof is in two parts. First, we relate the excess classification risk to the estimation error of the means $\mu_i(X)$. Second, we apply the bound from Theorem **??** to this relation.

Relating Excess Risk to Estimation Error Let $h^*(X) = \arg\max_i \mu_i(X)$ be the Bayes optimal classifier for the true (uncontaminated) means. The excess risk of our classifier $\hat{h}(X) = \arg\max_i \hat{\mu}_i(X)$ is upper-bounded by the probability of misclassification relative to the optimal one:

$$\mathcal{R}(\hat{h}) - \mathcal{R}^* \leq \mathbb{E}_X\left[\mathbb{I}(\hat{h}(X) \neq h^*(X))\right] = \Pr(\hat{h}(X) \neq h^*(X)) \quad (8)$$

A misclassification occurs if $\hat{h}(X) \neq h^*(X)$. Let $i^* = h^*(X)$ be the true optimal class. A misclassification implies that there exists some class $j \neq i^*$ such that $\hat{\mu}_j(X) > \hat{\mu}_{i^*}(X)$. We can establish a sufficient condition for this event. Consider the inequality:

$$\begin{aligned} \hat{\mu}_j(X) &> \hat{\mu}_{i^*}(X) \\ \mu_j(X) + (\hat{\mu}_j(X) - \mu_j(X)) &> \mu_{i^*}(X) + (\hat{\mu}_{i^*}(X) - \mu_{i^*}(X)) \\ \mu_{i^*}(X) - \mu_j(X) &< (\hat{\mu}_j(X) - \mu_j(X)) \\ &\quad - (\hat{\mu}_{i^*}(X) - \mu_{i^*}(X)) \end{aligned}$$

By the triangle inequality, the right-hand side is bounded:

$$\begin{aligned} \mu_{i^*}(X) - \mu_j(X) &\leq |\hat{\mu}_j(X) - \mu_j(X)| + |\hat{\mu}_{i^*}(X) - \mu_{i^*}(X)| \\ &\leq 2 \cdot \max_k |\hat{\mu}_k(X) - \mu_k(X)| \end{aligned}$$

The margin of the true means is defined as $\text{margin}_\mu(X) = \mu_{i^*}(X) - \max_{j \neq i^*} \mu_j(X)$. Since a misclassification implies the existence of a $j \neq i^*$ satisfying the above, it must be that the margin itself is bounded by this quantity. Therefore, the event $\{\hat{h}(X) \neq h^*(X)\}$ is a subset of the event $\{\text{margin}_\mu(X) \leq 2 \cdot \max_k |\hat{\mu}_k(X) - \mu_k(X)|\}$. This gives the first inequality of the corollary:

$$\mathcal{R}(\hat{h}) - \mathcal{R}^* \leq \Pr\left[\text{margin}_\mu(X) \leq 2 \cdot \max_i |\hat{\mu}_i(X) - \mu_i(X)|\right]$$

Applying the High-Probability Bound Theorem **??** provides a bound on $|\hat{\mu}_i(X) - \mu_i(X)|$ for a single class i that holds with probability $1-\delta$. We require a bound that holds simultaneously for all K classes. Let $\delta' = \delta/K$. Applying Theorem **??** for each class $i \in \{1, \ldots, K\}$ with failure probability δ', we have that with probability at least $1-\delta'$,

$$|\hat{\mu}_i - \mu_i| \leq C_1\sigma_i\rho_i + C_2\sigma_i\sqrt{\frac{\log(1/\delta')}{M_i}} + \frac{C_3\sigma_i}{k} \quad (9)$$

By the union bound, the probability that at least one of these K bounds fails is at most $\sum_{i=1}^K \delta' = K(\delta/K) = \delta$. Therefore, with probability at least $1-\delta$, the bound holds for all $i = 1, \ldots, K$ simultaneously.

On this high-probability event, we can bound the maximum estimation error:

$$\begin{aligned} \max_i |\hat{\mu}_i(X) - \mu_i(X)| &\leq \max_i (C_1\sigma_i\rho_i + C_2\sigma_i\sqrt{\frac{\log(K/\delta)}{M_i}} + \frac{C_3\sigma_i}{k}) \\ &\leq C_1\sigma_{\max}\rho_{\max} + C_2\sigma_{\max}\sqrt{\frac{\log(K/\delta)}{M_{\min}}} + \frac{C_3\sigma_{\max}}{k} \end{aligned}$$

where $\sigma_{\max} = \max_i \sigma_i$, $\rho_{\max} = \max_i \rho_i$, and $M_{\min} = \min_i M_i$. Substituting this high-probability bound into the result from **Relating Excess Risk to Estimation Error**, we arrive at the final expression for the excess risk:

$$\mathcal{R}(\hat{h}) - \mathcal{R}^* \leq \Pr\Big[\text{margin}_\mu(X) \leq 2\big(C_1\sigma_{\max}\rho_{\max} + C_2\sigma_{\max}\sqrt{\tfrac{\log(K/\delta)}{M_{\min}}} + \tfrac{C_3\sigma_{\max}}{k}\big)\Big] \quad (10)$$

This bound connects the classifier's performance to the intrinsic difficulty of the problem, captured by the distribution of $\text{margin}_\mu(X)$, and the quality of our robust estimation procedure. □

4. Limitations

CGBC yields the largest gains on datasets where CLIP can reliably interpret LLM-generated concepts (e.g., EuroSAT, DTD), whereas on harder domains (e.g., Aircraft) improvements are smaller and some prior prompts can match or slightly outperform CGBC. We attribute this variation to two factors: (i) occasional outlier concepts produced by LLMs and (ii) dataset-dependent limitations of CLIP in mapping text concepts to visual evidence. When CLIP's concept understanding is adequate, filtering and reweighting reduce the impact of outliers and CGBC improves over *CGBC Prior* (e.g., EuroSAT: 59.3→60.3; ImageNet: 68.3→69.4). When CLIP's zero-shot baseline is weak, CLIP's representational limitations can dominate and CGBC provides limited benefit (e.g., Aircraft: 26.1→26.0), where more descriptive prompts (e.g., CUPL) may be easier for CLIP to align with images.

From a theoretical standpoint, our method is motivated by a Bayesian view of zero-shot classification but relies on practical approximations. A fully Bayesian treatment would require marginalizing over the space of all concepts C to compute $p(Y \mid X)$, which is computationally infeasible; we instead approximate this integral using a small set

of high-quality concepts synthesized by an LLM pipeline. Moreover, the likelihood term $p(X \mid C)$ is generative in nature, while CLIP is a discriminative model; consequently, we use CLIP image–text similarity as an empirical proxy for $p(X \mid C)$. Developing a more principled probabilistic grounding of this likelihood approximation remains an open direction.