EFFECTS OF PRETRAINING CONCEPT FREQUENCY AND SEMANTIC SIMILARITY AMONG CLASSES ON CLIP GENERALIZED ZERO-SHOT ACCURACY

Anonymous authors
Paper under double-blind review

ABSTRACT

Foundation vision-language models (VLMs), such as CLIP, exhibit impressive generalized zero-shot recognition capabilities. However, their performance is shown to be affected by imbalances in their pretraining data. Understanding the factors that affect downstream zero-shot recognition accuracy under such pretraining imbalance is essential. Recent work has shown that across many downstream datasets, the zero-shot accuracy of individual concepts follows a log-linear trend with concept frequency in the pretraining dataset, and that achieving linear improvements in accuracy requires exponentially more training examples. In this paper, we (i) demonstrate that semantic similarity among target classes is another strong predictor of zero-shot accuracy, (ii) analyze how this semantic effect interacts with concept-frequency trends across different dataset sizes, model capacities and pre-training scales, showing that the frequency-accuracy link diminishes for larger pretraining corpora while the semantic-similarity signal does not collapse, and (iii) propose a graph-based zero-shot accuracy-estimation framework for a given list of class names that jointly takes (a) concept frequency estimates and (b) a pair-wise semantic-similarity matrix as input; allowing for a more flexible and realistic setup than the fixed 1000-way classification. Overall, our results provide a more comprehensive understanding of how both semantic complexity and class cardinality influence generalized zero-shot accuracy of vision-language models.

1 INTRODUCTION

Pretraining on large-scale image-language datasets has shown promising generalized zero-shot capabilities (Radford et al., 2021; Zhai et al., 2023) and these models have been widely adopted in both industry and academia (e.g. in open-vocabulary recognition tasks (Gu et al., 2022; Zhong et al., 2022), in anomaly detection (Jeong et al., 2023; Zhou et al., 2024), in concept bottleneck models (Yang et al., 2023; Oikarinen et al., 2023; Liu et al., 2025)). Therefore, it is important to understand the factors influencing the downstream zero-shot performance of these models.

Foundation Vision-Language Model (VLM) pretraining data inherently suffers from long-tailed concept distributions (Radford et al., 2021; Xu et al., 2024). Despite curation efforts to cap the number of examples per concept, rare concepts remain scarce, ultimately degrading the zero-shot performance of VLMs (Parashar et al., 2024; Udandarao et al., 2024; Wen et al., 2024).

Similar to how Long-tail Recognition (LTR) literature emphasizes class cardinality in supervised learning (Zhang et al., 2023), recent research links pretraining concept frequency directly to VLM zero-shot accuracy. Since web-scale data is not readily organized under the categories of the target zero-shot task, calculating these frequencies requires specialized quantification methods to analyze their impact on performance (Parashar et al., 2024; Udandarao et al., 2024; Wen et al., 2024).

These studies consistently reveal a positive correlation between estimated concept frequency and downstream accuracy (Parashar et al., 2024). Specifically, Udandarao et al. (2024) report a log-linear relationship, while also noting that achieving linear accuracy gains requires exponentially more training data. However, Wen et al. (2024) observe that this frequency-accuracy correlation weakens as the overall scale of the pretraining dataset increases.

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.

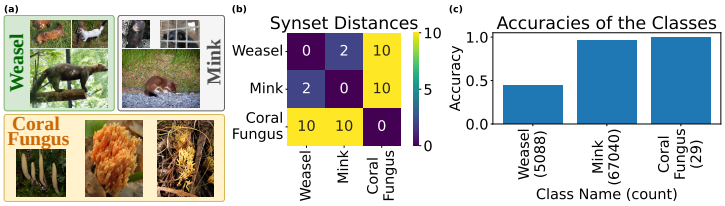

Figure 1: **Pre-training concept frequency does not always explain zero-shot accuracy.** We select three classes: Weasel, Mink, and Coral Fungus. Weasel and Mink have high semantic similarity (b). Frequencies calculated by Wen et al. (2024) for these classes are given in (c). Although Coral Fungus has just a few images (only 29), it has high zero-shot accuracy; and, although Weasel frequency is in the thousands, it has low zero-shot accuracy (c).

Motivation and Preliminary Analysis. Although class frequency has been used to predict the class accuracies, in this work, we argue and demonstrate that semantic similarity among the classes of the downstream dataset is also a major determinant of zero-shot performance. To illustrate this, we analyze three ImageNet (Russakovsky et al., 2015) classes (Fig. 1a): two that are semantically similar (Weasel and Mink), and one that is semantically distant (Coral Fungus). We specifically chose Coral Fungus because it is also one of the least frequent pretraining classes according to various quantification methods (Parashar et al., 2024; Udandarao et al., 2024; Wen et al., 2024). We calculate "semantic similarity" using the shortest path length on the WordNet hierarchy (Miller, 1995; Fellbaum, 1998)[1]. Fig. 1b visualizes these synset distances between the selected classes.

Despite its rarity in pretraining data (e.g. estimated at just 29 image-text pairs by Wen et al. (2024)), Coral Fungus achieves 100% zero-shot accuracy because it is highly distinct from the other classes (Fig. 1c). In contrast, Weasel and Mink struggle to achieve perfect accuracy despite being significantly more frequent (estimated at 5,088 and 67,040 pairs, respectively (Wen et al., 2024)). This highlights how semantic similarity heavily influences downstream zero-shot performance.

When evaluating downstream datasets like ImageNet, it is crucial to recognize that the degree of semantic competition varies significantly per class. In our 3-class example, Coral Fungus faces minimal semantic competition, whereas Weasel and Mink strongly compete with one another. We refer to this uneven distribution of similar competitors as a dataset lacking *semantic balance*.

Contributions. **(i)** We show that the log-linear trend between frequency and accuracy reported by prior work, which is obtained by *binning*: grouping classes into frequency intervals and plotting the average or median accuracy within each bin, conceals substantial scatter: the underlying per-class, unbinned data is far noisier and only weakly correlated. The correlation is substantially amplified both by binning (averaging accuracy within frequency intervals) and by aggregating datasets of different granularity. **(ii)** We show that semantic similarity among classes is an important factor for downstream zero-shot performance, alongside concept frequencies, and that combining the two explains accuracy variance substantially better than either alone. **(iii)** We find that semantic similarity remains a significant, robust predictor as class-subset size shrinks, pretraining data scales up, or model size increases. Under these conditions, the frequency-accuracy relationship weakens or breaks down entirely. **(iv)** We introduce a modified Graphormer architecture (ZAGT) that formulates zero-shot accuracy prediction as a graph-regression problem, treating target classes as nodes, their frequencies and embeddings as node attributes, and their pairwise semantic similarities as edge weights. ZAGT predicts zero-shot accuracy for a class list of arbitrary length and composition, a step toward tools that let practitioners forecast model performance before collecting data.

2 Related Work

Quantification of Concept Frequencies has been studied extensively in the literature. It is used both for balancing pretraining data during curation (Radford et al., 2021; Xu et al., 2024) and for estimating per-class frequencies for specific downstream tasks such as ImageNet1K classification (Russakovsky et al., 2015). Because associating image-text pairs with zero-shot classes is complicated by synonyms, repeated class-name mentions, and image-text inconsistencies, prior work assigns occurrences to concepts in different ways.

[1]Since all ImageNet classes map directly to WordNet synsets, we can easily estimate semantic distance by finding the shortest path between their corresponding nodes on the WordNet graph.

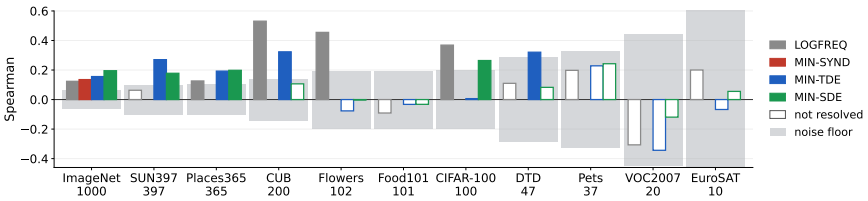

Figure 2: Spearman's rank correlation coefficients between zero-shot accuracies and different factors over several downstream datasets. Gray bands show the thresholds obtained from a permutation test, taking the 95th percentile of the resulting null distribution of absolute correlations. Bars that pass this threshold are filled with their respective colors. Class counts are given under dataset names.

Parashar et al. (2024) take a text-only approach, generating and filtering class synonyms with an LLM, then using it again to verify each match and quantify how often a concept occurs in the pretraining data. Similarly, Wen et al. (2024) use the synonyms in the WordNet (Miller, 1995; Fellbaum, 1998) to query the texts, and for MetaCLIP (Xu et al., 2024) they use the already available 500k concept frequencies employed during the data curation. Udandarao et al. (2024) extract concept names directly from the images with an image-tagging model (Zhang et al., 2024; Huang et al., 2025) and match them against text.

While these quantification methods estimate the class frequencies in different ways, their overall trends appear similar, especially among those using a text-only approach as evidenced by the high correlation between them (Appendix A Table 3).

From Concept-Frequencies to Accuracy Estimation. Several studies have demonstrated that VLM performance can be predicted from the frequency of concepts in the pretraining data (Udandarao et al., 2024; Parashar et al., 2024; Wen et al., 2024). Such a predictive relationship of Wen et al. (2024) is displayed in Fig. 2 in the form of the correlation between estimated log concept frequencies and class accuracies on several downstream datasets. Although for datasets with larger number of classes show significant positive Spearman correlations, as the number of classes decreases, the correlations become non-significant according to permutation test. We detail how we conducted this statistical test in Appendix A. The trends between different frequency quantification methods are mostly similar and we provide their results in Appendix throughout the text when they do not enrich the discussion.

These studies forecast subpar zero-shot performances on rare concepts (Parashar et al., 2024; Udandarao et al., 2024). To anticipate these failures, recent methods introduce probing techniques to estimate how underrepresented specific domains are in the pretraining data (Vorster et al., 2024). Recognizing this bias, newer adaptive techniques attempt to correct it at test time. These methods mitigate imbalances by adapting prompt distributions (Shu et al., 2022), correcting pretraining label bias (Zhu et al., 2024), or replacing rare concepts via retrieval (Parashar et al., 2024).

Beyond single-concept recognition, pretraining data statistics are critical for estimating compositional zero-shot generalization. Models exhibit pronounced brittleness when tasked with recognizing rare concept pairs, even if the individual constituent concepts are highly frequent in the pretraining data (Qu & Xie, 2026). Consequently, the co-occurrence frequency of these compositional elements during pretraining has emerged as a strong predictor of how well models like CLIP will generalize to real-world tasks (Wiedemer et al., 2024).

VLM Zero-Shot Performance Prediction. Zohar et al. (2023), in their work LOVM, introduce the novel task of predicting and ranking the performance of VLMs for a specific downstream task. They calculate different factors that may affect the performance by using LLMs to generate class captions and using them as if they are images as texts and images share the same embedding space in CLIP. Interestingly, the factors we use share many similarities which is discussed in Appendix B. Although their task is very similar to ours, they predict/rank the overall accuracy for a dataset while we are interested in class-level accuracies to uncover the factors that affect the accuracy on concept-level which, we hope, gives researchers further insight about the inner dynamics of VLMs. Instead of using texts as image proxies, Robbins et al. (2025) generate class images for performance prediction to see if a model would work on a specific class, i.e., at the same class-level granularity as ours. In contrast, we analyze how semantic and frequency-based factors, and their interaction, determine class-level zero-shot performance.

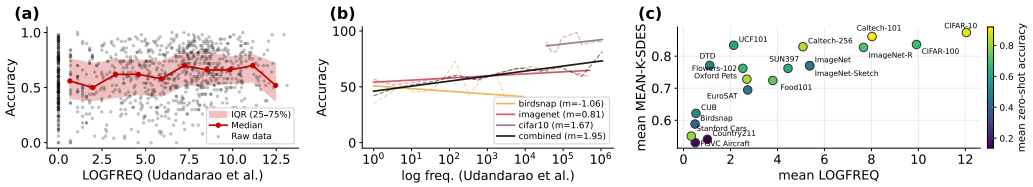

Figure 3: **Binning and pooling inflate the trend.** Frequencies from Udandarao et al. (2024). (a) ImageNet1K per-class scatter with binned medians and interquartile range. (b) Birdsnap, ImageNet and CIFAR10 separately and pooled, lines are fitted to the bin means, slopes m in the legend. (c) Dataset-level mean LOGFREQ against mean MEAN-K-SDE across Udandarao's 19 datasets. ViT-B/32, LAION-400M.

3 Revisiting the Trend Between Frequency & Accuracy

Although Udandarao et al. (2024) and Parashar et al. (2024) depict a clear trend between log-frequency and accuracy, the plotted values exhibit a high degree of variability (Fig. 3a). This is expected, as class frequency is not the only factor determining the accuracy. The smooth trend in those works is obtained by binning the data (as in Fig. 3a). However, a clear trend in binned plots does not imply a strong correlation in the original data. For example, the highest Spearman's rank correlation coefficient among those works on ImageNet is 0.19 (Appendix A Table 4), which could be considered weak.

Average accuracy varies across downstream datasets. For example, on CIFAR10 (Krizhevsky et al., 2009), the model achieves higher accuracy than ImageNet1K (Russakovsky et al., 2015) or Birdsnap (Berg et al., 2014) (Fig. 3b). Many factors play a role in this, from image resolution to decisions made on the data acquisition methods. Here, we focus on semantic similarity between classes. Providing that the other conditions are the same, it would be easier to differentiate a dog and a cat than to distinguish fine-grained categories such as similar dog breeds. Therefore, fine-grained datasets tend to obtain lower accuracies. In Fig. 3b, we select three of the datasets used by Udandarao et al. (2024) and draw their log-frequency versus accuracy plots separately. CIFAR10 obtains higher accuracies around 85-90% while ImageNet is around 55-70% and Birdsnap is around 30-50%. This is expected as Birdsnap is a fine-grained visual recognition dataset on birds.

Stemming from this semantic granularity, another plotting decision that may exaggerate the apparent trend is aggregating multiple datasets in a single plot. To illustrate the effect we again turn our focus on Fig. 3b. A quick glance shows that the zero-shot model exhibits a negative slope on Birdsnap drawn over this specific bins but positive slopes on ImageNet and CIFAR10. When we combine all three datasets by pooling log-frequency, accuracy pairs and applying the same binning procedure, we obtain a steeper trend. This effect arises from the entanglement of class selection during the dataset creation and concept frequency. When a broad category, like bird, is subdivided into finer classes, each subclass typically has a lower concept frequency than the parent class. Consequently, fine-grained datasets such as Birdsnap appear at lower concept frequencies in Fig. 3b. This is important because we also expect fine-grained datasets to obtain lower accuracies due to the stronger semantic opposition their classes encounter. This is illustrated in Fig. 3c where we plot dataset-level-averaged log-frequency vs a semantic factor (we discuss it in Sec. 4) and see a Spearman correlation of 0.81. Consequently, both factors correlate with dataset-level accuracy by itself (0.59 semantic, 0.55 frequency); however, since they are themselves strongly correlated, their individual contributions cannot be told apart. In the end, combining multiple datasets of varying granularity may reveal a more pronounced trend than any individual dataset, as illustrated in Fig. 3b. Note that this alone does not imply any causation, and does not mean semantic information explains away the effect of frequency or vise versa. This just highlights the importance of incorporating semantic similarity into the equation while analyzing the relation between concept frequencies and accuracies.

Even if the datasets are inspected individually, Parashar et al. (2024) demonstrate that most downstream tasks exhibit the log-frequency and accuracy trend. However, as with dataset aggregation, a single dataset is not generally semantically balanced, meaning that the semantic opposition against a class is not the same for all classes throughout the whole dataset. This raises the question whether the trend between concept frequency and accuracy could be fully explained away by semantic similarity. In Sec. 5, we provide findings supporting that this is not the case: jointly using semantic similarity and concept frequency yields a better understanding of the zero-shot accuracy.

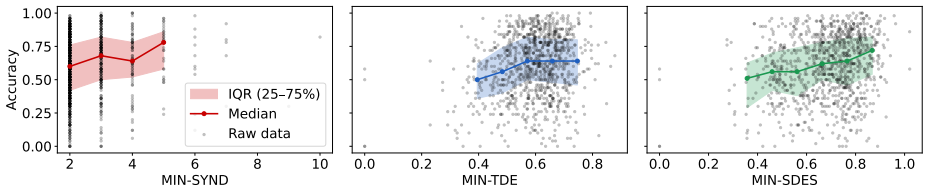

Figure 4: Relation between summarized semantic similarity quantifications and zero-shot class accuracies on ImageNet1K. For MIN-TDE, MIN-SDE, the results are obtained by linearly binning the values and for MIN-SYND, the bins are determined directly by its discrete values. The line shows the medians of bins. If example count in a bin is ≤ 15, it is not shown in the median line.

4 Quantification of Semantic Similarity

4.1 Methods of Quantification

We employ three different ways to quantify semantic similarity between a pair of classes. The first one is **Synset Distance** (**SYND**), which is the length of shortest path on the WordNet graph (Miller, 1995; Fellbaum, 1998) between two synsets that correspond to the classes. One limitation is that not every class in every downstream task has a corresponding synset in WordNet. Therefore, synset distance is mainly limited for analyses on ImageNet (Russakovsky et al., 2015). A further drawback stems from the fact that the noun hierarchy of WordNet contains is-a relations, which can yield large distance values even for classes that frequently co-occur. For example, ImageNet classes "bakery" and "baguette" illustrate this problem. One of them is a place of business and the other is a food, so the shortest path between them is long; although, these classes contain very similar images.

The second way is directly using the text embeddings of class names, which we refer to as **Text Distance Euclidean** (**TDE**) in the paper. We compute it as the Euclidean distance between the text embeddings of a concept pair, using the text encoder of the VLM. We follow the common practice of averaging templates (Radford et al., 2021; Ilharco et al., 2021).

One downside of TDE is that it relies on the embeddings of the same VLM where we are performing zero-shot classification. One may argue that it creates a circular dependency for the analysis or application before the training of the actual VLM. As an alternative, we use SBERT (Reimers & Gurevych, 2019) to obtain an embedding for a class, as we did in TDE, to obtain **SBERT Distance Euclidean** (**SDE**).

TDE and SDE are easier to apply to any set of class names as there is no requirement for them to occur in WordNet. Moreover, SYND is limited to discrete steps and in some cases it cannot distinguish finer degrees of similarity once the synset distance reaches its minimum (e.g., a distance of 2). In contrast, Euclidean distance is continuous.

4.2 Summarizing the Total Semantic Opposition

There is an important difference between concept frequency and semantic similarity quantifications. Concept frequencies directly belong to individual classes whereas semantic similarities belong to pairs of classes. Thus, for a set of classes, frequencies can be represented as 1D array, whereas pairwise semantic information forms a matrix. Consequently, it is not straightforward to summarize the "total semantic opposition" a single class encounters, with a scalar value.

In our transformer-based solution, we can incorporate the entire matrix, but a scalar summary is often more convenient for analysis. One simple way to perform such an operation is to consider only the distance to the nearest class, which is a reasonable starting point because it does not introduce any hyperparameter. Mean or median are not suitable since there are too many distant classes which would dominate the statistics. Using the nearest distance is meaningful because, generally, the most similar classes mainly determine the class performance while distant classes are unlikely to be confused, as shown in Fig. 1.

The trend between accuracy and summarization of semantic information with distance to the minimum class is shown in Fig. 4. We add **MIN-** prefix to the name of the semantic information and obtain **MIN-SYND**, **MIN-TDE**, and **MIN-SDE**. Apart from that, we also consider using the distances to the closest K classes and take their semantic distance averages, which we call **MEAN-K-SYND**, **MEAN-K-TDE**, and **MEAN-K-SDE**. (Fig. 9). K could depend on the dataset but throughout this

work, we use 5. Similar to LOGFREQ plots, we can observe a trend between semantic similarity and accuracy, but again, individual points show a high degree of scatter.

5 Generalized Zero-Shot Accuracy Prediction

Experimental Setting. To evaluate the models on downstream tasks, we divide the classes of each dataset evenly into two: one part used during training and the other part is left for testing. Because the particular split can affect the results, we repeat the experiment for 50 different random divisions. The divisions are kept constant throughout different experiments.

Not to limit ourselves to a single class set size, we also evaluate the model on sampled subsets of classes of various cardinalities for both training and testing. We employ two different sampling strategies: (i) fixed, uniform subset sizes of 8, 16, 32, 64, 128, and 256, (ii) mixed subset sizes logarithmically sampled between 2 and 256. For comparability, the sampled subsets are kept constant, except when training the Transfomer-based architecture in which we sample more subset for the training data (but still using the same division and the same test set). All factors and accuracies are calculated within subset.

We conduct experiments on ViT-B/32, ViT-B/16, ViT-L/14, pretrained on LAION400M (Schuhmann et al., 2021), LAION2B (Schuhmann et al., 2022), MetaCLIP400M and MetaCLIP-FULLCC (Xu et al., 2024). As downstream task, we use several: ImageNet1K (Russakovsky et al., 2015), SUN397 (Xiao et al., 2010; 2014), Places365 (Zhou et al., 2017), CUB Wah et al. (2011), Flowers Nilsback & Zisserman (2008), Food101 (Bossard et al., 2014), CIFAR100 (Krizhevsky et al., 2009), DTD (Cimpoi et al., 2014), OxfordPets (Parkhi et al., 2012), VOC2007 Everingham et al., EuroSAT (Helber et al., 2018; 2019).

5.1 Frequency and Semantic Similarity Together Estimate Accuracy Better

Udandarao et al. (2024) depicted a log-linear trend between concept frequencies and zero-shot accuracy of CLIP models. Our initial findings in Sec. 4.2 indicate that a similar relation exists for semantic-similarity-based factors as well. How about their combination? Would it explain the accuracy better or as we mentioned in Sec. 3, would incorporating the semantic information fully explain the relation between log-frequency and accuracy and the only reason literature found such a relation because of the entanglement between dataset class creation and concept frequencies? Initial observations on Fig. 5 shows a diagonally increasing trend, indicating semantic similarity and concept frequency together affects accuracy. To investigate whether combining semantic-based and frequency-based factors explains the accuracy better, we train linear models to estimate the accuracies in test divisions, using different sets of factors. Then, we measure Mean Squared Error (MSE), Mean Absolute Error (MAE), Coefficient of Determination (R^2), and Spearman's Rank Correlation Coefficient between predictions and ground truths. The results can be seen in Table 1 for ImageNet1K. As can be seen, combining frequency-based and semantic-based factors gives better estimations (MSE, MAE, Spearman) and explains the variance better (R^2).

Table 1: Performance of linear models trained on different factors. Using both concept frequency and semantic similarity together obtains better results in every combination. Related statistical tests are provided in the Appendix. LOGFREQ is calculated from Parashar et al. (2024). The values are averaged over 50 different divisions on ImageNet1K. The model is ViT-B/32.

Factors	MSE	MAE	R^2	Spearman
LOGFREQ only	0.0406	0.1636	0.0359	0.1757
MIN-SYND only	0.0418	0.1649	0.0088	0.1417
LOGFREQ + MIN-SYND	**0.0402**	**0.1620**	**0.0456**	**0.2149**
MIN-TDE only	0.0410	0.1638	0.0285	0.1486
LOGFREQ + MIN-TDE	**0.0389**	**0.1600**	**0.0764**	**0.2408**
MIN-SDE only	0.0408	0.1632	0.0325	0.1851
LOGFREQ + MIN-SDE	**0.0393**	**0.1606**	**0.0669**	**0.2427**

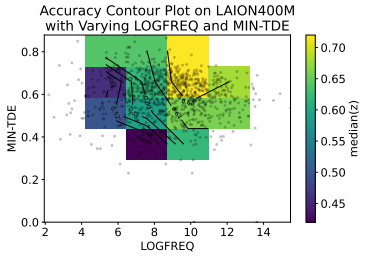

Figure 5: Contour plot and heatmap of binned median accuracies over the MIN-TDE and LOGFREQ grid. Accuracies are computed over complete ImageNet1K using ViT-B/32 pretrained on LAION-400M. Bins with < 10 examples are not shown.

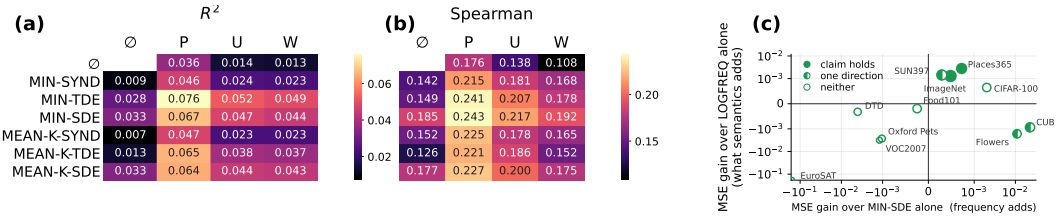

Figure 6: R^2 (a) and Spearman's Coefficients (b) of accuracy predictions of linear models trained on different combinations of factors. Vertical axis shows different semantic similarity summarization methods and horizontal axis shows three different frequency quantification methods are used: P (Parashar et al. (2024)), U (Udandarao et al. (2024)), W (Wen et al. (2024)). Combining frequency and semantic similarity (a) explains the variance better and (b) predicts the actual accuracy ranks better. The results are averaged over 50 different divisions on ImageNet1K. (c) MSE gains of LOGFREQ+MIN-SDE compared to their individual constituents on several downstream datasets. Statistical significance in both directions is depicted with filled circles. This result is significant for the datasets with larger number of classes. The model is ViT-B/32.

To see whether this result is generalizable to other concept frequency quantification works and other semantic summarization techniques, we draw heatmaps of combinations for ImageNet1K in Fig. 6 , observing that nearly all combinations (only exception is Wen + MEAN-K-SDE out of 72) obtain better results than their individual constituents. The same holds for MAE and MSE, which can be seen in the Appendix E. MIN-TDE and MIN-SDE seem to perform the best among the semantic similarity quantifications both when used individually and when combined with frequency-based factors. This may be caused by their ability to distinguish very similar classes while a very large portion of MIN-SYND values pile up at the smallest SYND, which is 2.

To evaluate the significance of the results, we perform two-way tests to check if a frequency factor puts on top of a semantic one and the other way around. We calculate paired t-statistics and compare them to that of 20000 random permutations. We find that all datasets that have ≥ 200 classes yield a statistically significant improvement in both directions, except CUB and SUN397 have only one (opposite) directions significant. The details of the test are provided in the Appendix E. Fig. 6c summarizes the significance test results where the filled circles show the significant results and as can be seen, all significant results support the claim.

Our results show that combining semantic and frequency information estimates the zero-shot accuracy better and this is a strong indication of the relation between frequency and accuracy not being caused only from class selection.

5.2 Relation Between Factors and Zero-Shot Accuracy Under Varying Conditions

One of the striking findings of Wen et al. (2024) is the weakening relation between concept frequency and accuracy as the pretraining data scales. In this section, we investigate how semantic-similarity-based factors behave under similar varying conditions and if frequency and semantic-based factors still work better together on 4 datasets that have ≥ 200 classes. We again train linear models and compare the Spearman Correlation Coefficients to the zero-shot accuracies.

As one of our goals is to estimate the zero-shot accuracy given a set of class names, we first analyze the change of relation strength under different sizes of class subsets. Fig. 7a shows that as subset size gets smaller, the relation between frequency-based factors and accuracy weakens significantly. On the other hand, semantic factors do not share that direction consistently: MIN-TDE is flat, MIN-SYND (ImageNet-only) is stronger on small subsets, and MIN-SDE moves in opposite directions on different datasets. The reason for the weakening relation of frequency could be the fact that the sampling is done randomly and it becomes rare for the small subsets to contain semantically too similar classes. On ImageNet and SUN397, LOGFREQ obtains near-zero correlation on 8-class subsets whereas CUB can be considered as an exception: as all of its classes are bird species, even an 8-class subset stays semantically crowded. The mean MIN-SDE inside 8-class subsets is 0.82 where as this value is 0.96 for Places365, 0.97 for SUN397, and 0.99 for ImageNet1K. Therefore, as in Fig. 1, when the classes are distant, even low frequency classes can get high accuracies. Note that combining factors gives at least as good results as its constituents and mimic their overall trend.

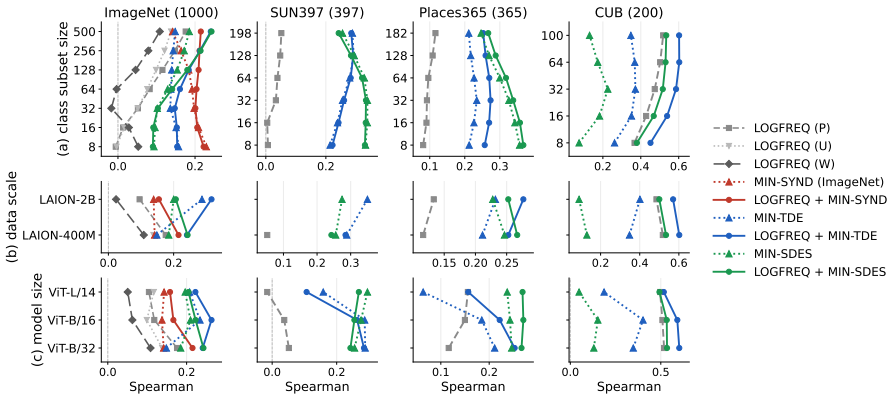

Figure 7: Spearman Coefficients under varying subset size, pretraining data scales, and model sizes. (a) Frequency-based factors lose their predictive ability as the subset size decreases while semantic-similarity-based ones remain relatively stable. (b) Similar to findings of Wen et al. (2024), as the pretraining data size increases, frequency and accuracy relation generally weakens. This is not always the case for semantic-similarity-based factors. (c) As the model size increases, the relation between frequency and accuracy generally weakens. The results are averaged over 50 different divisions on 4 datasets. The model is ViT-B/32. For the subset experiment, LAION400M is used. For the data scale experiment, subset size of half of the dataset is used.

We also vary the pretraining dataset scale in Fig. 7b. The frequency-based factor results are in accordance with Wen et al. (2024) (except Places365). Although scaling the data may lessen the detrimental effects of frequency imbalance, MIN-SYND (ImageNet-only) is not affected by scaling and interestingly MIN-TDE achieves even higher correlations in the scaled data. It may indicate that as data scales, the unexplainable factors captured by MIN-TDE lessen and the zero-shot accuracy becomes more explainable by the model's own embeddings. MIN-SDE shows different trends on different datasets but remains stable. When the pretraining dataset is smaller (Fig. 5 vs Fig. 10), the contours are diagonal, whereas for larger pretraining datasets, their direction is more aligned with the semantic similarity factor. A similar trend is observed in MetaCLIP (Xu et al., 2024), although the decline in the correlation between LOGFREQ and accuracy is less pronounced as the correlations are closer to 0 (0.019 at MetaCLIP400M against 0.108 on LAION400M, Appendix F). Similar to the previous findings, combining factors gives at least as good results as its constituents.

Finally, under increasing model sizes, frequency-based factors again show a weakening trend (except Places365) while this is not the case for all semantic-similarity-based factors (Fig. 7c). In particular, MIN-SYND(ImageNet-only) is not affected by model scaling and MIN-SDE mostly preserves a stable correlation (except CUB). However, MIN-TDE seems to get affected from model scaling negatively, consistently through all datasets with ViT-L/14 compared against ViT-B/16.

Aside from the fact that semantic-similarity-based factors can be used as predictors, they can be used as a part of the explanation when frequency performs well. For example, frequency survived at small subset sizes on CUB because the subsets stayed semantically crowded whereas you generally do not obtain similar classes when you sample 8 classes out of 1000 because of the semantic distribution of ImageNet classes. Additionally, under all of those variations, combining the two families is the safe and generally beneficial choice.

5.3 A Graph-Transformer to Predict Zero-shot Accuracy

We now explore whether a model that reads the whole semantic structure can predict accuracy better and whether a single model can generalize to multiple datasets.

Problem Formulation. To handle inputs representing a variable number of class frequencies and relationships, we formulate accuracy prediction as a graph regression problem and use graph-based networks. We represent the problem as a graph $\mathcal{G} = (V, E)$ where the vertices V correspond to classes encoded by their frequencies (and optionally text embeddings, e.g., using SBERT) and the edges E represent the degree of semantic similarity between the classes. With this approach, we predict the accuracies of all $N = |V|$ classes at once from N frequency (+ text embedding) and $N \times N$ semantic similarity information where N does not have to be a constant.

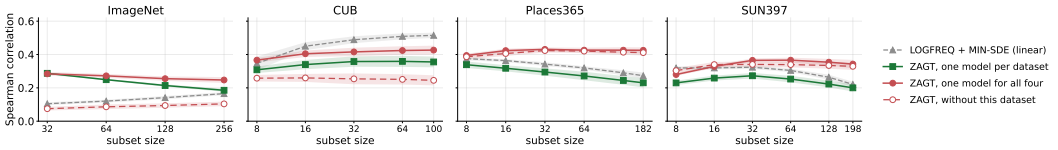

Figure 8: **Training one ZAGT for four datasets obtains better results for all datasets compared to training individual models per each.** Spearman's coefficients are shown over different models and different training settings. Each is a mean over ten class splits. *one model for all four* is the mixture, *one model per dataset* is the same architecture trained on that dataset alone, and *without this dataset* is the mixture trained on the other three, so the dataset is never seen. The mixture is above the per-dataset model in every panel and at all but one subset size. Wen et al. (2024) is used as the frequency source as it is available in all datasets and SDE is used for the semantic source.

ZAGT Architecture. We introduce a graph Transformer, called ZAGT (Zero-shot Accuracy Graph-Transformer), combining Graphormer's (Ying et al., 2021) spatial encoding and Shaw et al. (2018)'s relation-aware value encoding (similar to Hussain et al. (2022)). The frequency and text embedding are given as the input sequence and semantic distances describe the relation encoding. The model predicts all class accuracies in a single pass. The architecture and related formulas are discussed in Appendix G. As SDE is readily available for all datasets and, unlike TDE, does not depend on the VLM's own text encoder, we use SDE throughout this section. In Appendix I, we show the results of our ablation studies over our decisions.

Experiment Setup. Each batch has uniform subset size which is selected from log-uniformly between 2 and 256. We provide further details and the PyTorch-like code for the modified attention in Appendix H. We train 2 types of ZAGT: (i) one model per dataset, (ii) one model for multiple datasets. We use 4 datasets that have number of classes ≥ 200: ImageNet1K, SUN397, Places365, CUB, and when mixed, they are sampled with 0.4, 0.2, 0.2, 0.2 ratios, respectively. We use Wen et al. (2024) as the frequency source since we have its estimations for all 4 datasets.

Results. Fig. 8 shows that training a single model over all 4 datasets obtains better results compared to training one per dataset (except ImageNet subset size 32 where the results are close). This supports the idea that the learned information from one dataset is still valid for the others. To further test this, for each dataset, we leave that dataset out from training and use it as a held-out test set. As can be seen in Fig. 8, predictions always show positive correlation with accuracy and even bests ZAGT trained specifically on that dataset on Places365 and SUN397. We also compare ZAGT (trained on all datasets) to the linear baseline. Mixture model obtains better or comparable results, except for CUB dataset. In Sec. 5.2, our analysis showed that semantic distance inside a subset was significantly lower for CUB compared to the others and the most correlating factor was LOGFREQ. Therefore, there may be not much to add with our graph-base approach.

6 Conclusion, Discussion & Limitations

In this paper, we analyzed how the zero-shot accuracy of various CLIP models is influenced by both pretraining concept frequency and the semantic competition among target classes. While previous studies highlight a log-linear trend between frequency and accuracy, we demonstrate that this correlation weakens significantly for smaller subsets. Furthermore, without binning the raw data points, this linear trend is largely obscured by high variance. Ultimately, we establish that semantic similarity is a crucial determinant of zero-shot performance. By developing predictive models, ranging from linear regression to a modified Graphormer (Ying et al., 2021), we show that jointly leveraging semantic similarity and concept frequency yields more accurate estimates than relying on either metric alone.

Since our approach relies solely on semantic similarity and concept frequencies, it cannot distinguish between datasets that share identical class names, such as ImageNet and ImageNet Sketch (Wang et al., 2019). While adjusting prompt templates for specific variants might improve performance, our current framework does not address template-based factors. Similarly, our model cannot account for the low accuracy of CIFAR10's cat and dog classes, as it lacks awareness of input image resolution. Ultimately, our work contributes a critical new perspective to the frequency-accuracy relationship, but we acknowledge that a comprehensive zero-shot accuracy estimator must eventually incorporate these additional visual and structural variables.

AI USE STATEMENT

In this work, we had a human-built codebase which was read-only to LLM agents. We performed fast exploration together with LLM agents on a separate codebase where the functionalities mostly depended on the read-only human-maintained code-base so that we were aware of what those experiments were doing. LLM agents conducted the experimentation, running the scripts and tracking the results of the experiments and sometimes proposed next steps. They only modified this separate codebase but they were free to use the compute resources. Whenever a fruitful path was found, it was re-implemented in the human-build codebase by humans. The final results in this paper are the results of this human-build codebase. This covers the numerical results and base of the statistical tests. For plotting purposes, both LLM and human written code were used as it is easy to cross-check with the actual results. Although the sentences were written by humans, for grammar checking or writing suggestions, LLMs were used. We have also discussed the statistical robustness of our framework with LLMs to make it statistically sound.

In the terms of the disclosure list, this means we used LLMs to design and give feedback on our experiments and statistical methodology, to implement methods in the exploratory codebase and in the plotting code, and to interpret results. We also used them to draft some figure captions and appendix passages and to edit the manuscript for readability. We did not use them to find or survey related work, to clean or reformat any dataset, nor for translation. The remaining tasks on that list do not arise in this work: it contains no synthetic data, no theoretical model or conceptual framework, no mathematical claims or proofs, and no qualitative or thematic analysis.

We have reviewed all AI-assisted work. Every number reported in the paper comes from the human-built codebase, run by the authors, and claims made about those numbers were checked against the underlying results before being written down. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Although our analysis focuses on large, web-crawled datasets, we only analyzed the pretrained models on these corpora, used pre-calculated concept counts on them and ran frequency quantification methods where a specific downstream dataset is missing from their framework. Imbalance itself can be seen as a bias, and our work is a step to uncover possible biases in these datasets. We did not collect any data or conduct a human study.

REPRODUCIBILITY STATEMENT

The class divisions and the sampled subsets are generated from fixed seeds. We had to ensure that was the case for paired tests and different models had to be evaluated under the same data. We will release our code on GitHub after the anonymity restrictions are lifted.

The protocol behind every result is described in Sec. 5; the permutation floor in Appendix A, the two-way permutation test and its multiple-comparison correction in Appendix E, and ZAGT's architecture, training setup, pseudo-code and ablations in Appendices G and I.

Semantic information was extracted from openly available CLIP models and SBERT (Reimers & Gurevych, 2019). Synset distances can be obtained from openly available WordNet.

Frequency counts are mostly obtained from the works of Parashar et al. (2024), Udandarao et al. (2024), Wen et al. (2024), except we extracted the frequencies for SUN397 using the framework of Wen et al. (2024). However, this run used the original LAION400M which is currently not available. Therefore, we will share the frequencies we have extracted so that all of the tests can be run.

REFERENCES

Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2011–2018, 2014.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

Christiane Fellbaum. *WordNet: An electronic lexical database*. MIT press, 1998.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference Learning Representations*, 2022.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Introducing eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 204–207. IEEE, 2018.

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*, 2019.

Xinyu Huang, Yi-Jie Huang, Youcai Zhang, Weiwei Tian, Rui Feng, Yuejie Zhang, Yanchun Xie, Yaqian Li, and Lei Zhang. Open-set image tagging with multi-grained text supervision. In *ACM International Conference on Multimedia*, pp. 4117–4126, 2025.

Md Shamim Hussain, Mohammed J Zaki, and Dharmashankar Subramanian. Global self-attention as a replacement for graph convolution. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 655–665, 2022.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL `https://doi.org/10.5281/zenodo.5143773`.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 19606–19616, 2023.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7482–7491, 2018.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *International Conference on Computer Vision*, pp. 4015–4026, 2023.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Technical Report, Toronto, ON, Canada*, 2009.

Yang Liu, Tianwei Zhang, and Shi Gu. Hybrid concept bottleneck models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 20179–20189, 2025.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

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*, Dec 2008.

Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *International Conference Learning Representations*, 2023.

Shubham Parashar, Zhiqiu Lin, Tian Liu, Xiangjue Dong, Yanan Li, Deva Ramanan, James Caverlee, and Shu Kong. The neglected tails in vision-language models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12988–12997, 2024.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

Helen Qu and Sang Michael Xie. Worse together: Understanding the brittleness of multimodal models on rare concept pairs. In *International Conference Learning Representations*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Kevin Robbins, Xiaotong Liu, Yu Wu, Le Sun, Grady McPeak, Abby Stylianou, and Robert Pless. Will it zero-shot?: Predicting zero-shot classification performance for arbitrary queries. In *2025 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 1513–1523. IEEE, 2025.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115:211–252, 2015.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *NeurIPS Workshop on Data Centric AI*, 2021.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 464–468, 2018.

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, pp. 14274–14289, 2022.

Vishaal Udandarao, Ameya Prabhu, Adhiraj Ghosh, Yash Sharma, Philip Torr, Adel Bibi, Samuel Albanie, and Matthias Bethge. No "zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. *Advances in Neural Information Processing Systems*, 37:61735–61792, 2024.

Chris Vorster, Mayug Maniparambil, Noel E. O'Connor, Noel Murphy, and Derek Molloy. Underrepresented in foundation model pretraining data? a one-shot probe. *arXiv preprint*, 2024.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pp. 10506–10518, 2019.

Xin Wen, Bingchen Zhao, Yilun Chen, Jiangmiao Pang, and Xiaojuan Qi. What makes clip more robust to long-tailed pre-training data? a controlled study for transferable insights. In *Advances in Neural Information Processing Systems*, volume 37, pp. 36567–36601, 2024.

Thaddäus Wiedemer, Yash Sharma, Ameya Prabhu, Matthias Bethge, and Wieland Brendel. Pretraining frequency predicts compositional generalization of clip on real-world tasks. 2024.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010. doi: 10.1109/CVPR.2010.5539970.

Jianxiong Xiao, Krista A. Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Exploring a large collection of scene categories. *International Journal of Computer Vision*, 119: 3–22, 2014. URL https://api.semanticscholar.org/CorpusID:10224573.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. In *International Conference Learning Representations*, 2024.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 19187–19197, 2023.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? In *Advances in Neural Information Processing Systems*, volume 34, pp. 28877–28888, 2021.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *International Conference on Computer Vision*, pp. 11975–11986, 2023.

Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(9):10795–10816, 2023.

Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1724–1732, 2024.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 16793–16803, 2022.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *International Conference Learning Representations*, 2024.

Xingyu Zhu, Beier Zhu, Yi Tan, Shuo Wang, Yanbin Hao, and Hanwang Zhang. Enhancing zero-shot vision models by label-free prompt distribution learning and bias correcting. *Advances in Neural Information Processing Systems*, 37:2001–2025, 2024.

Orr Zohar, Shih-Cheng Huang, Kuan-Chieh Wang, and Serena Yeung. Lovm: Language-only vision model selection. *Advances in Neural Information Processing Systems*, 36:33120–33132, 2023.

TABLE OF CONTENTS: APPENDIX

A Single-Factor Spearman Correlation Coefficient Analysis **14**

B Factor Comparison with LOVM **15**

C Semantic Factors under MEAN-K Summarisation **16**

D Contour Plots **16**

E Details of the Statistical Tests Performed for the Linear-Model-based Analysis **17**

F Effect of Pretraining Data Size on Factors on MetaCLIP **18**

G ZAGT Architectural Details **19**

H Modified Attention Pseudo-Code **20**

I ZAGT Ablation Studies **20**

A SINGLE-FACTOR SPEARMAN CORRELATION COEFFICIENT ANALYSIS

As the number of classes in a dataset drops, it becomes more probable that a correlation exists by chance. The p ermutation test gives us a way to assess what would be the probability of such a correlation occurring randomly (even in the case of repetitive values in accuracy). We basically take our factor and shuffle the corresponding accuracy vector 20000 times and take absolute value of the correlations there. Afterwards, 95th percentile of the resultant absolute correlations become our threshold corresponding to p value of 0.05. If a correlation passes that threshold, we can say that the probability of that occurring is less than 0.05 under the assumption of null hypothesis. The correlations that pass their thresholds, and therefore, significant, are shown in Table 2. As we do not claim anything inter-dataset, we do not perform correction.

Table 2: Class-level Spearman correlation between each factor and per-class zero-shot accuracy (ViT-B/32, LAION-400M), over all classes of each dataset. The last column is the 95th percentile of $|\rho|$ under 20000 permutations of the accuracy vector. Correlations that clear it are in bold. MIN-SYND needs WordNet synsets and is therefore only defined on ImageNet. LOGFREQ values are taken from Wen et al. (2024).

Dataset	#classes	LOGFREQ	MIN-SYND	MIN-TDE	MIN-SDE	95% floor
ImageNet	1000	**0.124**	**0.135**	**0.156**	**0.195**	0.062
SUN397	397	0.063	–	**0.270**	**0.178**	0.099
Places365	365	**0.126**	–	**0.193**	**0.198**	0.103
CUB	200	**0.531**	–	**0.323**	0.106	0.139
Flowers	102	**0.455**	–	-0.077	-0.002	0.194
Food101	101	-0.091	–	-0.033	-0.032	0.196
CIFAR-100	100	**0.369**	–	0.006	**0.264**	0.197
DTD	47	0.110	–	**0.321**	0.083	0.286
Oxford Pets	37	0.198	–	0.228	0.243	0.328
VOC2007	20	-0.307	–	-0.344	-0.119	0.444
EuroSAT	10	0.200	–	-0.067	0.055	0.636

Table 3: Spearman Rank Correlation Coefficients between concept frequency quantification methods on ImageNet1K (Russakovsky et al., 2015). The methods are P (Parashar et al. (2024)), W (Wen et al. (2024)), and U (Udandarao et al. (2024)). $p < 0.001$ for all.

Method 1	**Method 2**	**Spearman**
P	W	0.83
P	U	0.66
W	U	0.68

Table 4: Pearson and Spearman Correlation Coefficients between estimated concept log-frequencies and class-wise accuracies on ImageNet1K (Russakovsky et al., 2015). The model is ViT-B/32 and the pretraining dataset is LAION400M (Schuhmann et al., 2021). $p < 0.001$ for all.

Method	**Pearson**	**Spearman**
Parashar et al. (2024)	0.22	0.19
Wen et al. (2024)	0.15	0.12
Udandarao et al. (2024)	0.15	0.15

B Factor Comparison with LOVM

LOVM (Zohar et al., 2023) predicts the zero-shot accuracy of a (model, dataset) pair from text alone, fitting one linear model per VLM–dataset combination over 35 VLMs and 23 datasets. Its scores are computed in the evaluated model's own text space, from the class names and from captions and synonyms generated by a language model. Two of them are dataset-level counterparts of our factors.

Fisher criterion and MIN-TDE. LOVM's Fisher criterion is the mean over classes of the largest cosine similarity between a class's prompt embedding and that of any other class. As they use VLM's own text encoder for that, it is very similar to our TDE-based scores. If prompt embeddings are unit-length, then selecting the maximum cosine similarity is the same as selecting the minimum TDE. Therefore, LOVM calculates average of MIN-TDE for the dataset with this score.

Synonym consistency and LOGFREQ. To measure the familiarity of a VLM to a specific concept, LOVM generates a set of synonyms for that concept and calculate if VLM encodes them to similar locations. In our case, degree of familiarity is inferred from the number of occurrences that a concept has in the pretraining dataset.

What does not overlap. LOVM additionally generates captions using an LLM to use them as if they are images and calculate scores that correspond to variances of classes, which we do not have a similar score here. On the other hand, we have our model-independent scores using SBERT.

Why the correlations are not comparable. LOVM specifically works on dataset-level and therefore, they average these scores. Averaging over classes removes exactly the variation we try to explain. LOVM's reported R^2 of 0.55 is across (model, dataset) pairs; ours are across the classes of one dataset under one model.

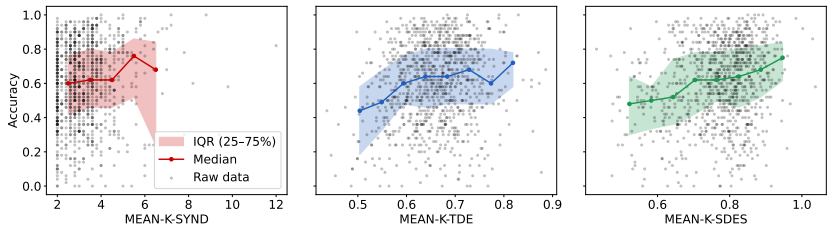

Figure 9: Relation between the MEAN-K summarized semantic similarity scores and zero-shot class accuracies on ImageNet1K. The line shows the medians of bins and the shaded band IQR. Bins with ≤ 15 classes are not drawn. The model is ViT-B/32 pretrained on LAION400M.

C Semantic Factors under MEAN-K Summarisation

Instead of taking the nearest, we try to summarize the vicinity of a class to get a better understanding of its semantic opposition. Averaging over the $K = 5$ nearest classes instead of taking only the nearest one smoothens the picture (Fig. 9).

D Contour Plots

We share the accompanying LAION2B figure of Fig. 5 in Fig. 10.

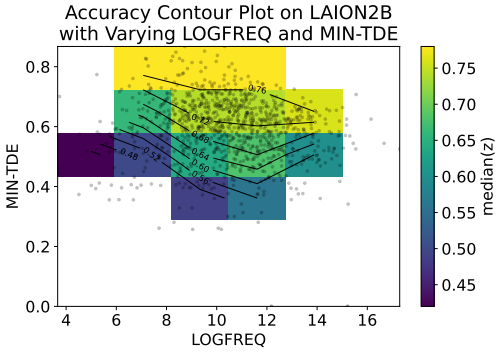

Figure 10: Contour plot and heatmap of binned median accuracies over the MIN-TDE and LOGFREQ grid. Accuracies are computed over complete ImageNet1K using ViT-B/32 pretrained on LAION2B. Bins with < 10 examples are not shown.

E Details of the Statistical Tests Performed for the Linear-Model-based Analysis

The ultimate goal is to understand whether a frequency factor X and a semantic factor Y estimate accuracy better than either factor alone. To do that, we divide the question into two:

- Does X add on top of Y?
- Does Y add on top of X?

and the result will be yes if both are true.

The most basic predictor would be a linear model and we go with that for this analysis. The training has 2 splits: train and test. How we split the dataset classes into two changes the outcome drastically. Therefore, we perform 50 different train test splits. Train and test splits are held constant throughout the tests so that we can calculate paired statistics. Directly applying a paired test would be optimistic: the 50 divisions reuse the same pool of classes, so their differences are not independent draws and the t-distribution is the wrong reference. We still calculate the t-test statistic; however, we perform permutation tests. Again we perform 20000 permutations and for each permutation, the same 50 split is re-evaluated and test statistics are calculated. With this, the distribution of test statistics are obtained under the assumption of null hypothesis. Permutation is done in two ways: while we are testing whether frequency-based factor X adds on top of semantic-based Y, we randomly shuffle X and perform the 50 split normally, as X is a property of a class. On the other hand, when we are testing if semantic factor Y adds on top of X, we first shuffle the text embeddings and re-calculate the factor Y (for example, MIN-SDE), as Y is not directly a property of a class and its value depends on the other classes that share the same split.

After obtaining statistics from the permutation test and sort them, we find the place of the original statistic in that list and estimate a p value for that. To obtain Fig. 6c, we perform an additional Holm correction over 11 datasets and 2 ways. We threshold the final p values with 0.05 to obtain the filtered set of datasets. We use MSE as our metric in this test while creating Fig. 6c. Detailed results can be seen in Table 5.

The MSE and MAE heatmaps can be seen in Fig. 11. Similar to Spearman and R^2 results, combining frequency and semantic metrics improves the accuracy estimation results.

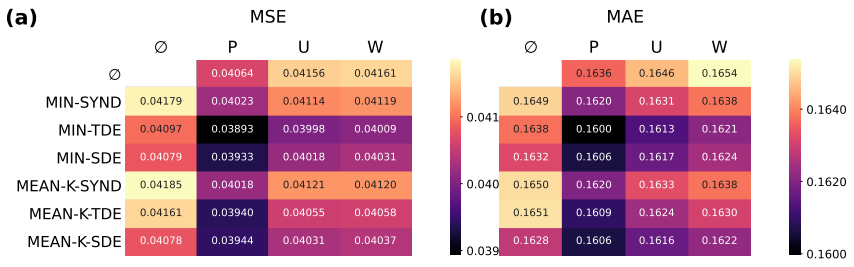

Figure 11: MSE (a) and MAE (b) of accuracy predictions of linear models trained on different combinations of factors. Vertical axis shows different semantic similarity summarization methods and horizontal axis shows three different frequency quantification methods are used: P (Parashar et al. (2024)), U (Udandarao et al. (2024)), W (Wen et al. (2024)). Combining frequency and semantic similarity estimates the accuracy better. The results are averaged over 50 different divisions on ImageNet1K. The model is ViT-B/32. Lower is better.

Table 5: Two-way permutation tests behind Fig. 6c: LOGFREQ + MIN-SDE against each of its parts, measured by MSE (ViT-B/32 on LAION-400M, Wen et al. (2024) counts, 50 class divisions, 20,000 permutations). *freq adds* is the gain of the combination over MIN-SDE alone, *sem adds* its gain over LOGFREQ alone; a positive gain means the combination has the lower MSE. p is the one-sided permutation p-value and p_{Holm} its Holm-corrected value over the 22 tests of the figure. A direction is marked ✓ when its gain is positive and $p_{\mathrm{Holm}} < 0.05$; the last column marks the datasets where both directions hold, which is the claim that the combination beats both of its parts.

		what frequency adds				what semantics adds				
Dataset	#cls	gain	p	p_{Holm}		gain	p	p_{Holm}		both
ImageNet	1000	$+4.8{\cdot}10^{-4}$	0.001	**0.010**	✓	$+1.3{\cdot}10^{-3}$	$< 10^{-4}$	**0.001**	✓	✓
SUN397	397	$+2.9{\cdot}10^{-4}$	0.011	0.149	–	$+1.4{\cdot}10^{-3}$	$< 10^{-4}$	**0.001**	✓	–
Places365	365	$+7.3{\cdot}10^{-4}$	0.002	**0.034**	✓	$+2.8{\cdot}10^{-3}$	$< 10^{-4}$	**0.001**	✓	✓
CUB	200	$+2.3{\cdot}10^{-2}$	$< 10^{-4}$	**0.001**	✓	$-9.4{\cdot}10^{-4}$	0.137	1	–	–
Flowers	102	$+1.1{\cdot}10^{-2}$	0.001	**0.023**	✓	$-1.6{\cdot}10^{-3}$	0.169	1	–	–
Food101	101	$-2.5{\cdot}10^{-4}$	0.324	1	–	$-1.9{\cdot}10^{-4}$	0.351	1	–	–
CIFAR-100	100	$+2.0{\cdot}10^{-3}$	0.007	0.107	–	$+6.5{\cdot}10^{-4}$	0.013	0.168	–	–
DTD	47	$-4.0{\cdot}10^{-3}$	0.288	1	–	$-3.2{\cdot}10^{-4}$	0.027	0.321	–	–
Oxford Pets	37	$-1.0{\cdot}10^{-3}$	0.156	1	–	$-2.7{\cdot}10^{-3}$	0.079	0.867	–	–
VOC2007	20	$-1.2{\cdot}10^{-3}$	0.085	0.867	–	$-3.2{\cdot}10^{-3}$	0.382	1	–	–
EuroSAT	10	$-1.6{\cdot}10^{-1}$	0.618	1	–	$-1.8{\cdot}10^{-1}$	0.716	1	–	–

F Effect of Pretraining Data Size on Factors on MetaCLIP

As we did in Sec. 5.2 for LAION, we perform the same dataset scaling test on MetaCLIP, which is a concept-frequency-wise curated dataset. The results are similar to Fig. 7: LOGFREQ drops as the dataset size increases, MIN-TDE shows an increasing trend and MIN-SDE shows more stable but different trends depending on the downstream dataset.

One of the important thing is that the correlation with the frequency is substantially lower than that of LAION on ImageNet and CUB.

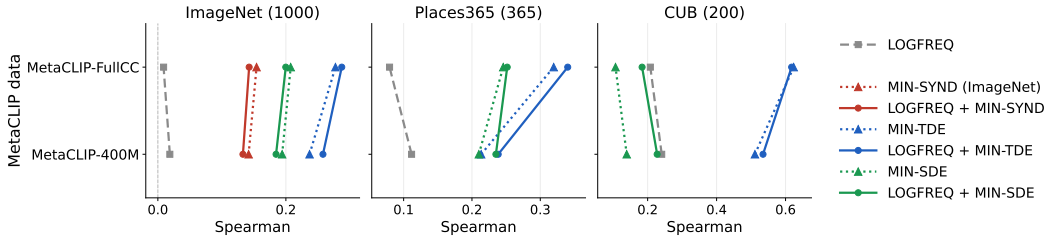

Figure 12: **On MetaCLIP too, frequency loses ground as the corpus grows while the semantic factors gain.** Spearman correlation between each factor and per-class zero-shot accuracy on MetaCLIP-400M and MetaCLIP-FullCC (Xu et al., 2024), averaged over 50 class divisions, counts from Wen et al. (2024). Models are the ViT-B/32-quickgelu. SUN397 has no Wen counts for these corpora and MIN-SYND needs WordNet synsets, so neither appears.

G ZAGT Architectural Details

Background: Graphormer. Graphormer (Ying et al., 2021) is a Transformer-based GNN that proposes three embeddings to be used in graph-specific self-attention across nodes and edges: (i) Centrality embedding, which represents a node's significance in the graph based on its in-degrees and out-degrees, (ii) spatial embedding, which encodes the "position" of a node in the graph, and (iii) edge embedding, which captures semantic length of a path connecting two nodes in the graph.

Background: Relation-aware Self-attention. Shaw et al. (2018) proposed a relation-aware self-attention which extends value calculation as follows:

$$z_i = \sum_{j=1}^{N} \alpha_{ij}(h_j W^V + a_{ij}), \tag{1}$$

where $\alpha_{ij} = \text{Softmax}((h_i W^Q)(h_j W^K)^T/\sqrt{d})$ is the attention weight; W^K, W^Q, W^V are the key, query and value matrices; $h_\cdot$ is the input to self-attention; and finally, a_{ij} is an embedding representing the relation between nodes i and j.

ZAGT: Method and Architecture. We introduce a graph Transformer, called ZAGT (Zero-shot Accuracy Graph-Transformer), combining Graphormer's (Ying et al., 2021) spatial encoding and Shaw et al. (2018)'s relation-aware value encoding (similar architectures have been explored for different tasks in the literature (Hussain et al., 2022)):

Spatial Embedding. Similar to Graphormer, based on the semantic similarity $\text{sim}(v_i, v_j)$ between two vertices v_i and v_j, we learn a scalar bias $b_{\text{sim}(v_i,v_j)}$ and add it to the attention weights:

$$\alpha_{ij} = \frac{(h_i W_Q)(h_j W_K)^T}{\sqrt{d}} + b_{\text{sim}(v_i,v_j)}. \tag{2}$$

Although SYND would directly fit into this definition, to incorporate TDE and SDE, we pass their sinusoidal encodings from a linear layer to obtain an single scalar.

Relation-aware Attention. Following Shaw et al. (2018), for every value of $\text{sim}(v_i, v_j)$, we introduce learnable embeddings of hidden-dimension size (d) and combine them with the attention values (similar to Eq. 1):

$$z_i = \sum_{j=1}^{N} \alpha_{ij}(h_j W_V + e_{\text{sim}(v_i,v_j)}), \tag{3}$$

where $e_{\text{sim}(v_i,v_j)}$ denotes semantic similarity embedding. This way, a node's representation is updated to incorporate its semantic similarity to other nodes. This time, to obtain an embedding for TDE and SDE, similar to what we did to obtain bias, we again pass the sinusoidal encodings through linear, but this time, the output of the linear layer is the embedding dimension of the transformer.

We do not use Graphormer's (i) centrality embeddings since all nodes have the same degree in our setting or (ii) edge embeddings since all nodes are directly connected and edge semantics is incorporated through the spatial embedding calculations.

Training Loss. We use $L2$ loss with loss attenuation (Kendall & Gal, 2017; Kendall et al., 2018) and take average over all classes in that subset:

$$\mathcal{L} = \sum_{i=1}^{N} \frac{1}{2\sigma_i^2} ||y_i - \hat{y}_i||^2 + \log \sigma_i, \tag{4}$$

σ_i is estimated by the model (for numerical stability, $\log \sigma$ is estimated), y_i is the ground truth accuracy and $\hat{y}_i$ is the estimated accuracy, for class i. We use the loss attenuation only to account for high noise and do not use σ.

Training and Evaluation Setup. ZAGT is a 6-layer Transformer with an embedding dimension of 128, 8 attention heads and a feed-forward width of $4d$. Nodes carry the class's log frequency, taken from Wen et al. (2024), together with its SBERT embedding; edges carry the SBERT distance (SDE), encoded as described above. We train with Adam (lr $= 10^{-4}$, weight decay 3×10^{-4}, cosine schedule) for 20000 iterations at batch size 64 in bfloat16, under the $L2$ loss with attenuation. Each batch uses a single subset size, drawn log-uniformly between 2 and the size of the training half, maxed at 256; one batch therefore holds 64 subsets of that size. The per-dataset models are trained on one dataset's training half, the mixture on all four with sampling weights 0.4 for ImageNet and 0.2 for each of CUB, Places365 and SUN397, and the leave-one-out models on the remaining three with those weights renormalised. No checkpoint selection is performed: the model at the final iteration is the one evaluated, and we verified on the training curves that performance saturates well before the budget ends.

At evaluation the model sees the held-out class half and is scored on the same frozen subsets as the linear models, 200 per size, with Spearman between predicted and true class accuracy. The subset sizes evaluated are 32 to 256 on ImageNet, 8 to 100 on CUB, 8 to 182 on Places365 and 8 to 198 on SUN397, the largest value in each case being the whole test half. ZAGT results are averaged over ten class splits rather than the fifty used for the linear analyses, because every split requires training a separate model; the ablations in Appendix I use three. The splits are the same ones the linear models use, so the two are directly comparable split by split.

H Modified Attention Pseudo-Code

We share the self-attention part of ZAGT in Listing 1. The function is adapted from the code of Kirillov et al. (2023).

```
class SelfAttentionWithSpatialBiasAndValue(nn.Module):
    def __init__(self, dim, num_heads=8):
        super().__init__()
        self.num_heads = num_heads
        head_dim = dim // num_heads
        self.scale = head_dim**-0.5
        self.qkv = nn.Linear(dim, dim * 3)
        self.proj = nn.Linear(dim, dim)

    def forward(self, x, spatial_biases, spatial_values):
        B, L, H  = x.shape
        # qkv with shape (3, B, nHead, H * W, C)
        qkv = self.qkv(x).reshape(B,L,3,self.num_heads,-1).permute(2,0,3,1,4)
        q, k, v = qkv.reshape(3, B * self.num_heads, L, -1).unbind(0)

        attn = (q * self.scale) @ k.transpose(-2, -1)
        attn = attn.reshape(B, self.num_heads, L, L)
        attn = attn + spatial_biases.unsqueeze(1)
        attn = attn.reshape(B * self.num_heads, L, L)
        attn = attn.softmax(dim=-1)
        spatial_values = spatial_values.reshape(B, L, L, self.num_heads, -1).permute(0, 3, 1,
    2, 4).reshape(B * self.num_heads, L, L, -1)
        attended_spatial_values = torch.einsum('bij,bijk->bik', attn, spatial_values)
        x = ((attn @ v) + attended_spatial_values).view(B, self.num_heads, L, -1).permute(0,
    2, 1, 3).reshape(B, L, -1)
        x = self.proj(x)
        return x
```

Listing 1: PyTorch-like implementation of the self attention used in ZAGT.

I ZAGT Ablation Studies

We ablate several of our architectural decisions. First of all, about whether to use text embedding itself (e.g., SBERT's class embeddings as input) in Table 6. In all of the three experiments, we

Table 6: Input ablation for ZAGT. Spearman between predicted and true class accuracy on the frozen subsets, averaged over the evaluated subset sizes and over three class splits. The mixture model trained on all four datasets. *semantic distances only* drops the class embedding from the nodes, which then carry only log frequency; *text embedding only* drops the distance graph, and with it the spatial and value encodings that carry it, leaving a plain Transformer over the node inputs.

	ImageNet	CUB	Places365	SUN397	mean
semantic distances + text embedding	**0.259**	**0.480**	**0.407**	**0.312**	**0.365**
semantic distances only	0.033	0.265	0.249	0.192	0.184
text embedding only	0.213	0.421	0.285	0.217	0.284

Table 7: Loss ablation for ZAGT. Spearman between predicted and true class accuracy on the frozen subsets, averaged over the evaluated subset sizes and over three class splits. The mixture model trained on all four datasets.

	ImageNet	CUB	Places365	SUN397	mean
$L2$ with attenuation	**0.259**	**0.480**	**0.407**	0.312	**0.365**
$L2$	0.169	0.316	0.303	0.257	0.261
$L1$ with attenuation	0.242	0.443	0.370	**0.317**	0.343
$L1$	0.212	0.406	0.337	0.250	0.301

still provide the frequency information to the model. Providing the text embeddings always seems to perform better in this test. Secondly, using our graph-like structuring increases the performance over just using text embeddings and letting the model find the structure itself.

Finally, in Table 7 we try L1 vs L2 and usage and absence of loss attenuation. On most of the datasets, L2+loss attenuation obtained the best or comparable to best results. Loss attenuation always increases the performance regarless the selection of L1 or L2 loss.