Do Multimodal Modals Show Evidence of Semantic Color Discriminability? Final Report

Laura Roettges   Ryan Moreno

1. Introduction

1.1. Motivation

Humans associate meanings with colors based on their sensory experiences, which we will refer to as **color-concept associations**. In data visualization, it is imperative to represent categorical data using colors that align with people's natural expectations or associations—colors that are **semantically interpretable**. Doing so reduces the cognitive effort of the viewer and can prevent confusion, improving communication effectiveness. Gathering color-concept association scores from human trials is costly, especially when scaling to the many concepts that require categorical color encodings for effective data visualizations. This demonstrates the need for automated methods to optimize color palette design for semantic interpretablity.

1.2. Literature Review

Recent studies have explored automated approaches to estimate color-concept associations, reducing reliance on expensive human trials. In the state-of-the-art study, Rathore et al. [9] generates color-concept association scores by analyzing the co-occurrence of specific colors based off images scraped from Google, augmented with clip-art. While effective, this approach requires supervised training utilizing human-annotated data to regress over expert-defined image features in order to hone in on meaningful image regions. Later, Ruizhen Hu et al. [2] explored a self-supervised method that generates color distributions from curated natural images via image colorization techniques. While this colorization method eliminates the need for ground truth color association scores, tailored image selection by concept remains necessary. Most recently, K. Mutherjee et al. [5] utilized GPT-4, a pretrained Large Language Model (LLM), curating tailored prompts to predict color-concept associations based on colors within the UW-71 color library. This method bypasses the need for supervision, image preselection, and ground truth association scores. This method resulted in comparable and often improved results when comparing predicted vs ground truth color association distributions using the Pearson Correlation Coefficient (PCC).

1.3. Goals

There has been limited exploration of multimodal models for the purpose of automating color concept association distributions. Multimodal models such as CLIP [8] integrate both imagery and language using contrastive learning, computing image and text embeddings in a shared latent space. Given that color associations stem from both visual experiences and related linguistic concepts, multimodal models may be well-suited to extracting meaningful color relationships.

Here we evaluate how CLIP and a model variant perform on deriving color concept associations in a zero shot setting. We compare the results with human trial data used in prior studies [9]. Similar to K. Mutherjee et al. [5]'s approach using GPT-4, this approach sidesteps training on ground truth data, offering an efficient and unsupervised approach to this challenge.

2. Method

2.1. Data

We utilize the full UW-71 set of colors as described in the methodology of Mutherjee et al. [5]. From these colors we convert the CIELAB coordinates to RGB image space and generate 71 uniformly colored 224x224 pixel images. We were provided a sample of 20 ground-truth human color concept associations based on the human trial conducted in [5] for the categories "Activities, Fruits2, Vegetables, and Properties," as denoted in Table 1 from Mutherjee et al. [5]. These values were provided directly from K. Mutherjee and K. Schloss from the Schloss Visual Reasoning Lab.

2.2. Zero Shot Framework

Our Zero Shot pipeline is visualized in Figure 1 and adapted from [8]. We experiment with two models

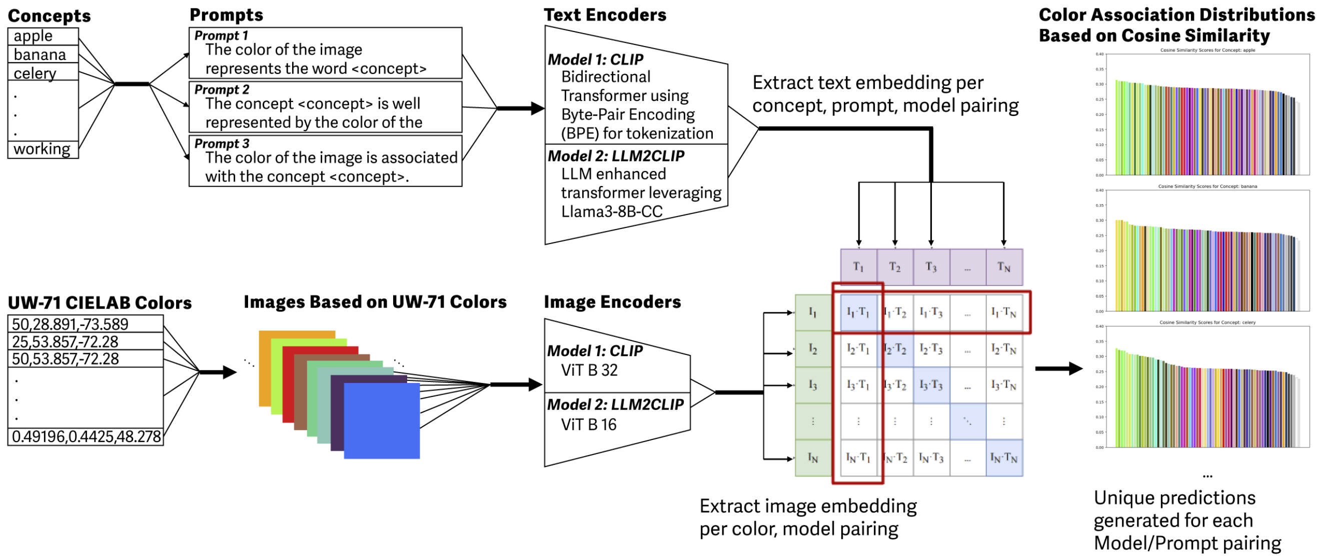

Concepts
apple
banana
celery
working
Prompts
Prompt 1
The color of the image represents the word <concept>
Prompt 2
The concept <concept> is well represented by the color of the
Prompt 3
The color of the image is associated with the concept <concept>.
Text Encoders
Model 1: CLIP
Bidirectional Transformer using Byte-Pair Encoding (BPE) for tokenization
Model 2: LLM2CLIP
LLM enhanced transformer leveraging Llama3-8B-CC
Extract text embedding per concept, prompt, model pairing
Color Association Distributions Based on Cosine Similarity
Cosine Similarity Scores for Concept: apple
Cosine Similarity Scores for Concept: banana
Cosine Similarity Scores for Concept: celery
UW-71 CIELAB Colors
50,28.891,-73.589
25,53.857,-72.28
50,53.857,-72.28
0.49196,0.4425,48.278
Images Based on UW-71 Colors
Image Encoders
Model 1: CLIP
ViT B 32
Model 2: LLM2CLIP
ViT B 16
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$
Extract image embedding per color, model pairing
Unique predictions generated for each Model/Prompt pairing

Figure 1. Zero shot pipeline: We generate a color association distribution for each concept. This is done for two models and three styles of text prompts.

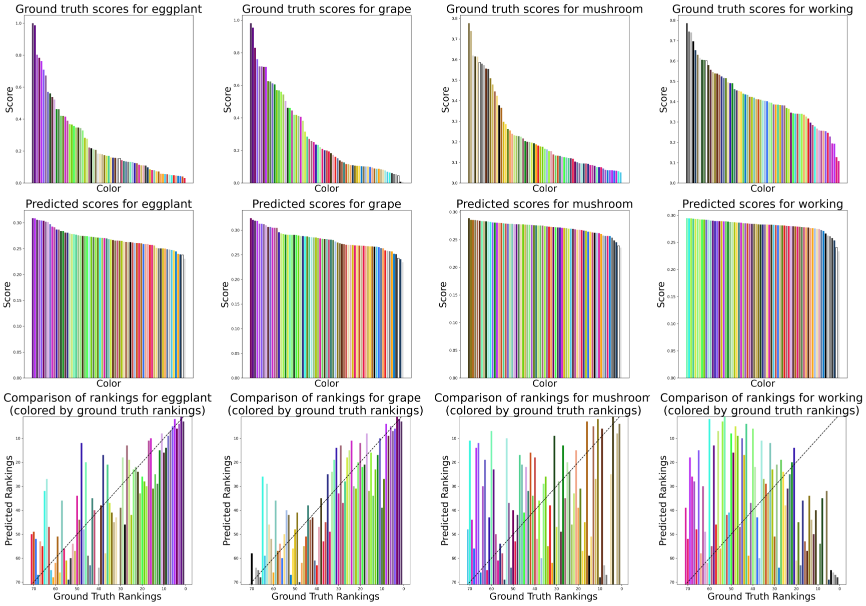


Figure 2. Sampling of ground truth distributions, predicted distribution, and rank correlations between the predicted and ground truth distributions generated using the CLIP model with prompt 1. Here we display two strongly performing concepts, grape and eggplant, followed by two under-performing concepts, mushroom and working. For the bottom row of plots, if the ranks aligned exactly, then the bars would follow the dashed line. For visuals of all predicted and ground truth distributions as well as rank alignment visualizations, see https://github.com/ryan-moreno/BMI_771_project/tree/main/output/evaluation.

and three styles of prompts, generating unique color concept association distributions for each concept per unique prompt-model pair. We refer to the two models we use as **CLIP** and **LLM2CLIP**.

We generate three unique text prompts referencing the concept of interest. The prompts are:

- **Prompt 1:** The color of the image represents the word *'concept'*.
- **Prompt 2:** The concept *'concept'* is well represented by the color of the image.
- **Prompt 3:** The color of the image is associated with the concept *'concept'*.

The text of a prompt is processed by the model's text encoder, producing a corresponding text embedding. CLIP utilizes a transformer with bidirectional attention, leveraging both left-to-right and right-to-left context for textual understanding. It employs Byte-Pair Encoding (BPE) for tokenization [8]. LLM2CLIP extends the transformer of CLIP by leveraging the LLama3-8B-CC LLM, which utilizes caption contrastive fine tuning to allow for extended token inputs (longer prompts) and deeper textual understanding [3]. Concurrently, the color blocks are fed into CLIP's pretrained vision transformer, accessible via

Huggingface [6–8]. This enables retrieval of the image's latent embedding.

- For the CLIP model, we use the ViT B/32 variant [7] which tokenizes images into patches of size 32x32 pixels
- For the LLM2CLIP model, we use the ViT B/16 variant [6], which tokenizes images into smaller 16×16 pixel patches for finer granularity.

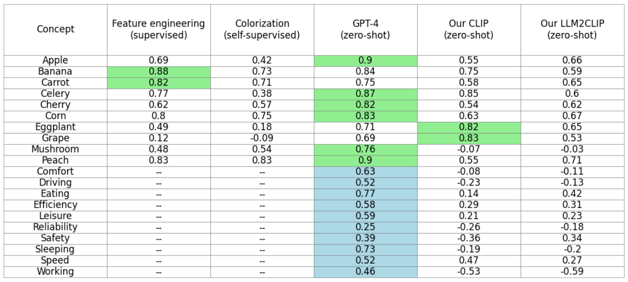
Using the normalized text and image embeddings within the shared latent space, we compute the color-concept association score as the cosine similarity between the image embedding I_j (j from 0 to the total number of colors) and the text prompt T_k (k from 0 to the total number of concepts). Cosine similarity is calculated as $\text{cosine\_similarity}(\mathbf{I_j}, \mathbf{T_k}) = \frac{\mathbf{I_j}\cdot\mathbf{T_k}}{\|\mathbf{I_j}\|\|\mathbf{T_k}\|}$

These similarity scores are grouped by concept in order to generate the color-concept association distribution over the UW-71 colors.

2.3. Evaluation

We used various evaluation metrics to compare an estimated color-concept association distribution $\hat{p} = \{\hat{p}_i\}_{i=1}^N$ to the ground truth user rating $r = \{r_i\}_{i=1}^N$, defined on the color library $c_i{}_{i=1}^N$, where N is the number of colors (71). We evaluated our results using Pearson Correlation [2, 4, 5, 9], Spearman Rank Correlation, Total variation [2], Earth Movers Distance [2], and Entropy Distance [2]. Results are shown in Figure 4. Pearson Correlation is our primary evaluation metric, as it was used across all previous studies, enabling comparison across previous methods.

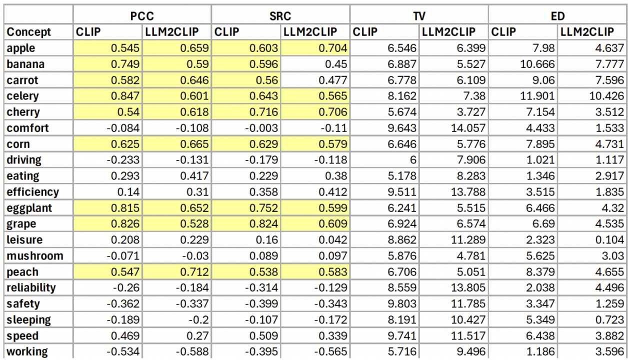
3. Results



Concept	Feature engineering (supervised)	Colorization (self-supervised)	GPT-4 (zero-shot)	Our CLIP (zero-shot)	Our LLM2CLIP (zero-shot)
Apple	0.69	0.42	0.9	0.55	0.66
Banana	0.88	0.73	0.84	0.75	0.59
Carrot	0.82	0.71	0.75	0.58	0.65
Celery	0.77	0.38	0.87	0.85	0.6
Cherry	0.62	0.57	0.82	0.54	0.62
Corn	0.8	0.75	0.83	0.63	0.67
Eggplant	0.49	0.18	0.71	0.82	0.65
Grape	0.12	-0.09	0.69	0.83	0.53
Mushroom	0.48	0.54	0.76	-0.07	-0.03
Peach	0.83	0.83	0.9	0.55	0.71
Comfort	--	--	0.63	-0.08	-0.11
Driving	--	--	0.52	-0.23	-0.13
Eating	--	--	0.77	0.14	0.42
Efficiency	--	--	0.58	0.29	0.31
Leisure	--	--	0.59	0.21	0.23
Reliability	--	--	0.25	-0.26	-0.18
Safety	--	--	0.39	-0.36	0.34
Sleeping	--	--	0.73	-0.19	-0.2
Speed	--	--	0.52	0.47	0.27
Working	--	--	0.46	-0.53	-0.59

Figure 3. Pearson Correlation Coefficients between predicted association scores and ground truth ratings, compared across different techniques: feature engineering [9], colorization [2], GPT-4 [5], and our CLIP and LLM2CLIP studies.

Samplings of resulting distributions and visualizations of rank correlations are provided in Figure 2, which seek to demonstrate our mixed success in aligning with ground truth results. As apparent in Figure 3, our results across both models are comparable to the feature engineering [9] and colorization approaches [2], but underperform against the GPT-4 approach [5]. We achieved improved performance for the grape and eggplant categories across all previous studies. Note that results for our studies depicted in Figures 3, 4, and 2 used prompt 1 because this prompt exhibited the highest average Pearson correlation for experiments across both models, with an average of 0.272 for the CLIP model and 0.266 for the LLM2CLIP model. Generally, the prompts perfromed similarly, as depicted in Appendix A Figure 7. However, for the LLM2CLIP model, prompt 3 led to a significant improvement for the 'mushroom' concept, which had been performing poorly across the other experiments. This improvement may indicate that further prompt tuning on LLM2CLIP could yield improved color-concept distributions.



	PCC		SRC		TV		ED	
Concept	CLIP	LLM2CLIP	CLIP	LLM2CLIP	CLIP	LLM2CLIP	CLIP	LLM2CLIP
apple	0.545	0.659	0.603	0.704	6.546	6.399	7.98	4.637
banana	0.749	0.59	0.596	0.45	6.887	5.527	10.666	7.777
carrot	0.582	0.646	0.56	0.477	6.778	6.109	9.06	7.596
celery	0.847	0.601	0.643	0.565	8.162	7.38	11.901	10.426
cherry	0.54	0.618	0.716	0.706	5.674	3.727	7.154	3.512
comfort	-0.084	-0.108	-0.003	-0.11	9.643	14.057	4.433	1.533
corn	0.625	0.665	0.629	0.579	6.646	5.776	7.895	4.731
driving	-0.233	-0.131	-0.179	-0.118	6	7.906	1.021	1.117
eating	0.293	0.417	0.229	0.38	5.178	8.283	1.346	2.917
efficiency	0.14	0.31	0.358	0.412	9.511	13.788	3.515	1.835
eggplant	0.815	0.652	0.752	0.599	6.241	5.515	6.466	4.32
grape	0.826	0.528	0.824	0.609	6.924	6.574	6.69	4.535
leisure	0.208	0.229	0.16	0.042	8.862	11.289	2.323	0.104
mushroom	-0.071	-0.03	0.089	0.097	5.876	4.781	5.625	3.03
peach	0.547	0.712	0.538	0.583	6.706	5.051	8.379	4.655
reliability	-0.26	-0.184	-0.314	-0.129	8.559	13.805	2.038	4.496
safety	-0.362	-0.337	-0.399	-0.343	9.803	11.785	3.347	1.259
sleeping	-0.189	-0.2	-0.107	-0.172	8.191	10.427	5.349	0.723
speed	0.469	0.27	0.509	0.339	9.741	11.517	6.438	3.882
working	-0.534	-0.588	-0.395	-0.565	5.716	9.496	1.186	3.596

Figure 4. All metrics based on prompt 1, with correlation coefficients over 0.5 highlighted.

Beyond Pearson correlation, the high entropy distances visible in Figure 4 suggest our model's predicted association scores lack *specificity* or "peakiness," the degree to which a concept is more strongly associated with some colors than others, an important aspect of color associations in the context of semantic discernibility [2, 4]. Our models are much flatter than the ground truth distribution scores; the largest cosine similarity score range within a single concept was 0.1003 for CLIP and 0.262 for LLM2CLIP. Even so, our models seem to generate consistent color ordering.

3.1. Review of Underperforming Concepts

We have considered several reasons for why our models underperformed on certain concepts.

Underperforming concepts are defined as the following: 'comfort', 'driving', 'mushroom', 'reliability', 'safety', 'sleeping', 'working', 'efficiency', 'leisure', and 'speed', which had PCC values of 0.2 or

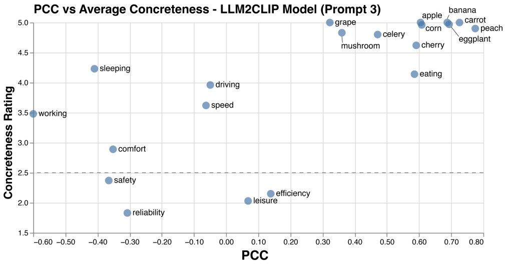


Figure 5. Concept concreteness and model performance. This plots the Pearson correlation of our model's predicted color distribution for a concept against the concept's average "concreteness." Average concreteness is based on the human trial study in[1], where participants rank concept concreteness on a 0-5 scale from most to least concrete. This exhibits a Pearson correlation of 0.712 with a p-value=0.0004 indicating a fairly strong relationship. This experiment was performed using prompt 3 and the LLM2CLIP model.

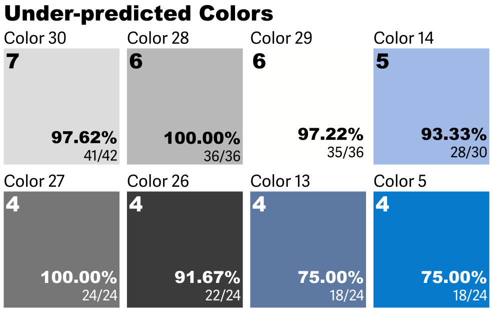


Figure 6. For underperforming concepts, we identified the top 17 colors in their ground truth distributions (approximately the top quartile of color-concept associations). For these colors, we measured how often they were incorrectly predicted in the bottom 17 (bottom quartile) instead of the top 17, highlighting instances of severe underestimation. In this plot, each color patch includes the following. Top-left: The number of the 10 underperforming concepts having this as a top 17 color according to ground truth. Bottom-right percentage: Frequency of severe underestimation across all experiments, models, and prompts. Bottom-right ratio: Count of severe underpredictions divided by the number of times the color is in the top quartile.

less. Most of these concepts are more abstract in nature and were also difficult for other studies [2, 5, 9], though they exhibited much better success in the GPT-4 study [5]. We evaluated the concreteness of these underperforming concepts as considered by Mutherjee et al. [5]. We extracted the mean concreteness ratings from Brysbaert et al. [1] and reviewed the correlation between the concreteness of the concept and the concept's color-association distribution PCC performance. Results are depicted in Figure 5 and demonstrate a positive correlation.

The concreteness hypothesis does not explain the underperformance on concepts such as 'mushroom.' Therefore, we also explored whether bias was leading to under-predicted colors. During an initial visual review of our results, we noticed that grays and white where often predicted among the least associated colors. We performed a more formal review of which colors were under-estimated for our underperforming concepts, as seen in Figure 6. We believe these colors could indicate a bias against common background colors (skys, ground, walls, etc). We also noted possible bias when reviewing the results of the "apple" concept. The ground truth color distribution has red as the highest associated color, followed by green, while our model predicted green followed by red. Since apples come in both red and green varieties, this could indicate an over inclusion of green apples in the CLIP training dataset. Investigating this would require a more thorough review of the CLIP training dataset.

4. Conclusion

Utilizing multimodal modals like CLIP to derive color-concept associations in a zero shot setting has the advantage of being unsupervised, but is outperformed by the most recent GPT-4 approach [5]. LLM2CLIP and CLIP models performed comparably; while CLIP achieved some of the highest Pearson correlations, the LLM2CLIP model outperformed the CLIP model on 35 out of 60 tests and showed lower average entropy distances. More extensive prompt exploration, especially with LLM2CLIP, could lead to further improvements to the predicted association distributions. Additionally, it would be valuable to substantiate our approach by testing our model within the context of color selection based on distribution differences, as described in [4]. Our code, results, analysis, and supplemental data visualizations can be found at `https://github.com/ryan-moreno/BMI_771_project/`.

References

[1] Marc Brysbaert, Amy B. Warriner, and Victor Kuperman. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46(3):904–911, 2014. 4

[2] Rui Hu, Ziqi Ye, Bin Chen, Oliver van Kaick, and Hui Huang. Self-supervised color-concept association via image colorization. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):247–256, 2023. 1, 3, 4

[3] Weiquan Huang, Aoqi Wu, Yifan Yang, Xufang Luo, Yuqing Yang, Liang Hu, Qi Dai, Xiyang Dai, Dongdong Chen, Chong Luo, and Lili Qiu. Llm2clip: Pow-

erful language model unlock richer visual representation, 2024. 2

[4] Kushin Mukherjee, Brian Yin, Brianne E. Sherman, Laurent Lessard, and Karen B. Schloss. Context matters: A theory of semantic discriminability for perceptual encoding systems. *CoRR*, abs/2108.03685, 2021. 3, 4

[5] Kushin Mukherjee, Timothy T. Rogers, and Karen B. Schloss. Large language models estimate fine-grained human color-concept associations, 2024. 1, 3, 4

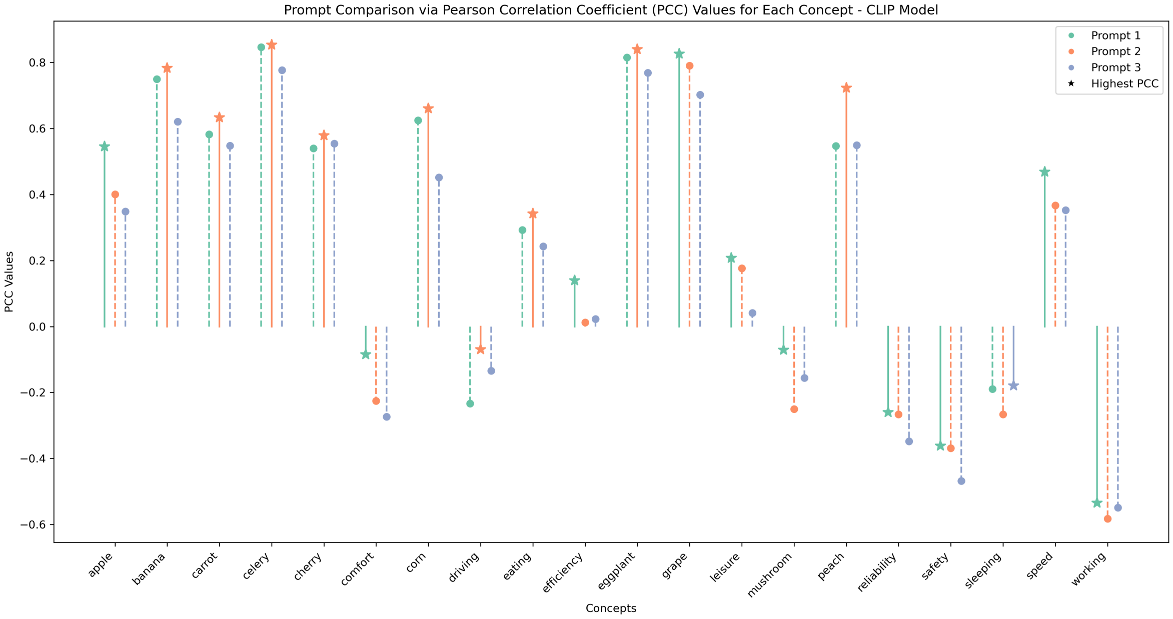
[6] OpenAI. CLIP - Vision Transformers (ViT-B/16), 2021. Accessed: [12/08/2024]. 3

[7] OpenAI. CLIP - Vision Transformers (ViT-B/32), 2021. Accessed: [12/08/2024]. 3

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1, 2, 3

[9] Ragini Rathore, Zachary Leggon, Laurent Lessard, and Karen B. Schloss. Estimating color-concept associations from image statistics. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1226–1235, 2019. https://schlosslab.discovery.wisc.edu/wp-content/uploads/2019/08/RathoreLeggonLessardSchlossinPress.pdf. 1, 3, 4

Appendix A



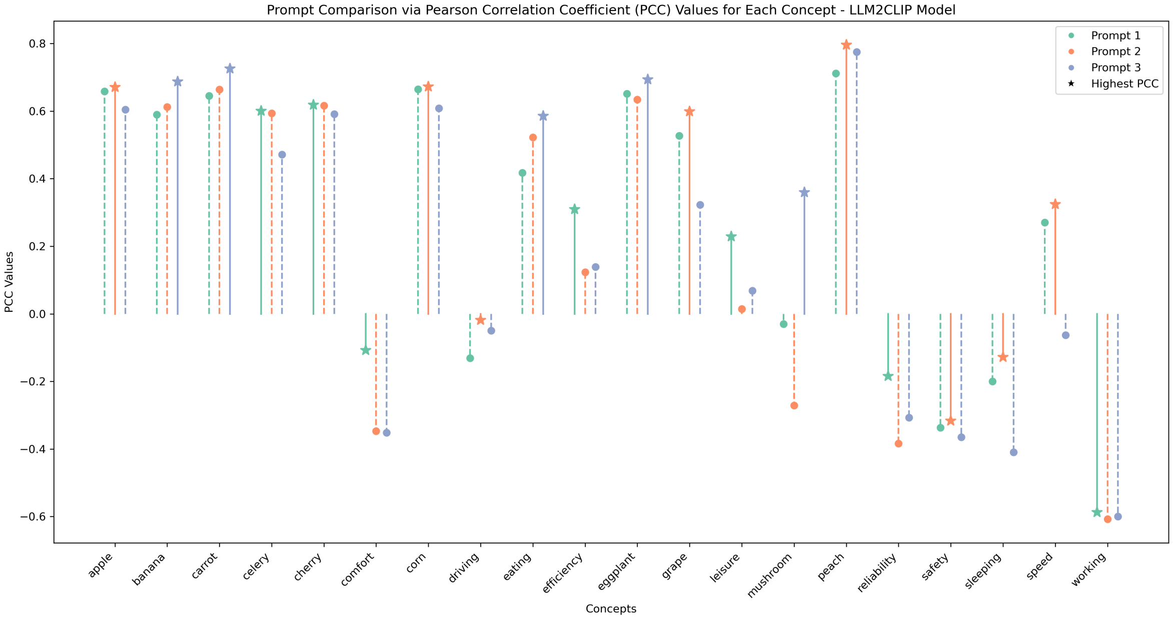


Figure 7. Pearson Correlation Coefficients (PCC) for each concept across three prompts. The top plot shows results using the CLIP model while the bottom plot shows results using the LLM2CLIP model.