Appendix to AlignCLIP: Navigating the Misalignments for Robust Vision-Language Generalization

Zhongyi Han[1,2*], Gongxu Luo[2], Hao Sun[3], Yaqian Li[4], Bo Han[5], Mingming Gong[6,2], Kun Zhang[7,2], Tongliang Liu[8,2*]

[1]CEMSE, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia.
[2]Department of Machine Learning, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates.
[3]School of Software, Shandong University, Jinan, China.
[4]OPPO Research Institute, OPPO, Shanghai, China.
[5]Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong.
[6]School of Mathematics and Statistics, The University of Melbourne, Melbourne, Australia.
[7]Department of Philosophy, Carnegie Mellon University, Pittsburgh, USA.
[8]School of Computer Science, The University of Sydney, Sydney, Australia.

*Corresponding author(s). E-mail(s): zhongyi.han@kaust.edu.sa; tongliang.liu@sydney.edu.au;

Appendix

Similarity Distribution and the Random Variable

In our paper, the term "similarity distribution" $\Pi(\mathbf{x}) = [\Pi(\mathbf{x}_1), \Pi(\mathbf{x}_2), \ldots, \Pi(\mathbf{x}_M)]$ refers to a probability distribution over the image tokens (patches) in an image $\mathbf{x}$. **The random variable here is the image token index i,** where each $\Pi(\mathbf{x}_i)$ represents the normalized similarity between the i-th image token embedding $\mathbf{I}_{\mathbf{x}}^i$ and the text embedding $\mathbf{T}_y$:

$$\Pi(\mathbf{x}_i) = \frac{\exp\left(S\left(\mathbf{I}_{\mathbf{x}}^i, \mathbf{T}_y\right)/\tau\right)}{\sum_{m=1}^{M} \exp\left(S\left(\mathbf{I}_{\mathbf{x}}^m, \mathbf{T}_y\right)/\tau\right)},$$

where $S(\cdot,\cdot)$ denotes the similarity function (e.g., cosine similarity), τ is a temperature parameter, and M is the number of image tokens excluding the class token. This distribution sums up to 1 ($\sum_{i=1}^{M} \Pi(\mathbf{x}_i) = 1$) and indicates how much each token contributes to the overall similarity between the image and the text. Essentially, $\Pi(\mathbf{x})$ captures the model's attention over the different parts of the image in relation to the text.

How Modifying Q, K, and V Affects Attention Focus

In the multi-head attention mechanism, the attention weights are calculated based on the interactions between the query (Q), key (K), and value (V) vectors. Specifically, the attention is computed as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V,$$

where Q, K, and V are linear transformations of the input embeddings, and d_k is the dimensionality of the key vectors. The attention weights are derived from the scaled dot-product between Q and K, and these weights are then applied to the V vectors to produce the output.

In the standard setting, the Q and K vectors are responsible for computing the attention scores between different tokens. The way Q and K are constructed influences which tokens attend to which other tokens.

Two seminal studies [1, 2] have observed the the opposite visualization in CLIP. Specifically, opposite visualization refers to the visualization results that are opposite to the ground-truth. They discovered that the major cause of this issue is the parameters (query and key) in the self-attention module, which heavily focus on opposite semantic regions.

In this paper, we present more clear evidence through empirical analysis as shown in Figure 1. This figure illustrates the tendency of CLIP models to disproportionately focus on background elements. This figure demonstrates that, despite various modifications to the original attention structures (from the third row to the last), the model struggles to correct the attention distributions effectively. These modifications, although diverse, are unable to bring about the desired alignment in attention distributions, highlighting the resilience of the misalignments present in the original model. We also found that even fine-tuning the image encoder in CLIP on downstream tasks, depicted in the second row, does not yield substantial alterations or improvements in the attention distribution. This underscores the limitations of fine-tuning as a strategy for addressing attention misalignments in such models, suggesting the need for more robust solutions to mitigate these inherent issues.

To elaborate on the modifications, the Q-Q-Q structure represents a scenario where both K and V vectors are replaced by Q, altering the dynamics of the attention mechanism. Similarly, the Q-K-K structure implies a condition where the V vector is

substituted by the K vector. These specific alterations, along with others not explicitly mentioned, follow a consistent rule of replacement, aiming to explore the potential of structural modifications in addressing the attention alignment problem. However, as the figure illustrates, these adjustments, while varied, do not succeed in resolving the attention alignment issues inherent to CLIP, indicating a pressing need for more innovative and effective approaches to overcome these challenges in attention-based models.

Causal Analysis of Attention Misalignment in CLIP Models

To understand why CLIP models tend to focus on background elements rather than salient foreground objects, we performed a causal analysis of the training data and process. CLIP is trained on a large dataset of image-text pairs, where images often act as the *effects* (Y) of *causes* (X), represented by the text descriptions. These image-text pairs frequently capture not only the main subjects (foreground objects) but also background elements and contextual information. During training, CLIP learns to associate images with their corresponding text descriptions, creating a potential bias.

From a causal perspective, images represent the *effects* of the *causes* (foreground objects) described in the accompanying text. However, because the text often captures both the foreground objects and background context, the model may inadvertently learn spurious correlations between background elements and class labels. This leads to two key issues:

- **Spurious Correlations**: CLIP may learn to associate background elements (e.g., sand in a beach scene) with certain labels, creating a reliance on these background elements for classification.
- **Overemphasis on Background**: Since background elements are prevalent across many images, the attention mechanism may prioritize these elements over the more relevant foreground objects.

This issue can be particularly problematic when CLIP is deployed in open-world settings, where background contexts may differ significantly from the training data. The bias in Q and K vectors exacerbates this issue, leading to attention maps that disproportionately focus on background elements rather than foreground entities.

Effect of Setting $Q = K$

By setting the query (Q) and key (K) vectors to be identical (e.g., employing a K-K-V or V-V-V attention structure), the attention scores effectively become based on **self-similarities**. This means that the attention weights are computed based on the similarity of each token to itself or to tokens with the same representation. Such a modification reduces the influence of spurious correlations learned by the original Q and K vectors, which may have been biased toward background elements due to the pre-training data. As a result, this adjustment promotes a more localized and coherent attention distribution, naturally emphasizing salient foreground objects.

Our experiments, illustrated in Figure 2, demonstrate that modifying the attention mechanism in this manner leads to attention maps that are more focused on the salient

foreground objects. The figure shows that synchronizing the Q and K vectors while maintaining the V vector in its original role results in more coherent similarity and attention distributions. This adjustment effectively addresses the inherent discrepancies observed in the original attention configurations, where the model disproportionately focused on background elements.

Motivated by this observation, we devised a specialized diagonal matrix $\mathbf{D}$ to replace the product of Q and K in the attention mechanism, leading to the $\mathbf{D}$-V attention structure:

$$\text{Attention}(\mathbf{D}, V) = \mathbf{D}V.$$

Here, $\mathbf{D}$ is a diagonal matrix that scales each token's own V vector, emphasizing the token's self-information and reducing the influence of irrelevant tokens. The $\mathbf{D}$-V structure not only preserves the structural and functional integrity of the attention mechanism but also significantly enhances computational efficiency. By mitigating the computational demands associated with calculating and normalizing the $QK^\top$ matrix, this method facilitates more streamlined processing, enabling the model to handle larger datasets and more complex computations with ease. Specifically, we found that the accuracy performance between the $\mathbf{D}$-V structure and the V-V-V structure is comparable; however, the $\mathbf{D}$-V structure notably reduces computational complexity.

Foreground objects in images typically have more coherent and distinctive features compared to background elements. When the attention mechanism relies on self-similarity (as in K-K-V or V-V-V), tokens belonging to the same salient object (foreground) reinforce each other's importance due to higher mutual similarities. This leads to stronger attention weights for these tokens. In contrast, background tokens, which are often more varied and less coherent, have lower mutual similarities and thus receive lower attention weights. Consequently, the attention mechanism naturally emphasizes the foreground objects, improving the model's focus on relevant features and enhancing overall performance.

By adopting the $\mathbf{D}$-V structure, we effectively address the attention misalignment issue in CLIP models. This adjustment aligns the attention mechanism to prioritize salient foreground entities rather than background elements, thereby enhancing the model's generalization capabilities in open-world settings.

Generation of the Expected Similarity Distribution $\Pi^*(\mathbf{x})$

To obtain the expected similarity distribution $\Pi^*(\mathbf{x})$, we utilize the modified attention mechanism with the $\mathbf{D}$-V structure. This approach focuses on each token's intrinsic features by simplifying the attention computation and reducing the influence of potentially biased interactions encoded in the original Q and K vectors.

Specifically, in the $\mathbf{D}$-V attention mechanism, we compute the new image token embeddings $\mathbf{I}_\mathbf{x}^{i*}$ as follows:

$$\mathbf{I}_\mathbf{x}^{i*} = \text{Attention}(\mathbf{D}, V_i) = \mathbf{D}_{ii}V_i,$$

where:

- V_i is the value vector of the i-th image token.

- $\mathbf{D}_{ii}$ is the i-th diagonal element of the diagonal matrix $\mathbf{D}$, which acts as a scaling factor for the i-th token.

Since $\mathbf{D}$ is a diagonal matrix, this operation effectively scales each token's value vector individually, emphasizing its self-information without considering interactions with other tokens.

Next, we compute the expected similarity distribution $\Pi^*(\mathbf{x})$ by calculating the similarity between each modified image token embedding $\mathbf{I}_{\mathbf{x}}^{i*}$ and the text embedding $\mathbf{T}_y$, and then normalizing these similarities:

$$\Pi^*(\mathbf{x}_i) = \frac{\exp\left(S\left(\mathbf{I}_{\mathbf{x}}^{i*}, \mathbf{T}_y\right)/\tau\right)}{\sum_{m=1}^{M} \exp\left(S\left(\mathbf{I}_{\mathbf{x}}^{m*}, \mathbf{T}_y\right)/\tau\right)},$$

where:

- $S(\cdot, \cdot)$ denotes the similarity function, such as cosine similarity.
- τ is a temperature parameter that controls the sharpness of the distribution.
- M is the number of image tokens (excluding the class token).

This similarity distribution $\Pi^*(\mathbf{x}) = [\Pi^*(\mathbf{x}_1), \Pi^*(\mathbf{x}_2), \ldots, \Pi^*(\mathbf{x}_M)]$ represents how much each token contributes to the overall similarity between the image and the text when relying solely on the tokens' inherent properties. By focusing on individual token representations, $\Pi^*(\mathbf{x})$ naturally emphasizes tokens corresponding to salient foreground objects, which have features more closely aligned with the textual description.

In summary, the steps to obtain $\Pi^*(\mathbf{x})$ are:

1. **Compute Modified Token Embeddings:** Use the $\mathbf{D}$-V attention mechanism to compute the new token embeddings $\mathbf{I}_{\mathbf{x}}^{i*}$ that emphasize each token's self-information.
2. **Calculate Similarities:** Compute the similarity between each modified token embedding $\mathbf{I}_{\mathbf{x}}^{i*}$ and the text embedding $\mathbf{T}_y$.
3. **Normalize to Form a Distribution:** Apply the softmax function with temperature τ to these similarities to obtain the normalized similarity distribution $\Pi^*(\mathbf{x})$.

By obtaining $\Pi^*(\mathbf{x})$ in this way, we create an expected similarity distribution that reflects a model focusing on the inherent properties of the tokens—thereby highlighting foreground elements—and provides a target for aligning the original similarity distribution $\Pi(\mathbf{x})$ through the Attention Alignment Loss.

References

[1] Li, Y., Wang, H., Duan, Y., Xu, H., Li, X.: Exploring visual interpretability for contrastive language-image pre-training. arXiv (2022)

[2] Li, Y., Wang, H., Duan, Y., Li, X.: Clip surgery for better explainability with enhancement in open-vocabulary tasks. arXiv (2023)

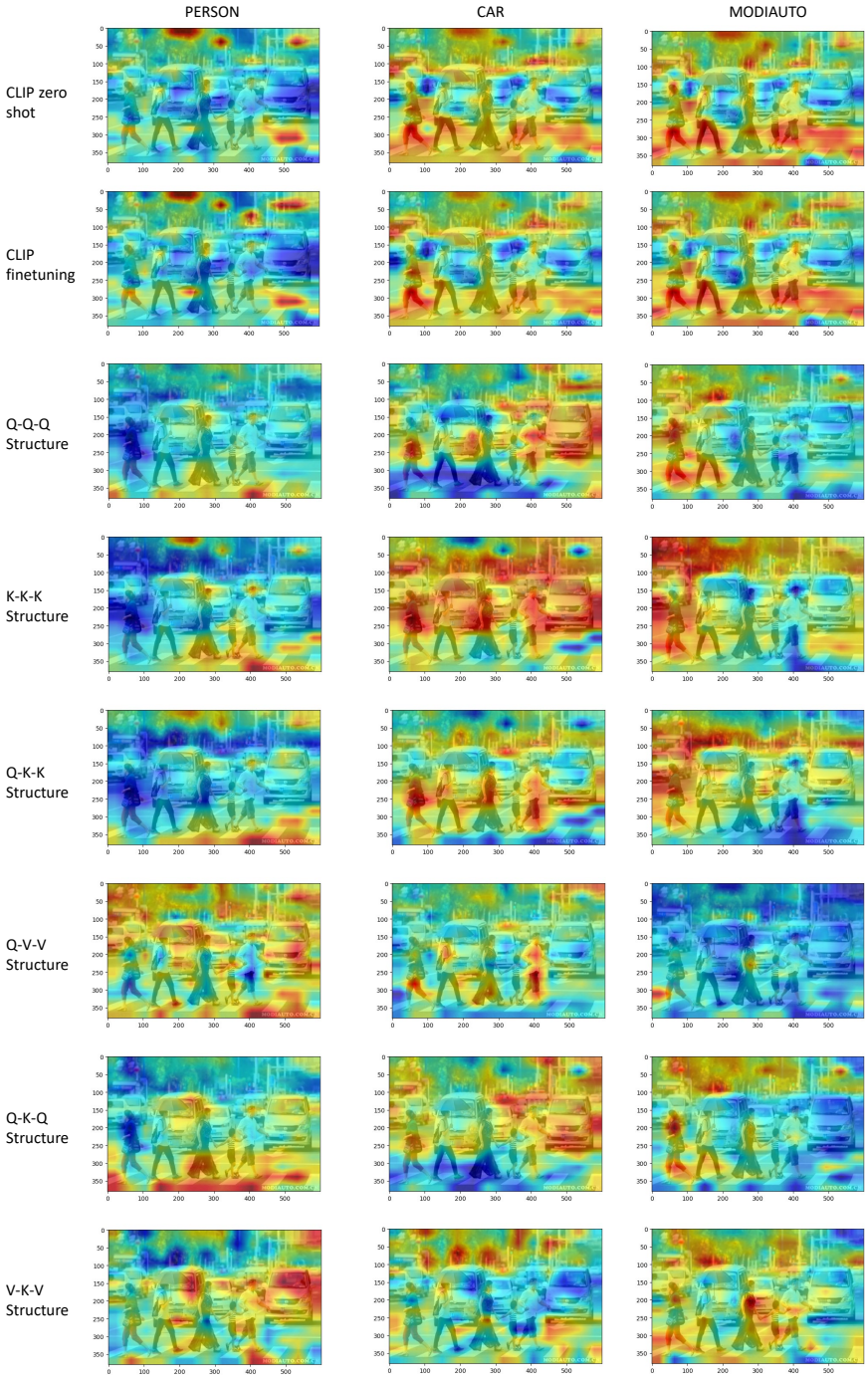

Fig. 1: Illustration depicting the inherent attention alignment issues in CLIP (first row). Subsequent adjustments to the original attention structures, spanning from the third row to the last, are unable to rectify the attention distributions effectively.

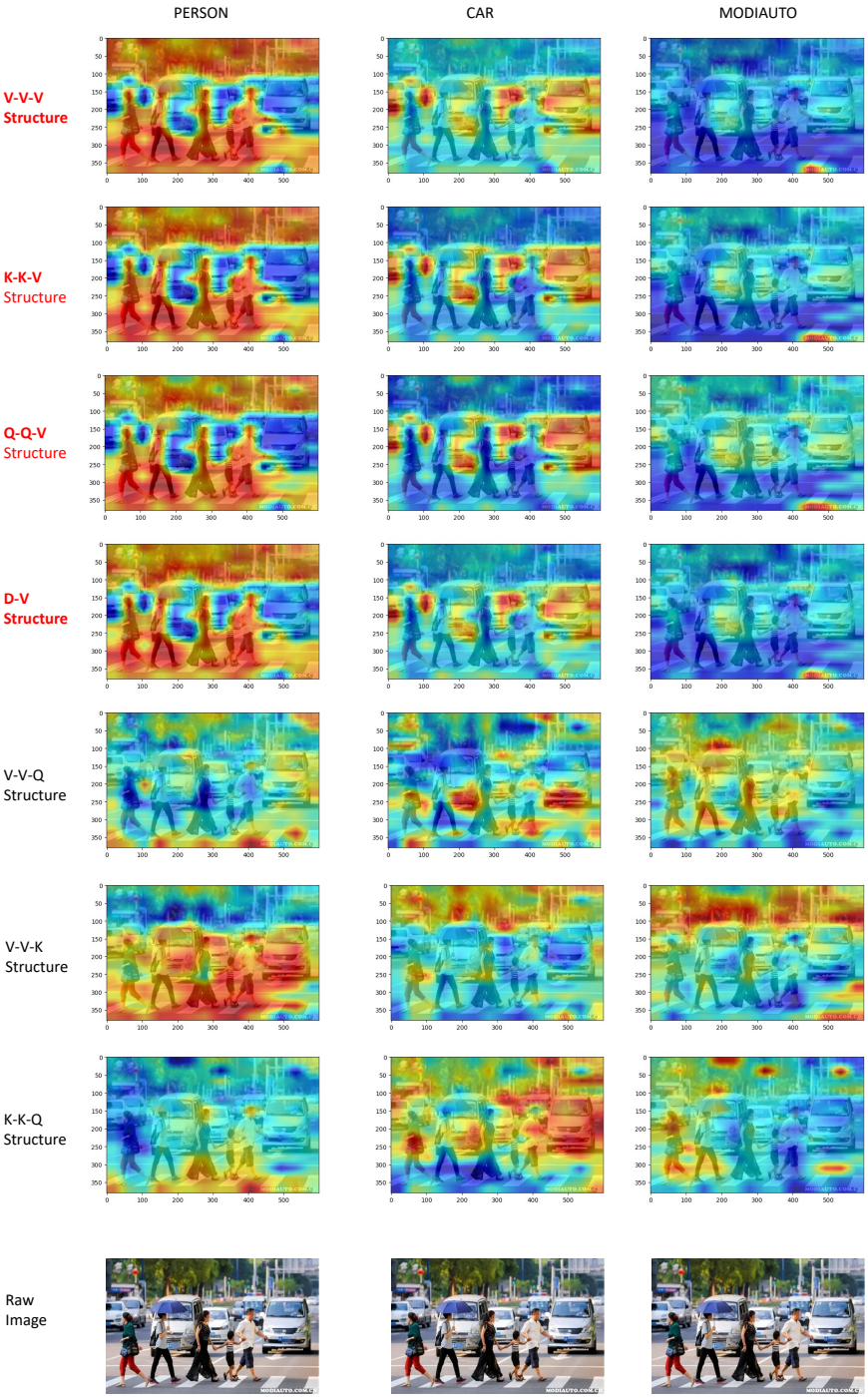

Fig. 2: This figure illustrates the impact of equating the Q and K vectors while retaining the V vector in its original placeholder, a modification that results in more coherent similarity and attention distributions. 'MODIAUTO' refers to the text contained in the images.