# Cross-modal Attention Congruence Regularization for Vision-Language Relation Alignment

**Rohan Pandey    Rulin Shao    Paul Pu Liang**
**Ruslan Salakhutdinov    Louis-Philippe Morency**
Language Technologies Institute and Machine Learning Department
Carnegie Mellon University
{rspandey,rulins}@cs.cmu.edu

## Abstract

Despite recent progress towards scaling up multimodal vision-language models, these models are still known to struggle on compositional generalization benchmarks such as Winoground. We find that a critical component lacking from current vision-language models is relation-level alignment: the ability to match directional semantic relations in text (e.g., 'mug **in** grass') with spatial relationships in the image (e.g., the position of the mug **relative** to the grass). To tackle this problem, we show that relation alignment can be enforced by encouraging the language attention from 'mug' to 'grass' (capturing the semantic relation 'in') to match the visual attention from the mug to the grass (capturing the corresponding physical relation). Tokens and their corresponding objects are softly identified using a weighted mean of cross-modal attention. We prove that this notion of soft cross-modal equivalence is equivalent to enforcing congruence between vision and language attention matrices under a 'change of basis' provided by the cross-modal attention matrix. Intuitively, our approach projects visual attention into the language attention space to calculate its divergence from the actual language attention, and vice versa. We apply our Cross-modal Attention Congruence Regularization (CACR) loss to fine-tune UNITER and improve its Winoground Group score by 5.75 points.

## 1 Introduction

Compositionality is the ability to combine meanings of constituents according to structured rules. Recent work shows that Vision-Language Models (VLMs) fail to construct compositional representations and generally ignore syntactic & structural information [Thrush et al., 2022, Milewski et al., 2022, Liang et al., 2022]. Winoground [Thrush et al., 2022] is a vision-language compositionality task that tests a VLM's ability to match syntactic permutations of text with their visual interpreta-
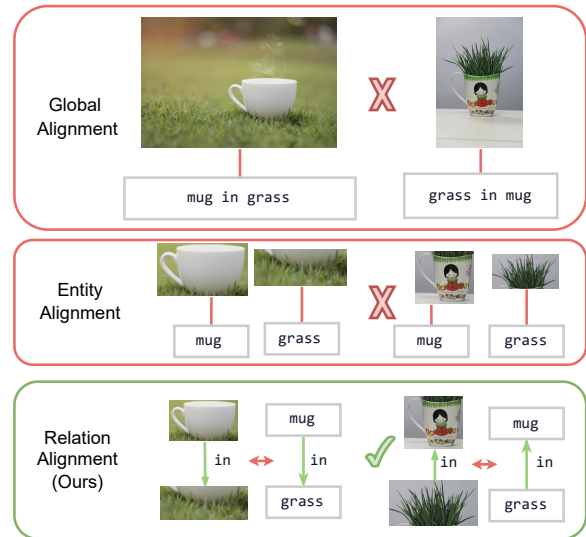


Figure 1: Global Alignment (GA) only aligns the entire image with the corresponding caption. Entity Alignment (EA) extracts entities from the image and caption for finer-grained alignment. Relation Alignment (RA) cross-modally aligns the intra-modal relations between entities in both the image and the text. We show RA is vital to improve compositional performance.

tions, for example correctly matching "grass in mug" and "mug in grass" to their corresponding images. Winoground finds that all recent state-of-the-art VLMs perform below chance levels on this compositionality task. Contemporaneously, Milewski et al. [2022] probe for structural knowledge in VLMs, finding that they encode significantly less linguistic syntax than Language Models (LMs) and virtually no visual structure. Recently, Yuksekgonul et al. [2022] built a large dataset confirming that VLMs treat images as a 'bag of objects' and don't adequately represent visuo-linguistic relations.

Since models must determine whether the compositional structure of an image matches that of the caption, it's important for the model to learn to cross-modally align intra-modal relations. That

5444

is, if the relation from 'mug' to 'grass' is 'in-ness', the model should recognize when the equivalent physical relation holds between a mug and grass in the image, and representationally align these relations such that an image-text matching head may more easily determine whether the relations are cross-modally equivalent. In simpler terms, the compositional structure of input for each modality should be represented such that they can be cross-modally matched even for difficult examples like Winoground.

Unfortunately, there has been less highly influential work on **relation alignment** between vision & language, and Thrush et al. [2022] did not benchmark any such models. In this work, we begin exploration of these relation alignment approaches by tentatively grouping them into 3 categories:

1. Structural Data: training a model on data that explicitly captures relational structure

2. Structural Model: infusing an inductive bias into the architecture of the model that enables more compositional representations

3. Structural Training: modifying the objective function or imposing a parameter constraint to encourage relation alignment

Since Structural Data approaches require complex annotations and Structural Model approaches are often incompatible with large transformers, we identify Structural Training as a promising avenue for providing compositional inductive biases to VLMs due to their architecture-agnostic compatibility and computational scalability.

In this work, we propose a Structural Training approach for relation alignment that uses the cross-modal attention matrix as a change of basis[1] to the opposite modality, which we then compare to the original modality to calculate a divergence loss, effectively measuring cross-modal congruence between intra-modal attentions.

We show how our approach, Cross-modal Attention Congruence Regularization (CACR), generalizes previous Structural Training work on cross-modal attention regularization (IAIS [Ren et al., 2021]) by taking into account all possible entity alignments and computationally simplifying relation alignment. The CACR regularization term can easily be dropped into most transformer-based Vision-Language model objectives with no

---

[1]not defined in a strict linear algebraic sense

added data and minimal computational overhead, to encourage relation alignment during training. Finally, we show that $CACR_{base}$ improves on $IAIS_{base}$—where $IAIS_{large}$ holds the current state-of-the-art on Winoground.

## 2  Related Work

Below, we categorize several relation alignment approaches following the framework in Sec. 1.

1. Structural Data [Wu et al., 2019, Zhang and Peng, 2019, Yu et al., 2021, Cui et al., 2021, Wan et al., 2021, Khan et al., 2022]

2. Structural Model [Andreas et al., 2016, Guo et al., 2019, Hong et al., 2021, Zhang, 2022, Wang et al., 2022b, Kim et al., 2022, Wang et al., 2022a]

3. Structural Training [Ren et al., 2021, Yang et al., 2021a,b, Xue et al., 2021]

While some of these works introduce ideas from multiple of these categories, we group them by their core contribution. For example, ROSITA proposes a graphical data pre-training approach, and a self-supervised objective to accompany it; we consider it a Structural Data approach since the training objective ultimately is just a necessity for the data being provided.

Unfortunately, many of these works do not provide publicly available code or pre-trained checkpoints, so we were unable to complete an exhaustive analysis of the compositional performance of these relation alignment approaches. Due to the added complexity of Structural Model approaches, we leave exploration of their compositional abilities to future work.

Regardless, we chose one exemplar for both Structural Data (ROSITA) and Structural Training (IAIS) that made their pre-trained image-text matching checkpoints available; we generated their scores on Winoground, which have not previously been calculated. In Tab. 1, we present these two relation alignment models' Winoground scores alongside a few entity alignment and global alignment models.

Notice that global alignment approaches tend to perform the lowest on Winoground, even when scaled considerably. Entity alignment approaches perform intermediately and OSCAR+ specifically held the state-of-the-art prior to our benchmarking of these relation alignment models. Of the

| Model | Text | Image | Group |
|-------|------|-------|-------|
| MTurk Human | 89.50 | 88.50 | 85.50 |
| IAIS (RA-ST) | **42.50** | **19.75** | **16.00** |
| OSCAR+ (EA) | 37.75 | 17.75 | 14.50 |
| ROSITA (RA-SD) | 35.25 | 15.25 | 12.25 |
| UNITER (EA) | 38.00 | 14.00 | 10.50 |
| CLIP (GA) | 30.75 | 10.50 | 8.00 |
| LXMERT (GA) | 19.25 | 7.00 | 4.00 |

Table 1: Comparison of Winoground scores for models [Ren et al., 2021, Zhang et al., 2021, Cui et al., 2021, Chen et al., 2020, Radford et al., 2021, Tan and Bansal, 2019] using Global Alignment (GA), Entity Alignment (EA), Relation Alignment with Structural Data (RA-SD), and Relation Alignment with Structural Training (RA-ST). We find that IAIS, a recent relation alignment approach that uses attention regularization for structural training achieves universal performance improvements.

two relation alignment approaches we benchmark, IAIS beats out OSCAR+ and achieves a new state-of-the-art on Winoground. But ROSITA, despite providing structural data to encourage cross-modal relation alignment, underperforms OSCAR+. We attribute this partly to the improved visual features OSCAR+ has access to as a result of VinVL, but further comparison of IAIS and ROSITA is explored in our recent work.

Based on these past results and analysis, we choose to further explore structural training approaches to relation alignment. In other words, our research question becomes: How can we infuse the vision-language model's training objective with an implicit structural prior that encourages cross-modal alignment of relations?

## 3 Cross-modal Attention Congruence Regularization

To attempt a solution to this question, we begin by noting that attention activations encode some degree of relational information. Attention values in transformers may be seen as an informational gating mechanism that implicitly encode how representations are composed [Abnar and Zuidema, 2020]. For example, past work in language has shown how syntax trees may be extracted [Mareček and Rosa, 2019] from attention across layers and used to guide attention [Bai et al., 2021, Li et al., 2020] for improved compositionality. In this section, we extend this intuition to the multimodal domain by proposing to use the cross-modal attentions, which as a change-of-basis matrix encode a

transformation from one modality's compositional structure to the opposite modality's, to encourage cross-modal relation alignment.

### 3.1 Relation Alignment Using Attention

In specific, we focus on the self-attention matrix $S$ computed in a transformer by

$$S = QK^\top = (XW^Q)(XW^K)^\top \qquad (1)$$

Then, some row $i$ in S corresponds to a distribution over columns $j_0, ..., j_n$ where $S_{i,j}$ tells us how much of the previous layer's entity representation $j$ we want to infuse into the current layer's entity representation $i$, intuitively their compositional relation. Since $X$ is a series of visual and linguistic tokens, we can segment $S$ into four submatrices for intra- and cross-modal relations [Bugliarello et al., 2021]. Denote the intra-modal attention submatrices in the last multimodal encoder layer as $S_{VV}$ (vision to vision) and $S_{LL}$ (language to language); the cross-modal attention matrices as $S_{VL}$ (vision to language) and $S_{LV}$ (language to vision).

$$S = \begin{pmatrix} S_{LL} & S_{LV} \\ S_{VL} & S_{VV} \end{pmatrix} \qquad (2)$$

If an image and caption have the same underlying compositional structure, the entities that cross-modally correspond to each other should bear similar intra-modal compositional structure. That is, a word $w$ should attend to other words (in $S_{LL}$) in a similar way that its visual object counterpart $o$ attends to other objects (in $S_{VV}$). Furthermore, we can use the cross-modal matrices ($S_{LV}$ and $S_{VL}$) to identify entities that cross-modally correspond as they will generally attend to each other [Aflalo et al., 2022]. Unfortunately, since representations are heavily contextualized by the final layer, clear bijective correspondences between words and objects may not always be identified using an argmax over the cross-modal attention matrix as Ren et al. [2021] attempts. Deeper analysis of when their model, IAIS, fails to identify cross-modal bijective correspondences is provided in Sec. 5.

### 3.2 Attention Congruence

We opt to use the cross-modal matrices ($S_{LV}$ and $S_{VL}$) as a whole to 'change basis' to the opposite modality, with which we can then calculate 'congruence' with the original modality. However, we use 'change of basis' and 'congruence' loosely since the cross-modal matrices are not guaranteed
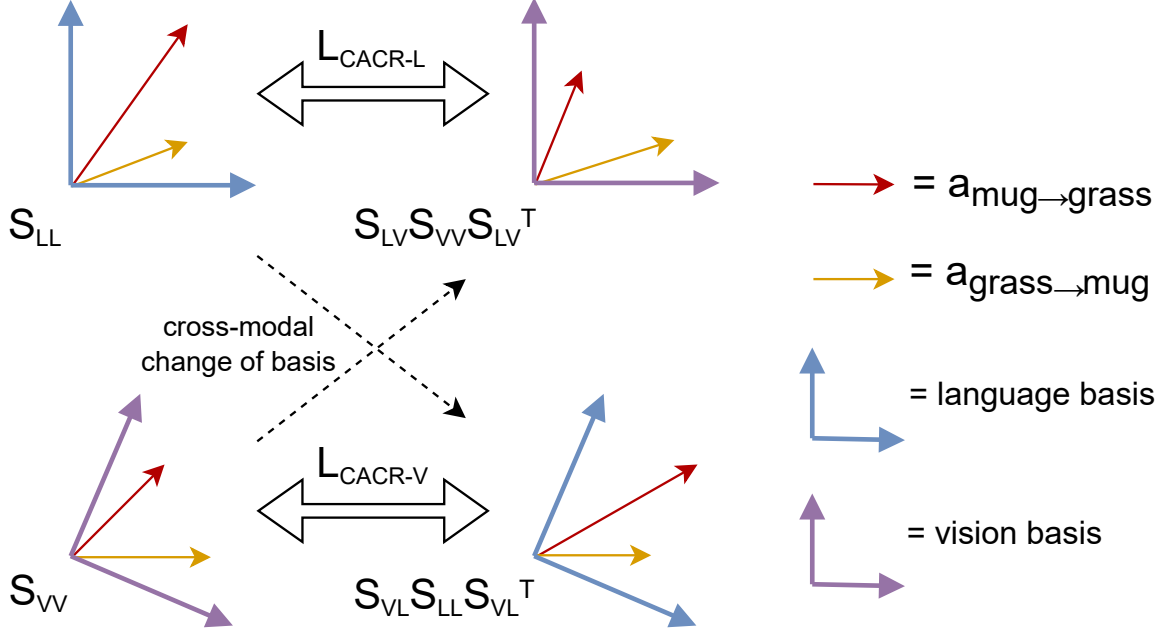
Figure 2: Top: language attention ($S_{LL}$) is aligned with the visual attention projected into the language basis ($S_{LV}S_{VV}S_{LV}^\top$) to calculate $\mathcal{L}_{CACR-L}$; specific attention values (yellow, red) capturing intra-modal relations are cross-modally aligned as a result. Bottom: as above, but in the vision basis.

to be square and thus do not satisfy strict linear algebraic definitions. We formulate $S_{VV}$ in the language basis as $S_{LV}S_{VV}S_{LV}^\top$, which we then encourage to be similar to $S_{LL}$.

Under the hood, this says that for each $a_{i \to j} \in S_{LL}$, we can use row vectors $S_{LV,i}$ and $S_{LV,j}^\top$ to calculate a weighted sum $a_{i \to j}^*$ over $S_{VV}$. If we were to do this for all $i, j$, we would construct a matrix of the same shape as $S_{LL}$ where each entry is $a_{i \to j}^*$, i.e. an approximation of the visual correspondent of the relation $a_{i \to j}$ taking into account all the possible cross-modal alignments of $i$ and $j$. Since this computation intuitively makes a lot of sense and may more easily be compared to previous approaches, we choose to illustrate it in Fig. 3. However, since this computation is relatively expensive, we instead use the $S_{LV}S_{VV}S_{LV}^\top$ formulation which produces the same matrix of $a_{i \to j}^*$ values but with considerably fewer operations. This also enables us to view the operation as a 'change-of-basis' to the opposite modality and the CACR loss as encouraging a sense of cross-modal 'congruence'.

Specifically, we align the original $S_{LL}$ with the language-basis $S_{VV}$ matrix using $\mathcal{L}_{CACR-L}$:

$$\mathcal{L}_{\text{CACR-L}} = \text{m-KL}(\sigma(S_{LV}S_{VV}S_{LV}^\top), \sigma(S_{LL})). \tag{3}$$

We apply a softmax to normalize both matrices since $S_{LV}S_{VV}S_{LV}^\top$ will generally be larger in scale due to summation. Additionally, m-KL($\cdot$) [Ren et al., 2021] is a symmetric matrix-based Kullback-Leibler Divergence (m-KL) which measures the distance between two matrices $S$ and $S'$:

$$\text{m-KL}(S, S') = \sum_i^N \text{KL}(S_i || S_i') + \text{KL}(S_i' || S_i), \tag{4}$$

where $(\cdot)_i$ stands for the $i^{\text{th}}$ row-vector in the matrix.

Similarly, we have $\mathcal{L}_{\text{CACR-V}}$:

$$\mathcal{L}_{\text{CACR-V}} = \text{m-KL}(\sigma(S_{VL}S_{LL}S_{VL}^\top), \sigma(S_{VV})), \tag{5}$$

Combining $\mathcal{L}_{\text{CACR-V}}$ and $\mathcal{L}_{\text{CACR-L}}$, we present our $\mathcal{L}_{\text{CACR}}$ objective, an attention activation regularizer for cross-modal relation alignment:

$$\mathcal{L}_{\text{CACR}} = \mathcal{L}_{\text{CACR-V}} + \mathcal{L}_{\text{CACR-L}}. \tag{6}$$

When the vision inputs and the language inputs have the same sequence length and $S_{VL}, S_{LV}$ are invertible, then $S_{VV}$ and $S_{VL}S_{LL}S_{VL}^\top$ (as well as $S_{LL}$ and $S_{LV}S_{VV}S_{LV}^\top$) can become strictly congruent. In this case, $S_{VL}S_{LL}S_{VL}^\top$ can be interpreted as the language view of $S_{VV}$. Aligning

$S_{VL}S_{LL}S_{VL}^\top$ and $S_{VV}$ leads to cross-modal relation alignment. It is similar for $S_{LV}S_{VV}S_{LV}^\top$ and $S_{LL}$. In the general case where the vision inputs and the language inputs may have different sequence lengths, the two forms are not linear algebraically congruent but the relevant intuition still holds.

### 3.3 Hard and Soft Cross-modal Equivalence

In this section, we show that CACR can be interpreted as leveraging cross-modal soft equivalences, where IAIS [Ren et al., 2021] uses hard bijective equivalences. In their approach, each element in the intra-modal attention matrix is aligned with a single counterpart in the opposite modality. This is built upon a strict assumption that there exists a one-to-one mapping (provided by an argmax over the cross-modal attention) from $S_{LL}$ to $S_{VV}$ and vice versa, which is unsatisfied in practical cases. CACR may be seen as a soft cross-modal equivalence method which instead uses the whole $S_{LV}$ (or $S_{VL}$) to implicitly build an 'equivalence weighting' which is then used to compute a weighted mean over $S_{VV}$ (or $S_{LL}$). We illustrate and compare hard cross-modal equivalence and our soft cross-modal equivalence in Figure 3, taking the language-side alignment as an example.

We note that IAIS could be seen as a special case of soft cross-modal equivalence by forcing the cross-modal attention map to be a one-hot matrix, i.e., taking the argmax of the attention matrix as the index of the cross-modal counterpart. We show in Section 5 that IAIS can have inferior performance when a clear bijective cross-modal correspondence isn't available.

In Alg. 1, we show the pseudo-code of the soft cross-modal equivalence method for calculating the vision-side loss. $\mathcal{S}_{CACR-V}$ can be computed similarly. Computing the hard and soft cross-modal equivalence is computationally complex and difficult to be parallelized due to indexing operations. For practical applications, we sought to simplify this soft cross-modal equivalence algorithm to a mathematical equivalent that would improve computational tractability. From here, we arrive at CACR, which is a closed-form formulation of soft cross-modal equivalence which utilizes only differentiable matrix multiplications. Therefore, our CACR is more **computationally efficient** and **easier to be parallelized** than soft cross-modal equivalence.

---

**Algorithm 1** Soft Cross-modal Equivalence (V)

**Require:** $S_{LL} \in N \times N, S_{VL} \in N \times M, S_{VV} \in M \times M$

1: $\mathcal{L} \leftarrow 0$
2: **for** $i, j \in S_{VV}$ **do**
3: $\quad W \leftarrow S_{VL}[i] \cdot S_{VL}^\top[j]$ $\quad \triangleright$ soft weighting
4: $\quad a_{i \to j}^* \leftarrow \overline{W \circ S_{LL}}$ $\quad \triangleright$ element-wise weighted mean
5: $\quad \mathcal{L} = \mathcal{L} + \text{m-KL}(a_{i \to j}^*, S_{VV}[i, j])$
6: **end for**
7: **return** $\mathcal{L}$

---

### 3.4 Proof of Equivalence Between CACR and Soft Cross-modal Equivalence

Computing the hard (IAIS) and soft cross-modal equivalence is computationally complex and difficult to parallelize due to indexing operations. However, CACR loss is mathematically equivalent to soft cross-modal equivalence but can be computed efficiently. We take $\text{CACR}_V$ for illustration of this equivalence, but $\text{CACR}_L$ can be proved in the same way.

Beginning with the visual-basis form of $S_{LL}$ in CACR, the attention at index $[i, j]$ in $S_{VL}S_{LL}S_{VL}^\top$ is

$$
\begin{aligned}
&(S_{VL}S_{LL}S_{VL}^\top)[i, j] \\
&= \sum_p^{N_L} \sum_k^{N_L} a_{v_i \to l_k} a_{l_p \to l_k} a_{v_j \to l_p} \\
&= \sum_p^{N_L} \sum_k^{N_L} \underbrace{S_{VL}[i, k] S_{VL}[j, p]}_{\text{soft weighting}} S_{LL}[p, k]
\end{aligned}
$$

(7)

where $a_{v_i \to l_j}$ stands for the attention from the $i$-th visual token to the $j$-th linguistic token, $N_L$ is the total number of language tokens and $N_V$ is the total number of the visual tokens. Comparing Eq.7 and Alg.1, we observe that the summation we arrive at above is equivalent to the content of the for-loop (line 3-5). Thus, although of seemingly different linear algebraic form, CACR generalizes IAIS by way of its equivalence to the Soft Cross-modal Equivalence formulation presented above.

## 4 Results

How does CACR compare to other vision-language models in its compositional ability?

We fine-tuned CACR on Flickr30k [Young et al., 2014] for 5000 epochs using PyTorch [Paszke et al., 2019] with a train-validation-test split of 90-5-5.
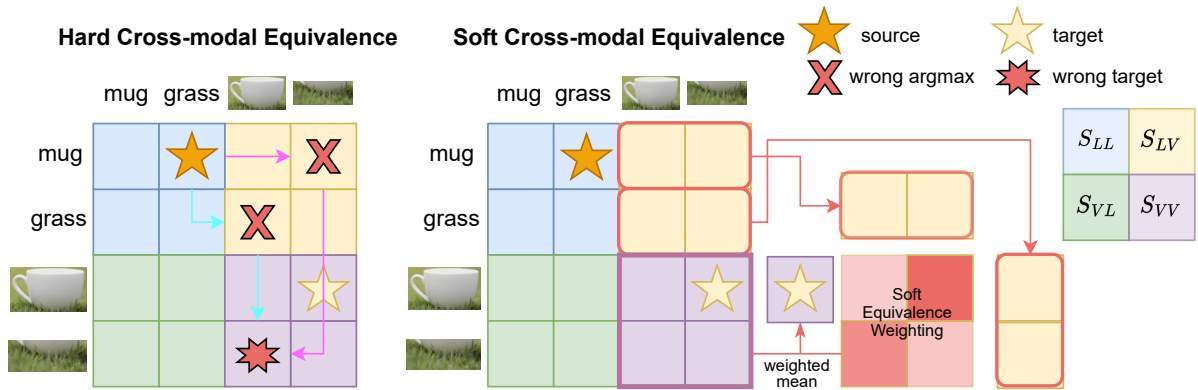
Figure 3: Comparison of the hard cross-modal equivalence used in IAIS (left) and the soft cross-modal equivalence we propose in CACR (right), with an example from Winoground to illustrate how the target visual equivalent (yellow star) of a source linguistic relation (orange star) is calculated. Hard cross-modal equivalence attempts to build a one-to-one mapping between language and vision by applying argmax (red crosses) to the $S_{LV}$ row vectors. Our soft cross-modal equivalence instead uses the whole $S_{LV}$ row vectors to calculate a weighted (red) mean over $S_{VV}$. The scalar that is produced corresponds to the attention from 'mug' to 'grass' but in a visual basis. We note that IAIS can be seen as a special case of soft cross-modal equivalence by forcing the attention matrix (red) to be a one-hot matrix where the max value is set to 1 and all others to 0. CACR linear algebraically simplifies soft cross-modal equivalence for computational efficiency.

The training batch size is 4 and 31 negative samples are provided for every individual positive sample in a standard image-text matching training setup. We use a learning rate of $5 \times 10^{-5}$, the $AdamW$ optimizer [Loshchilov and Hutter, 2017], and introduce $\mathcal{L}_{CACR}$ with an exponential warmup schedule. Training was completed on a node with 4 NVIDIA GTX 1080 Ti's, each with 11 GB of memory.

| Model | Text | Image | Group |
|---|---|---|---|
| MTurk Human | 89.50 | 88.50 | 85.50 |
| CACR<sub>base</sub> | 39.25 | **17.75** | **14.25** |
| UNITER<sub>large</sub> | **43.50** | 14.75 | 13.75 |
| IAIS<sub>base</sub> | 37.50 | 16.75 | 13.00 |
| UNITER<sub>base</sub> | 32.75 | 11.75 | 8.50 |

Table 2: CACR outperforms its pre-trained baseline (UNITER) and an alternative attention regularization approach (IAIS) across all Winoground scores.

In Tab. 2, we present our approach's scores alongside a few other models. Since we use CACR to fine-tune UNITER, we include scores for the two baseline UNITER sizes. We also include scores for IAIS<sub>base</sub> which is also built on UNITER.

The fact that CACR<sub>base</sub> outperforms IAIS<sub>base</sub> suggests that, with adequate computational resources, CACR<sub>large</sub> could similarly outperform IAIS<sub>large</sub>, potentially achieving a new state-of-the-art on Winoground. Furthermore, its performance compared to UNITER<sub>large</sub> is impressive consider-

ing that CACR<sub>base</sub> is approximately half its size in parameters. Despite our resource constraints, we were able to train a partly frozen (first 12 layers) version of CACR<sub>large</sub>, which achieves IAIS<sub>large</sub> levels of performance on Winoground (text: 37.5, image: 18.75, group: 15.75) with just 20% of the training time.

| Model | Image R@1 | Image R@10 | Text R@1 | Text R@10 |
|---|---|---|---|---|
| IAIS<sub>base</sub> | 73.54 | 96.32 | 86.10 | 99.10 |
| UNITER<sub>base</sub> | 72.52 | 96.08 | 85.90 | 98.80 |
| CACR<sub>base</sub> | 70.88 | 95.68 | 83.50 | 98.80 |

Table 3: CACR performance on Flickr30k has marginal reductions from UNITER, suggesting performance could be improved even further with a hyperparameter search.

Finally, we report Flickr30k retrieval scores in Tab. 3 to verify that we are not somehow overfitting to Winoground. Though CACR takes some minor losses to its retrieval scores, this may be attributed to imperfect hyperparameters, suggesting that CACR's performance on Winoground could be even higher with adequate hyperparameter tuning. It's also important to remember here that we're only training on Flickr30k, so this isn't a case of our model overfitting to Winoground and 'forgetting' its true image-text matching ability. Rather, it shows that the hyperparameters that we

adapted from IAIS need to be modified to more perfectly train CACR on Flickr30k, which would then carry over to compositional improvements on Winoground.

## 5 Analysis

Why does CACR's soft cross-modal equivalence approach outperform hard cross-modal equivalence?

### 5.1 Qualitative

Hard cross-modal equivalence, implemented by IAIS, assumes that cross-modal submatrices can be used to find a singular equivalent of an entity in the opposite modality. Specifically, if $i^* = \text{argmax}(S_{LV}[i])$ then $S_{LL}[i]$ should correspond to $S_{VV}[i^*]$. In simple terms, IAIS says the following: if word A attends most to object A and word B attends most to object B, then word A should attend to word B in a similar way that object A attends to object B. Underlying IAIS is the hard assumption that argmaxing over the cross-modal attention submatrix is an effective means of identifying the opposite modality equivalent of an entity. However, we show in this section that this is often not the case.

Given the argmaxes for rows in the $S_{LV}$ submatrix, we can identify the bounding box that each token maximally attends to, which IAIS assumes is its visual equivalent. In Fig. 4a, we visualize an example where 'clouds' maximally attends (green) to the ground, which would prevent IAIS from identifying the correct cross-modal equivalence. 'Turbines' (Fig. 4b), on the other hand, maximally attends to a bounding box that better matches our intuition. It is qualitatively clear from the several examples displayed that the argmax assumption often fails to identify the correct cross-modal equivalence. Since words may attend to several visual tokens for different reasons, we shouldn't assume that the cross-modal argmax provides us with a clear bijective correspondence.

Instead, the cross-modal matrices should be seen as providing useful high-level information about what visual entities are relevant to a word, and vice versa (as intuitively demonstrated by [Ilinykh and Dobnik, 2022]). We can certainly gain useful information about cross-modal correspondences using it, but it isn't as simple as using an argmax, due to words having multiple referents and entity representations being intermixed. Instead, our soft
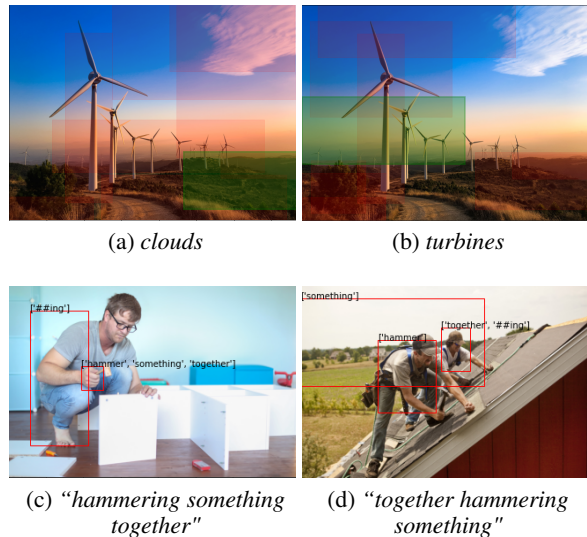


(a) *clouds*  (b) *turbines*



(c) *"hammering something together"*  (d) *"together hammering something"*

Figure 4: Top: UNITER $S_{LV}$ attention for caption *"a few clouds and many wind turbines"*, with the bounding box maximally attended to by the token in green; other highly attended boxes in red. Bottom: UNITER $S_{LV}$ attention with bounding boxes labeled with the tokens that maximally attend to them. Note that argmaxes often fail to precisely identify cross-modal equivalence.

cross-modal equivalence approach takes all the possible cross-modal alignments into account with a weighted sum.

To illustrate how the soft approach accounts for critical cross-modal alignment information, we present a few Winoground examples with UNITER's cross-modal attention activations in Fig. 4 and 5. We use UNITER since this is the baseline model from which attentional information is bootstrapped to calculate cross-modal alignments. For example, in Fig. 5c, using the representation for the bounding box covering the mug's handle may not adequately capture the visual referent of 'mug' and therefore disrupt our ability to calculate the visual-basis relation between 'mug' and 'grass' if restricted by an argmax.

### 5.2 Quantitative

In the absence of annotations, we attempted a quantitative measurement of whether overlap in argmaxes (several words attending to one bounding box or vice versa) as quantified by the Shannon Entropy of argmax indices inversely correlates with soft Winoground score. Intuitively, if an example has more like a one-to-one mapping between text and image, the entropy of its cross-modal argmaxes should be higher as each token will attend to a different box, which would suggest that the model is
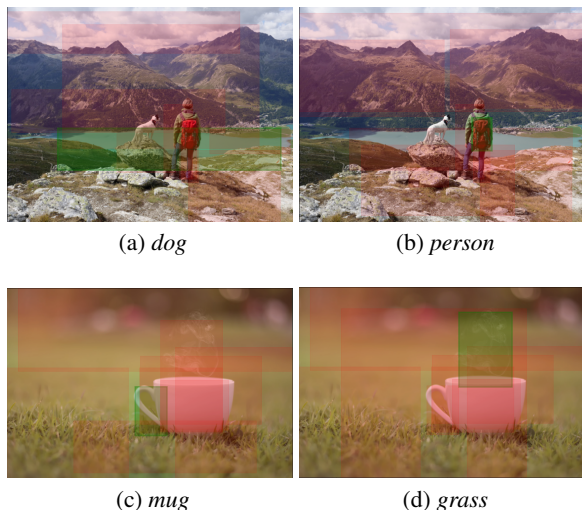
(a) *dog*  (b) *person*

(c) *mug*  (d) *grass*

Figure 5: UNITER $S_{VL}$ attention for captions *"a dog on a rock next to a person"* and *"there is a mug in some grass"*. Shown are boxes that attend highly to the displayed token, with the maximally attending bounding box in green; others in red. Observe that although the argmax often does pick up on a relevant bounding box, it is prone to missing critical visual information, e.g. focusing on only the backpack in (b).

better aligning entities. However, we found no significant correlation with Winoground score, which we attribute to the fact that high entropy on its own doesn't mean *correct* entity alignment.

Rather, high entropy in argmax indices could still be produced by a bad representation if 'mug' attends to the grass & 'grass' attends to the mug; conversely, low entropy could be produced by a good representation for an example like 'fire truck' where two tokens refer to a single object. Quantitative exploration of cross-modal attention is difficult without annotations and we leave this task to future work to explore in a multimodal compositionality context.

As a general takeaway, while the cross-modal argmax assumption of IAIS does hold in some cases and may be more meaningful during the course of IAIS training, it is clearly quite a strict assumption that could suffer if an entity attends to several cross-modal entities or there are no corresponding cross-modal entities. Furthermore, since IAIS is only active in the final self-attention layer, all the token representations are intermixed and therefore don't necessarily have a one-to-one correspondence with our intuitive notions of what they should be—the word 'turbine' may not solely represent the traditional meaning of that word but per-

haps the entire scene that includes the turbines, clouds, and ground.

We hypothesize that by removing the hard argmax assumption, our approach better accounts for varying cross-modal entity equivalences and thus enables stronger relation alignment. By also calculating alignment between all pairs of source and target modality entities, CACR should considerably improve sample efficiency, which is important considering that the final layer $S$ matrix of the converged IAIS model is largely flat. Therefore it's important to backpropagate as much alignment knowledge over the course of training as possible, which CACR's soft equivalence weighting implicitly enables.

# 6 Conclusion

In this work, we identified that a key factor holding back models from vision-language representational compositionality is cross-modal relation alignment. We categorized recent compositional inductive bias approaches into 3 categories: Structural Model, Structural Data, and Structural Training, showing that a previous Structural Training model (IAIS) achieves state-of-the-art performance on Winoground. We then identified a potential key weakness in IAIS, its hard argmax assumption, and developed a soft cross-modal equivalence approach to address it. Having linear algebraically simplified this approach, we arrived at CACR, an auxiliary loss that encourages cross-modal congruence of intra-modal attention. CACR improves on IAIS' performance on Winoground, and even outperforms a UNITER model nearly twice as large.

As computational scaling becomes more widespread, it's necessary to develop compositional inductive biases that do not require complex annotated data or exotic model architectures. Our work illustrates how taking advantage of the transformer's own attentional structure can improve the quality of fine-grained vision-language representations, opening the avenue for large scale approaches to visually-grounded compositionality.

# 7 Limitations, Risks & Ethics

Though CACR shows significant gains in compositional performance, results are limited in their exploration of only one pre-trained model and compositionality dataset. A significant risk of models is their tendency to be biased by distributions in their training data; vision-language models are

not free from this flaw, but we see our work as teaching VLMs to learn better structured representations rather than memorizing spurious correlations in data. We remain far from solving the vision-language compositionality problem, so biases must continue to be actively mitigated.

## Acknowledgements

## References

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, 2020.

Estelle Aflalo, Meng Du, Shao-Yen Tseng, Yongfei Liu, Chenfei Wu, Nan Duan, and Vasudev Lal. Vl-interpret: An interactive visualization tool for interpreting vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21406–21415, 2022.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 39–48, 2016.

Jiangang Bai, Yujing Wang, Yiren Chen, Yaming Yang, Jing Bai, Jing Yu, and Yunhai Tong. Syntax-bert: Improving pre-trained transformers with syntax trees. *arXiv preprint arXiv:2103.04350*, 2021.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs. *Transactions of the Association for Computational Linguistics*, 9:978–994, 2021. doi: 10.1162/tacl_a_00408. URL https://aclanthology.org/2021.tacl-1.58.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

Yuhao Cui, Zhou Yu, Chunqi Wang, Zhongzhou Zhao, Ji Zhang, Meng Wang, and Jun Yu. Rosita: Enhancing vision-and-language semantic alignments via cross-and intra-modal knowledge integration. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 797–806, 2021.

Longteng Guo, Jing Liu, Jinhui Tang, Jiangwei Li, Wei Luo, and Hanqing Lu. Aligning linguistic words and visual semantic units for image captioning. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 765–773, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368896. doi: 10.1145/3343031.3350943. URL https://doi.org/10.1145/3343031.3350943.

Yining Hong, Qing Li, Song-Chun Zhu, and Siyuan Huang. Vlgrammar: Grounded grammar induction of vision and language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1665–1674, October 2021.

Nikolai Ilinykh and Simon Dobnik. Attention as grounding: Exploring textual and cross-modal attention on entities and relations in language-and-vision transformer. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4062–4073, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.320. URL https://aclanthology.org/2022.findings-acl.320.

Zaid Khan, Vijay Kumar BG, Xiang Yu, Samuel Schulter, Manmohan Chandraker, and Yun Fu. Single-stream multi-level alignment for vision-language pretraining. *arXiv preprint arXiv:2203.14395*, 2022.

Taehyeong Kim, Hyeonseop Song, and Byoung-Tak Zhang. Cross-modal alignment learning of vision-language conceptual systems. *arXiv preprint arXiv:2208.01744*, 2022.

Zhongli Li, Qingyu Zhou, Chao Li, Ke Xu, and Yunbo Cao. Improving bert with syntax-aware local attention. *arXiv preprint arXiv:2012.15150*, 2020.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv preprint arXiv:2209.03430*, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

David Mareček and Rudolf Rosa. From balustrades to pierre vinken: Looking for syntax in transformer self-attentions. *arXiv preprint arXiv:1906.01958*, 2019.

Victor Milewski, Miryam de Lhoneux, and Marie Francine Moens. Finding structural knowledge in multimodal-bert. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5658–5671, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

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*, pages 8748–8763. PMLR, 2021.

Shuhuai Ren, Junyang Lin, Guangxiang Zhao, Rui Men, An Yang, Jingren Zhou, Xu Sun, and Hongxia Yang. Learning relation alignment for calibrated cross-modal retrieval. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 514–524, 2021.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. 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)*, pages 5100–5111, 2019.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022.

Bo Wan, Wenjuan Han, Zilong Zheng, and Tinne Tuytelaars. Unsupervised vision-language grammar induction with shared structure modeling. In *International Conference on Learning Representations*, 2021.

Yanan Wang, Michihiro Yasunaga, Hongyu Ren, Shinya Wada, and Jure Leskovec. Vqa-gnn: Reasoning with multimodal semantic graph for visual question answering. *arXiv preprint arXiv:2205.11501*, 2022a.

Zhecan Wang, Haoxuan You, Liunian Harold Li, Alireza Zareian, Suji Park, Yiqing Liang, Kai-Wei Chang, and Shih-Fu Chang. Sgeitl: Scene graph enhanced image-text learning for visual commonsense reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5914–5922, 2022b.

Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Hongwei Xue, Yupan Huang, Bei Liu, Houwen Peng, Jianlong Fu, Houqiang Li, and Jiebo Luo. Probing inter-modality: Visual parsing with self-attention for vision-and-language pre-training. *Advances in Neural Information Processing Systems*, 34:4514–4528, 2021.

Xu Yang, Chongyang Gao, Hanwang Zhang, and Jianfei Cai. Auto-parsing network for image captioning and visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2197–2207, 2021a.

Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9847–9857, June 2021b.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3208–3216, 2021.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bag-of-words models, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

Bryan Zhang. Improve MT for search with selected translation memory using search signals. In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track)*, pages 123–131, Orlando, USA, September 2022. Association for Machine Translation in the Americas. URL https://aclanthology.org/2022.amta-upg.9.

Junchao Zhang and Yuxin Peng. Hierarchical vision-language alignment for video captioning. In *International Conference on Multimedia Modeling*, pages 42–54. Springer, 2019.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021.

## A  For every submission:

☐ A1. Did you describe the limitations of your work?
*Left blank.*

☐ A2. Did you discuss any potential risks of your work?
*Left blank.*

☐ A3. Do the abstract and introduction summarize the paper's main claims?
*Left blank.*

☐ A4. Have you used AI writing assistants when working on this paper?
*Left blank.*

## B  ☐  Did you use or create scientific artifacts?

*Left blank.*

☐ B1. Did you cite the creators of artifacts you used?
*Left blank.*

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
*Left blank.*

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
*Left blank.*

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
*Left blank.*

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
*Left blank.*

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
*Left blank.*

## C  ☐  Did you run computational experiments?

*Left blank.*

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
*Left blank.*

---

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*Left blank.*

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
*Left blank.*

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
*Left blank.*

**D ☐ Did you use human annotators (e.g., crowdworkers) or research with human participants?**

*Left blank.*

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
*Left blank.*

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
*Left blank.*

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
*Left blank.*

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*Left blank.*

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
*Left blank.*