MM-SHAP: A Performance-agnostic Metric for Interpreting Multimodal Contributions in Vision and Language Models & Tasks

Anonymous ACL submission

Abstract

Vision and language models (VL) are known to exploit unrobust indicators in individual modalities (e.g., introduced by distributional biases), instead of focusing on relevant information in each modality. A small drop in accuracy obtained on a VL task with a unimodal model suggest that so-called unimodal collapse occurs. But how to quantify the amount of unimodal collapse, i.e., how multimodal are VL models really? We present MM-SHAP, a performance-agnostic multimodality score that quantifies the proportion by which a model uses individual modalities in multimodal tasks. MM-SHAP is based on Shapley values and will be applied in two ways: (1) we compare models for their degree of multimodality, and (2) measure the importance of individual modalities for a given task and dataset. Experiments with 6 VL models – LXMERT, CLIP and four ALBEF variants – on four VL tasks – image-sentence-alignment, Visual Question Answering, GQA and the more fine-grained VALSE benchmark – highlight that unimodal collapse can occur to different degrees and in different directions, contradicting the wide-spread assumption that unimodal collapse is one-sided. We recommend MM-SHAP to complement accuracy metrics when analysing multimodal tasks, as this can help guide progress towards multimodal integration.

1 Introduction

Vision and language (VL) tasks are dominated by general-purpose pretrained transformer-based VL models (Lu et al., 2019; Tan and Bansal, 2019; Li et al., 2019; Chen et al., 2020; Li et al., 2020, 2021a). But we are only starting to understand why these models work so well, and how they utilise and fuse the image and text modalities (Hessel and Lee, 2020; Cao et al., 2020). Even worse, these highly parametrised neural VL models, pretrained on large amounts of data, tend to exploit artefacts and statistical correlations in the data (Shekhar et al., 2019; Kafle et al., 2019), showing little to no evidence of detailed linguistic or visual understanding (Milewski et al., 2022; Parcalabescu et al., 2022; Thrush et al., 2022). Statistical biases towards indicators in one modality – to the detriment of others – can cause *unimodal collapse* (Parcalabescu et al., 2022), where seemingly multimodal models exploit one modality exhibiting biases, meaning that a multimodal system effec-

Figure 1: Image-sentence-alignment score (ISA) of three VL models with their textual degree `T-SHAP` (in %). Each text and image token (image patch) is colour-coded: Blue tokens contribute to a high ISA, while red ones lower the ISA. The visual degree is $100 - \texttt{T-SHAP}\%$. Cf. §4.5 for more explanation, App. B a more detailed analysis of this instance and for more samples. Note that the ISA of CLIP is an absolute score, while ALBEF and LXMERT predict ISA probabilities.

tively reduces to a unimodal model (Madhyastha et al., 2018) – e.g., a model answers "How many...?" questions with "two", the most frequent answer in the train set (Goyal et al., 2017). Unimodal collapse is severe, as it leads to loss of system reliability. It also shows that modality fusion is far from being solved. Hence the importance of measuring *multimodal degree* – the degree to which modalities are used for model predictions – with reliable metrics.

To test for unimodal collapse, research so far focuses on performance tests where a VL model is evaluated on a task – while one modality crucial for solving it correctly is missing, corrupted (Shekhar et al., 2017) or permuted (Gat et al., 2021).

But we argue that an appropriate contribution of each modality is not necessarily reflected in a VL model's measured performance. Clearly, accuracy reflects whether a model prediction is correct – but we cannot use it to identify cases where the model's prediction is wrong even though it considers relevant indicators in a given modality – or conversely, when a prediction is correct but derived on the grounds of inappropriate indicators. Fig. 1 shows how model responses, with almost identical image-sentence alignment (ISA) scores (and hence ISA accuracy), are concentrated on very different image regions and text tokens that contribute to the final model output, as indicated by Shapley values.

As an alternative to accuracy-based methods, we propose MM-SHAP as a *performance-agnostic metric* to quantify and interpret the contribution of individual modalities in VL models. MM-SHAP is based on Shapley values (Shapley, 1953), a theoretically well-founded interpretability method from cooperative game theory. They can be applied to measure the contribution of specific parts of the input towards a model prediction.

Our main contributions are:

i) We propose MM-SHAP, a performance-agnostic metric to measure the contribution of each modality in VL models (but which is not limited to V&L) to answer the question: *How much do VL models use individual modalities?* We combine MM-SHAP with model accuracy to analyse the degree to which each modality contributes to model predictions.
ii) We make use of MM-SHAP to (1) compare models in terms of their focus on different modalities, (2) to compare the relevance of different modalities for a given task and dataset, and to (3) zoom in at sample-level, to determine the contribution of each modality and each token in each modality for specific model predictions (see Fig. 1).
iii) We conduct experiments with six VL models: LXMERT (Tan and Bansal, 2019), CLIP (Radford et al., 2021a) and four ALBEF (Li et al., 2021a) variants – on four VL tasks: image-sentence-alignment, VQA (Goyal et al., 2017), GQA (Hudson and Manning, 2019) and on the more fine-grained VALSE🎭 VL benchmark (Parcalabescu et al., 2022).
iv) We identify VL models that are balanced in their usage of two modalities (CLIP), models that have a higher visual degree (LXMERT) or a stronger textual degree (ALBEF).
v) We show that (i) fine-tuning a model can affect its multimodal degree and that (ii) current VL models do not all collapse towards the same modality, as found by recent work (Frank et al., 2021; Gat et al., 2021), but that sides can differ from model to model.

2 Related Work

Testing for unimodal collapse Strong prediction indicators in either modality can cause multimodal models to ignore weaker indicators in another modality. Prior work has proposed different methods to identify and possibly remove such biases from the data (Goyal et al., 2017).

Foiling approaches introduce mistakes in image descriptions and test whether VL models notice the discrepancy between image and captions (Shekhar et al., 2019; Parcalabescu et al., 2022), finding that models are surprisingly insensitive to constructed foils. Gat et al. (2021), in a similar line of work, exchange the image with another image or the caption with a different caption. They assume that with inputs containing misleading information in one modality, model accuracy on the task decreases. They measure the decrease in task accuracy to calculate a perceptual score measuring the multimodal degree of the models. Their findings suggest that across the tested VL models, the textual input consistently matters more than the visual input.

Ablation methods remove information from either modality and test whether the model can still solve the task. Here, Frank et al. (2021) find an asymmetry: VL models suffer from removed image inputs when predicting masked text, but can predict masked visual inputs when textual input is ablated. Their findings contradict the conclusions of Gat

et al. (2021), but note that their investigations have only a single model in common, namely LXMERT.

We observe that the literature commonly agrees that VL models are not as cross-modal as expected – but we find considerable divergence in findings attributing models to rely more on the textual or on the visual side. In this work we argue that methods for measuring a model's multimodal degree should not rely solely on accuracy. This is because in robustness tests with ablated, permuted or corrupted inputs, accuracy-based methods can fail to capture cases where the model is right for the wrong reasons – or incorrect despite taking the right information into account. Moreover, accuracy-based methods cannot properly assess the contribution of each modality in cases where model accuracy is generally very low – as in out-of-domain or zero-shot settings. We therefore propose an *accuracy-agnostic* method for measuring the multimodal degree of VL models, using *SHAP* (Lundberg and Lee, 2017) as a theoretically well-founded interpretability method.

Interpretability Methods for explaining predictions of neural models can be classified into two categories: *White-box methods*, which require access to specific components of neural architectures and *black-box methods*, which are model-agnostic, requiring only access to model inputs and outputs.

Notable *white-box methods* are: Attention-based methods, which correlate high attention weights with high feature importance. However, the equivalence between importance score and attention is debated and has to be taken with care (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019). Cf. Appendix C for a discussion on why not to use attention for defining a multimodal score. Layer-wise relevance propagation (Binder et al., 2016) and gradient-based methods like Grad-CAM (Selvaraju et al., 2017) can also be used for determining the importance of inputs, but can be deceived by small changes in inputs (adversarial attacks).

Notable *black-box methods* are: LIME (Ribeiro et al., 2016), which approximates the vicinity of the input with a linear function that is interpretable. But depending on the choice of the size of the vicinity, LIME can lead to very different results. Methods like RISE (Petsiuk et al., 2018) and SHAP (Lundberg and Lee, 2017) compute importance scores by randomly masking parts of the input and determining the effect this has on the output. Among the latter two, SHAP exhibits great properties for interpretability, as detailed in Section 3.1.

3 Quantifying Multimodal Contributions

3.1 Background on Shapley Values

Shapley values were first introduced in a game theoretical setting to estimate fair rewards among cooperative players (Shapley, 1953). For machine learning, the outcome of a game is the model's prediction, the players are parts of the input and are assigned Shapley values that represent the importance of each player (Lundberg and Lee, 2017).

We compute Shapley values for pretrained transformer-based VL models at prediction time. Their input consists of p input tokens (image and text tokens alike). We create subsets $S \subseteq \{1, \ldots, n\}$ of tokens, where tokens not being part of the subset are masked, and all tokens contained in the subset form a coalition towards the model prediction $val(S)$. $val(\emptyset)$ is the output of the model when all tokens are masked. Then the Shapley value for a token j is computed by formula (1):

$$\phi_j = \sum_{S \subseteq \{1,\ldots,n\} \setminus \{j\}} \frac{val(S \cup \{j\}) - val(S)}{\gamma} \quad (1)$$

Here, $\gamma = \frac{|S|!(n-|S|-1|)!}{p!}$ is the normalising factor accounting for all possible combinations of choosing the subset S. When masking (or not masking) p tokens, the coalition possibilities grow exponentially, i.e. $n = 2^p$. Therefore we use a Monte Carlo approximation of the Shapley values by randomly sub-sampling $n = 2p + 1$ coalitions.

The Shapley value of a token measures its contribution towards the model prediction (e.g., the probability of image-sentence-alignment) and can be **positive** (increases the model prediction) or **negative** (decreases it) or **zero** (no effect). Shapley values exhibit four defining properties of a fair payout, which are all beneficial for model interpretability: (1) *Efficiency*: the contributions of all players sum up to the model outcome; (2) *Symmetry*: any two players that contribute the same are assigned the same payout; (3) *Dummy*: a non-contributing part is assigned zero value and (4) *Additivity*, enabling us to simply average the Shapley Values to determine the overall player contributions in a game with combined payouts (e.g., the two halves of a soccer match, or ensembling of decision trees).

Most importantly, Shapley values are not based on the model's accuracy or performance, but *solely on the model's input and its prediction*, such as the probability for an image and a caption to match. This is an important property for a multimodality

score to have, since its objective is to quantify *how much multimodal inputs of either modality matter for prediction* – even if the cooperation between (multimodal) inputs is not sufficient to reach success (i.e., yielding the correct outcome).

3.2 MM-SHAP

For a pretrained VL transformer with n_T text tokens and n_I image tokens, Eq. 2 defines the textual contribution Φ_T and the image contribution Φ_I towards a prediction as the sum of (absolute) Shapley Values (Eq. 1) of all textual resp. visual tokens:

$$\Phi_T = \sum_{j}^{n_T} |\phi_j| \quad ; \quad \Phi_I = \sum_{j}^{n_I} |\phi_j| \qquad (2)$$

We ignore the sign of the token contributions[1] and consider their magnitude since we are interested in measuring whether a token is active within a modality – irrespective of the correctness of the ensuing prediction. In Eq. 3 we define MM-SHAP as a proportion of modality contributions, which allows us to determine a model's textual degree `T-SHAP` and its visual degree `V-SHAP`:

$$\texttt{T-SHAP} = \frac{\Phi_T}{\Phi_T + \Phi_I}; \texttt{V-SHAP} = \frac{\Phi_I}{\Phi_T + \Phi_I} \qquad (3)$$

We can extend MM-SHAP to any number of modalities. Here we only use image and text.

When generating coalitions, i.e., subsets of tokens from which to compute Shapley Values, we do not differentiate between image and text tokens, because the idea of MM-SHAP is to fairly distribute potential token contributions first, and to aggregate contributions modality-wise in a second step, using Eq. 2. For **masking tokens**, text tokens are replaced with the [MASK] token,[2] while for images, we mask out patches by image space, replacing pixel values with zero (see Section 4.4 for details).

3.3 Ways of using MM-SHAP

Sample-level MM-SHAP is a sample-level score (cf. Fig. 1) based on the contribution of individual image and text tokens. It thus enables fine-grained analyses of the relevance of tokens from a single or various modalities, for individual instances.

Dataset and model level Sample-level MM-SHAP scores can be averaged to yield dataset-level multimodality scores, thanks to the additivity property of Shapley values. We use MM-SHAP at dataset level to analyse a given model on different datasets or different models on a given dataset, to gain insights about models, datasets and tasks.

Measuring fine-tuning effects An accuracy-based multimodality score reaches its limits when the model performance on a task is very low, and the difference between model accuracy with correct inputs vs. permuted inputs is small by default. In such cases, the Perceptual Score (Gat et al., 2021) will assign a low multimodal score, irrespective of the relevance of multimodal inputs. Since MM-SHAP is not based on task performance – measured by comparing model prediction to the gold standard (e.g., with accuracy) –, but on the actual model predictions, we can use MM-SHAP to measure multimodal scores of models with low performance. This allows us, e.g., to compare a pretrained model to a fine-tuned version of it that may have lost general abilities (of, e.g., image-sentence alignment) after specialising on another task.

Future work could focus on applying MM-SHAP on models accepting different or a wider range of modalities, or on data cleaning by filtering out samples with very unbalanced multimodal degree.

4 Multimodal Contributions across Models and Datasets

We use MM-SHAP to study multimodal contributions i) for different model types, ii) on different datasets and iii) tasks. In doing so we will re-evaluate findings from prior research on visual vs. textual modality collapse. We will also showcase MM-SHAP's abilities for interpreting predictions for individual samples, to enable deeper error analysis.

We evaluate pretrained VL models with MM-SHAP and complement this analysis with accuracy-based assessments of model performance on multiple tasks. Prior work has presented findings showing models to be either consistently more visual (Gat et al., 2021) or textual (Frank et al., 2021). But assessing multimodal contributions can be misleading when evaluating models using performance-oriented metrics. We thus use the performance-agnostic MM-SHAP metric to analyse whether we find such trends to be consistent across models and tasks, or whether models differ with respect to the modality they rely on the most.

[1] Contributions can be positive (increase the model prediction) or negative (decrease it) or zero (no effect), see §3.1.

[2] See App. A for details on the choice of masking.

To assess whether the multimodal degree of a model tends towards the textual or the visual modality, we compare MM-SHAP to a 50% `T-SHAP` : 50% `V-SHAP` baseline for image-sentence alignment, where we hypothesise that in average, V&L should contribute equally when the model predicts whether the contents of the modalities are aligned.

We investigate the setting where image and caption match, but also cases of discrepancy between modalities. We break down our incongruity analyses into *high discrepancy* cases, where image and caption are in a complete mismatch (Table 1), and cases of *low discrepancy*, where only a single word or phrase incurs a mismatch (Table 2).

4.1 Tasks

Visual Question Answering A canonical task where pretrained VL transformers have consistently increased state-of-the-art performance through fine-tuning is Visual Question Answering. We use the VQA v2.0 (Goyal et al., 2017) and the GQA (Hudson and Manning, 2019) datasets for testing the contribution of V&L in multimodal models.

Image-sentence alignment (ISA) VL models are usually pretrained on predicting an image-sentence alignment score. We are interested in assessing the multimodal contributions in such VL models when using them within their "comfort zone", by testing how well they predict the alignment of images and captions in contrast to misalignment between images and random captions.

We test this on 1,500 samples from the validation set of MSCOCO (Lin et al., 2014), and further evaluate ISA model performances on more uncommon image-caption pairs composed from questions and answers from the validation sets of the VQA and GQA datasets (1,500 image-caption pairs each).

ISA on fine-grained visio-linguistic phenomena In image-sentence alignment task settings models are usually confronted with negative samples (non-matching image–caption pairs) of high discrepancy. To evaluate VL models in a more fine-grained manner, we examine their multimodal contributions on the VALSE VL benchmark (Parcalabescu et al., 2022). It contains foiled captions targeting six linguistic phenomena: existence, counting, plurality, spatial relations, actions, coreference. Foiled captions were created by altering a word or phrase that realises a specific linguistic phenomenon, such that image and foiled caption do not match. For sake of completeness, we also test on foiled noun phrases in the FOIL it! dataset (Shekhar et al., 2017).

4.2 Models

LXMERT (Tan and Bansal, 2019) is a dual-stream transformer model that combines V&L through early fusion using cross-modal attention layers between image and language encoders. Its pretraining data consists of MSCOCO (Lin et al., 2014) images and captions, and VQA v2.0 and GQA images, questions and answers. Pretraining objectives were (i) multimodal masked word and object prediction, (ii) image-sentence alignment, and (iii) question-answering. For experiments on ISA, VQA and GQA, we use LXMERT's[3] corresponding heads and task-specific checkpoints.

CLIP (Radford et al., 2021b) processes image and text with two separate transformer-based encoders. The resulting image and text representations are combined in late fusion by cross-product. CLIP is trained on 400M image-text pairs to predict high scores for paired image-text examples and low scores when image-text samples are not paired in the dataset. With this simple contrastive learning objective, CLIP is capable of zero-shot capabilities in e.g. object classification, OCR, or activity recognition (Radford et al., 2021b). For our experiments, we use CLIP[4] for tests on image-sentence-alignment and VALSE, where we use the model's image-text alignment score to assess whether a higher image-text similarity is predicted for correct pairs or for foiled image-caption pairs.

ALBEF (Li et al., 2021b) uses early and late fusion to combine V&L. As in CLIP, the transformer-based image and text encoders map the two modalities to a common space. Subsequently, the representations are further combined through cross-modal transformer layers with objectives of (i) multimodal masked word prediction and (ii) image-sentence alignment. ALBEF is pretrained on Conceptual Captions (Sharma et al., 2018), SBU Captions (Ordonez et al., 2011), MSCOCO (Lin et al., 2014) and Visual Genome (Krishna et al., 2017).

To analyse to what extent the contribution of modalities can be affected by fine-tuning on diverse tasks and datasets, we compare four ALBEF[5] models fine-tuned on (1) image retrieval on MSCOCO,

[3] `github.com/huggingface/transformers`
[4] `github.com/openai/CLIP`
[5] `github.com/salesforce/ALBEF`

(2) image retrieval on Flickr30k (Plummer et al., 2015), (3) visual grounding on RefCOCO+ (Yu et al., 2016) and (4) VQA (Goyal et al., 2017).

4.3 Metrics

We use two main categories of metrics: **accuracy** to measure model performance on each task, and **MM-SHAP** to assess the proportion to which the different modalities contribute.

With MM-SHAP (as defined in Section 3.2), we aim to analyse the multimodal contributions in terms of visual degree `V-SHAP` and textual degree `T-SHAP`. Since these are complementary metrics for – in our case – two modalities, `V-SHAP` $=$ $100 -$ `T-SHAP`. We hence report only `T-SHAP` (in %). We distinguish $\texttt{T-SHAP}_c$ for the textual degree in image-caption pairs and $\texttt{T-SHAP}_f$ for image-foil pairs.

When evaluating VQA and GQA performance, accuracy measures the proportion of correct answers given pairs of images and questions. For ISA, we fan out the accuracy metric into three metrics: **caption accuracy** acc_c measures whether models correctly predict images and captions to match; **foil accuracy** acc_f quantifies whether models correctly predict mismatching images and captions; **pairwise accuracy** acc_r measures the proportion of samples where the ISA score is higher for a correct image-text pair compared to its foil. acc_r is more permissive than acc_c and acc_f as it does not require the ISA score to surpass a classification threshold, but only that image-foil pairs are ranked lower for ISA than their ground truth image-caption pairs.

4.4 Experimental Setting

We test all VL models as described in Section 4.2 without further tuning and assess both their task accuracy and their MM-SHAP scores on VQA, GQA and VALSE data.

For masking we ensure that text length and image sequence lengths are similar, i.e., for longer text, we have more smaller patches and vice versa. For the majority of samples in our data, this results in 16 image patches. See Appendix A for details.

4.5 Experiments and Results

We now report the results of our experiments for different model types at dataset level, in three settings: i) for the canonical VQA task on the VQA and GQA datasets; for image-sentence alignment ii) with high discrepancy between image and caption foils (on data from MSCOCO, VQA, GQA) and iii) where the discrepancy between images and caption foils is lower, on VALSE data; finally iv) we show sample-level analyses of MM-SHAP results. Results on VQA, GQA and ISA are presented in Table 1; Table 2 shows results for VALSE.

Note that individual MM-SHAP scores vary from sample to sample. In Tables 1 and 2, we report MM-SHAP mean values with a standard deviation of 11% to 13% across all our experiments.

High discrepancy ISA Unsurprisingly, acc_r scores for ISA on MSCOCO, VQA and GQA (Table 1) are high for all models since they have been pretrained for this task – ALBEF vqa being the odd one out, as it has lost its ISA performance during fine-tuning on VQA. LXMERT has highest acc_r for ISA on VQA and GQA, which is unsurprising, as its last 10 epochs contained data from the training sets of these datasets.

For image-sentence alignment, we observe that the models scatter around the hypothesised 50-50% balance for `T-SHAP`, with CLIP being the most balanced one, especially on MSCOCO. This is expected since CLIP is a two-branch model where the two modalities communicate in late fusion, in other words, CLIP keeps all information from the textual and visual branches separate until the very end. By contrast, LXMERT has a low textual degree of only 35.5%, while ALBEF models are more textual.

With highly diverging information in the two modalities, we observe that differences are prominent between $\texttt{T-SHAP}_c$ and $\texttt{T-SHAP}_f$, especially for LXMERT that moves from weak textual degree (35.5%) to higher textual degree (62.8%) and inversely for ALBEF mscoco (63.4% to 54.3%).

Canonical VL tasks Results on VQA and GQA in Table 1 – with ALBEF fine-tuned for VQA and LXMERT fine-tuned on VQA and GQA[6] – show high model accuracy. The textual degrees are higher for VQA than for ISA between image and matching caption. This is interesting to note, especially since LXMERT was more visually focused on ISA. It seems that fine-tuning on VQA and LXMERT performing a VQA task increases the impact of the textual to the detriment of visual input modality. This aligns with earlier findings that in VQA tasks, linguistic indicators (such as "How many...?") give away the most likely answer (two) (Goyal et al., 2017).

[6]We do not test CLIP and the other ALBEF models on VQA because they do not have corresponding VQA heads.

Model	VQA		GQA		Image-sentence alignment														
					MSCOCO					VQA					GQA				
	acc	T	acc	T	acc_c	acc_f	acc_r	T_c	T_f	acc_c	acc_f	acc_r	T_c	T_f	acc_c	acc_f	acc_r	T_c	T_f
Random	0.0	50.0	0.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
LXMERT	72.5	51.5	60.3	57.8	71.8	99.1	99.3	**35.5**	**62.8**	66.6	95.9	95.2	45.7	57.5	41.8	96.5	89.9	47.5	59.8
CLIP	-	-	-	-	-	-	99.5	50.3	52.9	-	-	94.0	48.4	47.6	-	-	83.4	47.0	46.0
A mscoco	-	-	-	-	95.9	99.6	99.8	**63.4**	54.3	28.0	99.9	91.0	60.3	59.2	13.1	99.7	83.6	58.3	57.2
A flickr	-	-	-	-	97.3	99.4	99.7	**61.1**	56.6	42.4	99.2	91.8	**61.3**	60.2	23.4	99.5	84.1	58.7	58.1
A refcoco	-	-	-	-	92.3	99.3	99.7	56.6	58.9	49.8	99.1	90.0	57.8	58.6	25.0	98.4	85.6	58.2	59.3
A vqa	76.0	66.7	-	-	99.9	0.0	33.4	**64.1**	**62.8**	100.0	0.0	60.2	58.2	60.0	100.0	0.0	52.6	61.7	**62.4**

Table 1: Task accuracy and multimodality score on canonical tasks. T is T−SHAP (in %). V−SHAP $= 100-$T−SHAP. acc_r is pairwise ranking accuracy, counting predictions as correct if $p(caption, img) > p(random, img)$. **A** stands for ALBEF fine-tuned for different tasks: image retrieval on MSCOCO and Flickr30k; visual grounding on RefCOCO+ and VQA. Overall foil task performance is the mean of acc_c and acc_f (equal nb. of samples, all pairs).

Metric	Model	Existence quantifiers	Plurality number	Counting bal.†	Counting sns.†	Counting adv.†	Sp.rel.‡ relations	Action repl.†	Action actant swap	Coreference std.†	Coreference clean	Foil-it! nouns	Avg. ± stdev.
	Random	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0±0
acc_r	CLIP	66.9	56.2	62.1	62.5	57.5	64.3	75.6	68.6	52.1	49.7	88.8	64.0±11
	LXMERT	78.6	64.4	62.2	69.2	42.6	60.2	54.8	45.8	46.8	44.2	87.1	59.6±15
	A mscoco	78.6	80.1	71.8	74.3	68.9	74.6	79.8	62.6	62.2	59.6	**97.0**	**73.6**±11
	A flickr	80.6	78.9	71.0	73.6	64.3	73.3	82.4	55.5	59.9	57.7	96.6	72.1±12
	A refcoco	73.1	69.0	67.9	70.7	45.7	68.6	79.9	58.9	52.7	43.3	96.5	66.0±15
	A vqa	40.8	63.3	49.0	49.2	23.2	61.9	51.7	52.0	55.9	43.3	67.2	50.7±12
acc_c	LXMERT	41.6	68.0	50.9	50.0	61.5	73.1	35.8	36.8	81.2	80.8	72.3	59.3±17
	A mscoco	18.4	93.2	26.7	23.7	34.6	95.9	66.2	64.9	87.0	89.4	96.1	63.3±32
	A flickr	28.7	94.0	43.1	41.2	50.8	96.8	65.1	64.2	91.5	96.2	97.5	69.9±26
	A refcoco	33.7	89.8	41.8	31.0	57.2	93.1	72.5	75.0	81.4	90.4	92.7	69.0±24
	A vqa	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0±0
acc_f	LXMERT	70.1	42.2	53.0	60.8	37.3	28.4	66.4	60.2	18.4	17.3	69.3	47.6±20
	A mscoco	91.5	27.1	82.0	87.2	80.9	9.2	61.7	42.3	16.1	12.5	52.1	51.1±32
	A flickr	82.4	18.5	66.4	70.9	58.6	7.1	63.3	38.8	8.2	4.8	42.4	41.9±28
	A refcoco	71.3	19.4	62.0	72.9	41.8	10.5	53.2	29.7	18.4	8.7	61.19	40.8±25
	A vqa	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0±0
T−SHAP$_c$	CLIP	44.7	52.3	51.5	51.8	52.1	50.9	50.0	49.7	52.1	52.6	49.9	50.7±2
	LXMERT	51.7	**37.1**	46.5	47.3	46.4	**36.6**	42.1	42.2	**38.2**	**37.2**	**36.1**	41.9±5
	A mscoco	56.7	**63.5**	58.3	58.0	59.5	**64.1**	**61.7**	**61.5**	**61.9**	**61.4**	**63.9**	60.9±3
	A flickr	59.5	**61.7**	59.6	59.8	59.5	**61.6**	59.8	58.9	60.9	**61.9**	**63.5**	60.6±1
	A refcoco	53.3	57.2	55.4	55.1	55.8	57.0	54.5	54.4	57.9	58.9	56.8	56.0±2
	A vqa	**64.6**	**63.6**	**62.5**	**61.4**	**63.4**	**63.0**	59.3	60.3	**63.6**	**63.1**	**62.1**	**62.4**±2
T−SHAP$_f$	CLIP	45.2	53.0	50.8	51.7	51.1	51.0	48.3	48.2	52.4	52.1	50.0	50.3±2
	LXMERT	52.3	**39.4**	48.2	48.8	45.8	**36.5**	43.9	42.7	**39.1**	**38.6**	45.0	43.7±5
	A mscoco	57.2	**62.8**	57.7	56.0	57.0	**64.6**	**61.9**	**63.2**	**61.9**	**61.8**	**65.8**	60.9±3
	A flickr	56.1	**61.9**	57.8	57.8	58.5	**62.5**	59.3	**61.9**	**61.1**	**62.1**	**61.7**	60.1±2
	A refcoco	56.1	58.5	56.2	55.6	57.8	57.6	55.5	56.9	58.4	58.4	**61.3**	57.5±2
	A vqa	**64.0**	**64.7**	**61.9**	**60.9**	**61.2**	**63.2**	59.9	60.1	**63.4**	**62.4**	**62.2**	**62.2**±2

Table 2: Performance and multimodal score of VL models on the instruments of the VALSE benchmark. We bold-face high accuracies and multimodally unbalanced models on tasks. acc_r is the pairwise ranking accuracy, considering predictions as correct if $p(caption, img) > p(foil, img)$. Overall foil task performance is the mean of acc_c and acc_f (equal number of samples, all pairs). **A** stands for ALBEF models fine-tuned on different tasks and datasets: image retrieval on MSCOCO and Flickr30k, visual grounding on RefCOCO+ and VQA. †**bal.** Counting balanced. †**sns.** Counting small numbers. **adv.** Counting adversarial. **repl.** Action replacement. ‡ **Sp.rel.** Spatial relations. †**std.** Coreference standard. We test CLIP in pairwise ranking mode only (CLIP works contrastively).

Low discrepancy ISA The tests on VALSE (Table 2) are all based on ISA, where we expect a 50%-50% balance between V−SHAP and T−SHAP. We mark high deviations from this baseline in bold-face (above 61% and below 40% T−SHAP). Indeed, we observe that the scores generally do not deviate much from this baseline. CLIP is by far the multimodally most balanced model, with an average T−SHAP$_c$ of 50.7% across all instruments, which is expected, as argued for high discrepancy ISA above. By contrast, LXMERT is skewed towards the visual modality with an average T−SHAP$_c$ of 42%, while ALBEF is generally more focused on text, its variants showing T−SHAP$_c$ values of 57% to 62%. These findings are consistent with our results for high discrepancy ISA in Table 1.

We do not find notable differences between foils and captions in terms of MM-SHAP, while we find clear differences in accuracies. A notable exception on VALSE is the difference between T−SHAP$_c$ and T−SHAP$_f$ for LXMERT and ALBEF refoco on Foil-it! (underlined numbers in Table 2).

Accuracy vs. MM-SHAP Overall, accuracies do not correlate with MM-SHAP (see Appendix A for details). Hence, our experiments strongly suggest that MM-SHAP is *complementary to accuracy* for assessing multimodal contributions.

Comparing results per model across VALSE instruments, we note that models are better with some instruments (noun phrases, existence) as opposed to others (actions, coreference). While this was already observed by Parcalabescu et al. (2022), our work adds the multimodal score MM-SHAP as a new dimension of analysis. Some models exhibit pronounced differences in `T-SHAP` score across instruments: LXMERT is especially visually focused for plurality, spatial relations and noun phrases, while ALBEF's general strong focus on text is especially concentrated on text on these phenomena.

Dataset bias As for the relationship between accuracy, MM-SHAP and dataset bias, we observe different behaviour between ISA and VQA.

For ISA on VALSE, in Table 2, we see that despite varying model accuracies (standard deviation across phenomena ranges from 11 to 15%), MM-SHAP is relatively stable across phenomena (1 to 5% stdev.), even when the data distribution is very different: For example, the *adversarial* piece in the counting instrument contains foils of small numbers from 0 to 3, while captions involve numbers higher than 4. The piece serves as a sanity check against biased models that may prefer small numbers, being more frequent in datasets. We note for LXMERT and ALBEF refcoco that acc_r drops for counting *small numbers* to counting *adversarial* from 69.2% to 42.6% for LXMERT and 70.7% to 45.7% for ALBEF, while $\texttt{T-SHAP}_c$ stays remarkably constant (47.3% to 46.4% and 55.1% to 55.8%) – see encircled numbers in Table 2.

For VQA we have conducted further experiments beyond those listed in Table 1 on the balanced set of GQA, which controls the answer distribution bias of questions in GQA *balanced*. While LXMERT shows 57.8% `T-SHAP` on GQA (cf. Table 1), our experiments on GQA *balanced* show a much more harmonic MM-SHAP score of 51.4% (+6.4 points), which is much closer to LXMERT's 51.5% `T-SHAP` on VQA v.2 (cf. Table 1), demonstrating that MM-SHAP can capture dataset biases.

Fine-tuning effects Comparing the four ALBEF models fine-tuned on different tasks and datasets on VALSE, we observe that the capacity of the models to predict ISA is high for the ALBEF models fine-tuned for image retrieval (73.6% average acc_r for ALBEF mscoco) and lower for VQA (ALBEF vqa 50.7%) and referring expressions (ALBEF refcoco 66.0%). This is expected, since ISA and image retrieval are very similar tasks, while VQA and referring expressions differ more. Interestingly, not only accuracy, but also the multimodal score changes, making ALBEF for VQA more focused on text (62.4% average $\texttt{T-SHAP}_c$ across VALSE) compared to referring expressions (ALBEF refcoco 56.0%). Notably, MM-SHAP being accuracy-agnostic, we can compute indicative scores even in cases where a fine-tuned model fails the task completely, such as ALBEF vqa that always predicts the foil class on captions.

Sample-level analysis Fig. 1 shows ISA predictions of CLIP, ALBEF mscoco and LXMERT and their `T-SHAP` values for caption and foil. LXMERT correctly predicts high ISA between image and *caption* (left), although the regions contributing most (in blue) are not all reasonable, since the 'phone' token is not correctly grounded. ALBEF mscoco and CLIP also assign very high ISA scores, while using well-justified image regions for thumb and phone. On the foil (right), LXMERT's contributing tokens change, with the phone region in the image mistakenly contributing to a high ISA. Favourably for LXMERT and ALBEF, the 'keyboard' text token contributes towards raising the ISA, unlike for CLIP, where the 'keyboard' token lowers the ISA. For more examples see App. B.

5 Conclusions and Future Work

We presented MM-SHAP, a performance-agnostic metric that measures the multimodal degree of VL models at dataset and sample level. Our analyses show that VL models vary in which modality they rely on most: ALBEF is rather textual, CLIP is balanced, LXMERT shows higher visual than textual degree. This confirms findings in Gat et al. (2021) and contradicts Frank et al. (2021). Using MM-SHAP we are the first to quantify changes in a model's multimodal degree through fine-tuning. Our experiments and analyses show that degrees of multimodal contributions can be orthogonal to task performance, supporting the need for performance-agnostic metrics. MM-SHAP is applicable to further modalities. It enables model-based data cleaning and bias removal. It can serve as a diagnostic tool for improving multimodal fusion methods.

6 Ethical Considerations

This paper uses publicly available datasets and models and therefore could carry on their potential biases and imperfections. However, the method presented in this paper enables model and dataset interpretation and *can help future work locate harmful biases*.

References

Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. 2016. Layer-wise relevance propagation for neural networks with local renormalization layers. In *International Conference on Artificial Neural Networks*, pages 63–71. Springer.

Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. 2020. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *European Conference on Computer Vision*, pages 565–580. Springer.

Hila Chefer, Shir Gur, and Lior Wolf. 2021a. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406.

Hila Chefer, Shir Gur, and Lior Wolf. 2021b. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *ECCV*.

Ian Covert, Scott M Lundberg, and Su-In Lee. 2020. Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33:17212–17223.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. 2021. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9847–9857, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Itai Gat, Idan Schwartz, and Alex Schwing. 2021. Perceptual score: What data modalities does your model perceive? *Advances in Neural Information Processing Systems*, 34.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Jack Hessel and Lillian Lee. 2020. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 861–877, Online. Association for Computational Linguistics.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Théo Jaunet, Corentin Kervadec, Romain Vuillemot, Grigory Antipov, Moez Baccouche, and Christian Wolf. 2021. Visqa: X-raying vision and language reasoning in transformers. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):976–986.

Kushal Kafle, Robik Shrestha, and Christopher Kanan. 2019. Challenges and prospects in vision and language research. *Frontiers in Artificial Intelligence*, 2:28.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 11336–11344. AAAI Press.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021a. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021b. Align before fuse: Vision and language

representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. In *Arxiv*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Pranava Swaroop Madhyastha, Josiah Wang, and Lucia Specia. 2018. Defoiling foiled image captions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 433–438, New Orleans, Louisiana. Association for Computational Linguistics.

Victor Milewski, Miryam de Lhoneux, and Marie-Francine Moens. 2022. Finding structural knowledge in multimodal-bert. *arXiv preprint arXiv:2203.09306*.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, Dublin, Ireland. Association for Computational Linguistics.

Letitia Parcalabescu, Albert Gatt, Anette Frank, and Iacer Calixto. 2021. Seeing past words: Testing the cross-modal capabilities of pretrained V&L models on counting tasks. In *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*, pages 32–44, Groningen, Netherlands (Online). Association for Computational Linguistics.

Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. RISE: randomized input sampling for explanation of black-box models. *CoRR*, abs/1806.07421.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021a. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021b. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

L. S. Shapley. 1953. *17. A Value for n-Person Games*, pages 307–318. Princeton University Press.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. FOIL it! find one mismatch between image and language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265, Vancouver, Canada. Association for Computational Linguistics.

Ravi Shekhar, Ece Takmaz, Raquel Fernández, and Raffaella Bernardi. 2019. Evaluating the representational hub of language and vision models. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 211–222, Gothenburg, Sweden. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2019. 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, Hong Kong, China. Association for Computational Linguistics.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. *arXiv preprint arXiv:2204.03162*.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. 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 11–20, Hong Kong, China. Association for Computational Linguistics.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer.

A Experimental Details

Masking VL models predict their outputs (such as ISA) on full and uncorrupted image and text inputs. To compute Shapley values and with them the MM-SHAP score, we create coalitions by masking image and text tokens.

For masking, we aim for a balance between text and image sequence length. Therefore we use the text length to dynamically determine patch sizes: For longer text, we use more and smaller patches and for shorter text, less but bigger patches. In the majority of our experiments, we have 16 image patches. We illustrate the image tiling in the top right of Figures 2 to 9.

This masking procedure has several advantages: i) It adapts to variable caption lengths and variable image sizes, and ii) it directly applies to different types of VL model architectures, since some apply transformers directly on the image (CLIP and ALBEF), while others compute image tokens with a different CNN-based backbone (LXMERT).

Special tokens When computing token-wise contributions, we do not take [SEP] and [CLS] tokens into account (i.e. they are always assigned zero contribution), since their functionality is to aggregate cross-modal information, e.g. for classification, and hence they cannot be attributed to one modality exclusively.

Correlation between accuracy and MM-SHAP For each model and instrument on VALSE, we computed the Spearman correlation coefficient between the sample's accuracy and textual degree. The correlations are very low, e.g., the correlation between acc_c and $\texttt{T-SHAP}_c$ is around 0.02 for most instruments and models, rising to 0.12 in rare cases.

B Sample-level Analyses with MM-SHAP

SEE FIGURES ON FOLLOWING PAGES!

Figures 2 to 9 contain sample-level visualisations for each model for images and i) captions that match and ii) foils / random captions that show low / high discrepancy mismatch with the images, as introduced in Section 4.5:

- There is **low discrepancy** between images and foils obtained from VALSE targeting specific linguistic phenomena, with only a phrase differing between the caption and the foil. We selected examples for different phenomena: Figure 2 (noun phrase), 3 (action replacement, easy example), 4 (counting), 5 (positive existence), 6 (negative existence), 9 (action replacement, hard example).

Figure 2: **Low discrepancy** *noun phrase* foil: Image-sentence-alignment score (ISA) of the six VL models with their textual degree T-SHAP (in %). Each text and image token (image patch) is colour-coded: Blue tokens contribute to a high ISA, while red ones lower the ISA. The visual degree is $100 - \texttt{T-SHAP}$. Note that the ISA of CLIP is an absolute score, while ALBEF and LXMERT predict ISA probabilities. With ✓ we mark correct ISA and highlight the correct / foil token that contributes in the right direction for aligning the image and the caption. With ⚡, we mark incorrect ISA and wrong contribution directions.

- There is **high discrepancy** between MSCOCO images and randomly chosen captions in terms of low ISA between image and random caption – Figures 7 (easier

example) and 8 (harder example).

In Figure 2 we reiterate Figure 1 from the main paper with more detail:

- CLIP correctly predicts a foil in the pairwise accuracy setting, since the ISA score for the caption (30.3) is higher than for the foil (29.9), but fails to identify that "keyboard" should not contribute towards a high ISA. It successfully predicts caption alignment, but seems to misunderstand the meaning of the word "shines" and its instantiation in the image.

- ALBEF mscoco is the only model to predict ISA (99.4%) on the caption with coherent – but mostly textual – indicators. It fails on foil prediction, still relying on the same textual indicators, and on the visual side *focuses on counter-evidence regions*, erroneously taking them as positive support for ISA.

- LXMERT predicts correct ISA for the caption (99.5% ISA), using few relevant textual tokens as indicators, and possibly useful supporting visual tokens (focuses the fingers of the two hands). It fails to detect the foil (99.4% ISA which is higher than a 50% classification threshold and just slightly below the ISA for the caption): counterevidence from textual tokens is out-weighted by a single strong indicator (thumb); visual tokens confirm ISA despite focusing on counterevidence (the phone).

On the following pages we present Figures 4 to 9 with more samples and their analyses.

We sampled the instances based on the following criteria: i) low / high discrepancy; ii) interesting VALSE instruments; iii) easier (no cluttering, no dark spots, no blur) and iv) harder examples (e.g., hard to recognise the statue as such in Figure 9).

Through Fig. 4 to 9, we observe some patterns:

Model performance does not tell much about the multimodal degree. A correct ISA score (high for the caption, low for the random caption/foil) is not always accompanied by a sensible contribution pattern in terms of Shapley values as seen for example in Figures 2 and 4 for CLIP and LXMERT. The Shapley values computed on the image and text side deliver much better intuition about what was successfully aligned and what was not grounded correctly. Among all models, LXMERT seems to be most affected by high discrepancy between performance and image and text token contributions.

Easy examples deliver more robust contribution patterns. On easy examples (Figures 3 and 4), where the model generally performs well, we can see how in the low discrepancy cases where caption and foil differ in only one word, the one word difference does not change the contribution patterns much. In contrast, low discrepancy hard examples (Figures 8 – unusual bed and bedroom arrangement and 9 – hard to recognise the goat as a statue without world knowledge) deliver different patterns on caption and foil, indicating confusion from the models.

Positive existence is easier than negative existence. When comparing Figures 5 and 6 we get some insight into how the models' image-sentence-alignment pretraining objective affects their behaviour:

For positive existence, where the caption indicates that **an object is present in the image** – as in Fig. 5: *There are children.* – is better handled by the models, delivering more sensible patterns for image-caption pairs. The contribution patterns on the negated version of the existence sentence – the foil *There are no children.* – show that some models handled the negation correctly (CLIP, LXMERT, ALBEF mscoco and refcoco), while the rest do not.

Negative existence, where the caption indicates that **an object is not present in the image** – as seen in Fig. 6: *There are no humans in the picture.* – seems more difficult to align, since the objects are not present in the image and to assign a high ISA for text mentions that cannot be located, the model needs to understand the negation. The foil, changing the sentence to affirmative – *There are humans in the picture.* – turns the instance into a much simpler case of no image-sentence-alignment, as is often seen during pretraining. Unsurprisingly, all models correctly predict a low ISA in Figure 6.

Counting is hard. In Figure 4 for the counting foils in VALSE, CLIP is the only model that assigns higher ISA for the image-caption pair and not to the image-foil pair. Overall, the contribution patterns look scattered: High visual contributions in the image indicate that the models align the plane object to its mention in the sentence, but we see confused textual contributions from the mentioned number of planes (0 or 4) in the text. This is unsurprising, given the low performance of VL models in counting as highlighted by Parcalabescu et al. (2021).

Figure 3: **Low discrepancy** (VALSE *action replacement*): Image-sentence-alignment score (ISA) of the six VL models with their textual degree `T-SHAP` (in %). Each text and image token (image patch) is colour-coded: Blue tokens contribute to a high ISA, while red ones lower the ISA. The visual degree is $100 - \texttt{T-SHAP}$. Note that the ISA of CLIP is an absolute score, while ALBEF and LXMERT predict ISA probabilities. With ✓ we mark correct ISA and an highlight the correct / foil token that contributes in the right direction for aligning the image and the caption. With ⚡, we mark incorrect ISA and wrong contribution directions.

C Why not to use Attention for defining a Multimodality Score

For defining a multimodality score that aims at quantifying each modality's contribution to any model prediction, we need an interpretability method that has crucial properties to do so. With their properties of efficiency, symmetry, dummy variable, additivity (see §3.1), Shapley values provide important ingredients for *sample-based explanations* that can be aggregated in a straightforward way into *dataset-level explanations* for machine learning methods (Covert et al., 2020). Other interpretability methods lack the robustness and theoretical foundation to produce a multimodality score that is comparable to the one proposed in our work.

In particular, for attention – while being widely used for generating visually appealing heat-maps – it is questionable how much high/low attention scores correlate with high/low contributions of input features for system predictions (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019).[7] While

[7] Arguably this may be the case when attention weights are

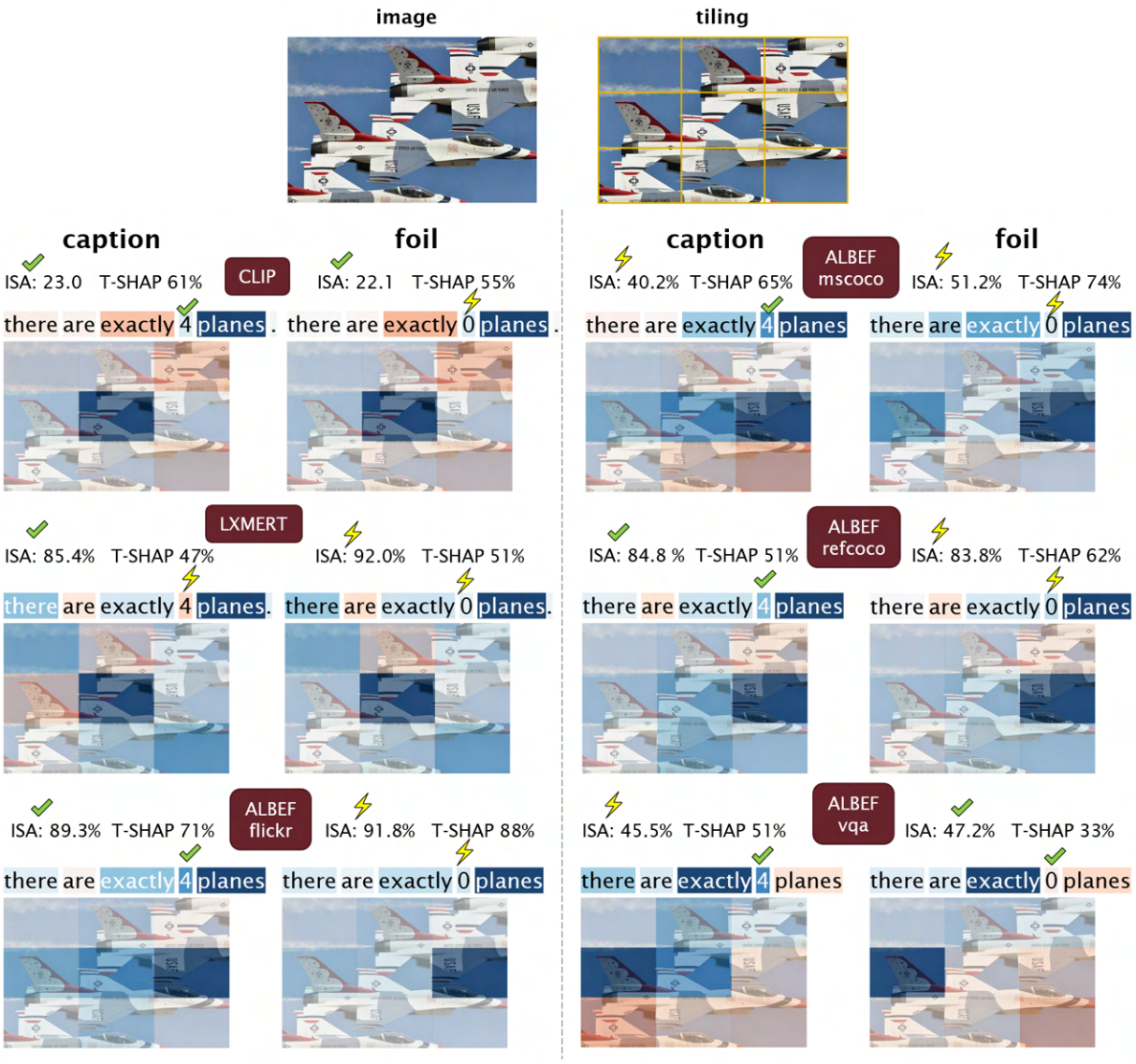

Figure 4: **Low discrepancy** (VALSE *counting*): Image-sentence-alignment score (ISA) of the six VL models with their textual degree `T-SHAP` (in %). Each text and image token (image patch) is colour-coded: Blue tokens contribute to a high ISA, while red ones lower the ISA. The visual degree is 100 − `T-SHAP`. Note that the ISA of CLIP is an absolute score, while ALBEF and LXMERT predict ISA probabilities. With ✓ we mark correct ISA and an highlight the correct / foil token that contributes in the right direction for aligning the image and the caption. With ⚡, we mark incorrect ISA and wrong contribution directions.

attention does linearly combine input features and determines how much of each token is mixed with every other token, it does not necessarily mean that a low attention value cannot have a large impact on the decision of the model. In other words, a pinch of salt is enough to make food taste good: Even if the attention score for salt is low, its contribution to the taste of the food (captured by Shapley values) is high.

Attention is present in transformers in multiple layers and to complicate the matter even further, each attention layer contains multiple attention heads. Hence, to visualize attention we need a carefully designed interface, as proposed, e.g., by Jaunet et al. (2021) `https://visqa.liris.cnrs.fr/` to keep a reasonable overview of all attention values. When integrating the multiple attention values and propagating them back to the input to assign relevancy values for image and text tokens, research strives to generate simple explanations that represent the most important tokens and tend to inhibit the rest, as can be seen on the progress from Chefer et al. (2021b) to Chefer et al. (2021a) (cf. Figure 4 in Chefer et al. (2021a)). Fur-

high, but it is clearly not the case when attention weights are low.

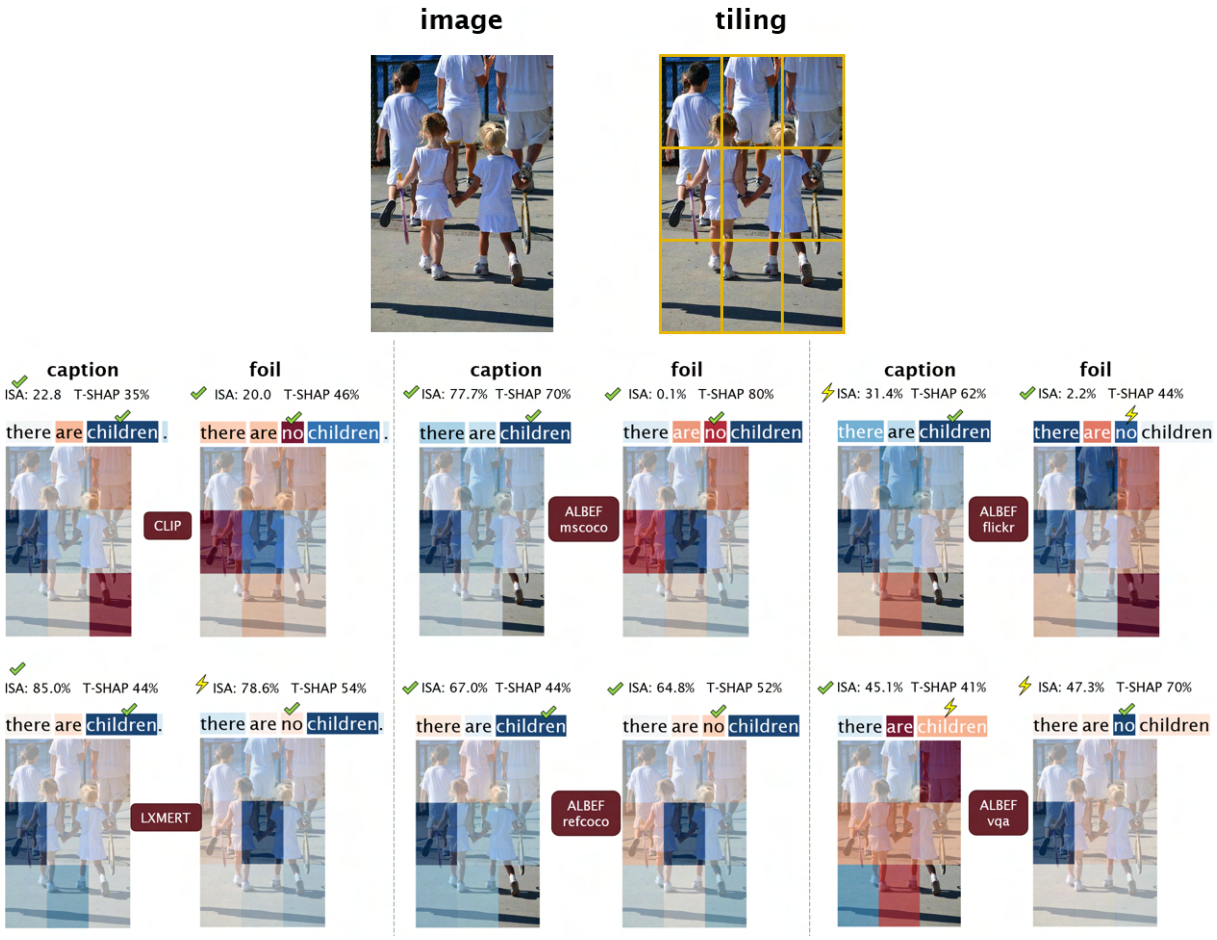

Figure 5: **Low discrepancy** (VALSE *existence positive*): Image-sentence-alignment score (ISA) of the six VL models with their textual degree `T-SHAP` (in %). Each text and image token (image patch) is colour-coded: Blue tokens contribute to a high ISA, while red ones lower the ISA. The visual degree is $100 - \texttt{T-SHAP}$. Note that the ISA of CLIP is an absolute score, while ALBEF and LXMERT predict ISA probabilities. With ✔ we mark correct ISA and an highlight the correct / foil token that contributes in the right direction for aligning the image and the caption. With ⚡, we mark incorrect ISA and wrong contribution directions.

thermore, while Shapley values estimate both the positive and the *negative contributions* of input tokens towards the model prediction, Chefer et al. (2021a) allows for positive-only relevance assessments.

In Figures 10 and 11, we have visualised CLIPs attention-based relevancy for the image-caption and foil examples shown in Figures 2 to 7 using the method of Chefer et al. (2021a). On the image side, we observe little to no changes in the attention visualisation, when comparing image-caption to image-foil pairs (cf. Figure 10). Even more, on the text side, both the correct and the foil word carry relatively similar attention scores, with no indication whether this contributes positively or negatively towards the model prediction. Shapley values however, are sensitive to foil words and we can visualise whether the word contributes towards raising the ISA (high image-sentence match) or lowering the ISA (e.g., Figure 3).

Besides the problematic interpretation of attention as feature contribution and the many ways of aggregating and propagating the different attention values to the input, another problem with attention is that it is unclear how to disentangle and aggregate the textual self-attention, visual self-attention, text-to-image attention and image-to-text attention into a single multimodality score that assesses the degree to which a given modality contributes towards the model prediction.

All things considered, we argue that attention is not well-suited as a basis for a multimodality score we aim for in this work, but that Shapley values – as presented in this paper – are, thanks to their theoretical properties (efficiency, symmetry, dummy variable, additivity) and their property

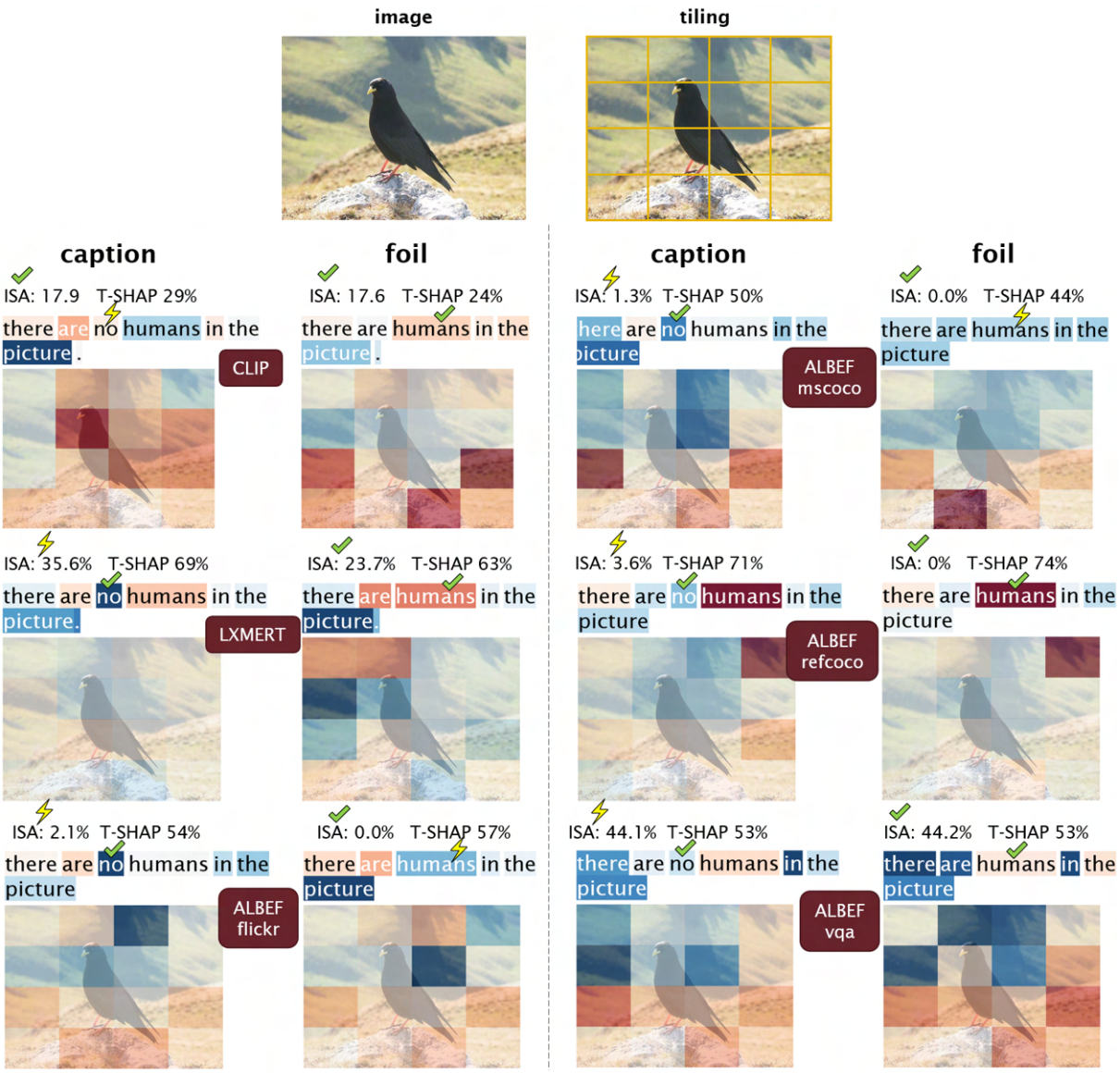

Figure 6: **Low discrepancy** (VALSE *existence negative* – harder phenomenon than positive existence): Image-sentence-alignment score (ISA) of the six VL models with their textual degree `T-SHAP` (in %). Each text and image token (image patch) is colour-coded: Blue tokens contribute to a high ISA, while red ones lower the ISA. The visual degree is $100 - \texttt{T-SHAP}$. Note that the ISA of CLIP is an absolute score, while ALBEF and LXMERT predict ISA probabilities. With ✓ we mark correct ISA and an highlight the correct / foil token that contributes in the right direction for aligning the image and the caption. With ⚡, we mark incorrect ISA and wrong contribution directions.

of being model-agnostic measurements of input feature contributions.

D Compute footprint

Computing all possible coalitions between input tokens for Shapley Values is infeasible because their number is exponential in the number of tokens (2^p). Therefore we perform Monte Carlo approximation by randomly sub-sampling $2p + 1$ coalitions. This results in approximate MM-SHAP scores per sample. We argue that as an alternative, one can simply increase the number of sampled coalitions for more exact measurements (as we did for Fig. 1 and the examples in Appendix B) – at the cost of increasing the environmental footprint. But increasing the number of samples is not necessary when estimating MM-SHAP at dataset level, because the number of coalitions has very little effect on a data-set wide range – given that approximation fluctuations average out.

MM-SHAP is computed while running models in inference mode $2p + 1$ times, where p is the number of tokens to mask (around 30 in average for MSCOCO-sized captions). On an NVIDIA Ti-

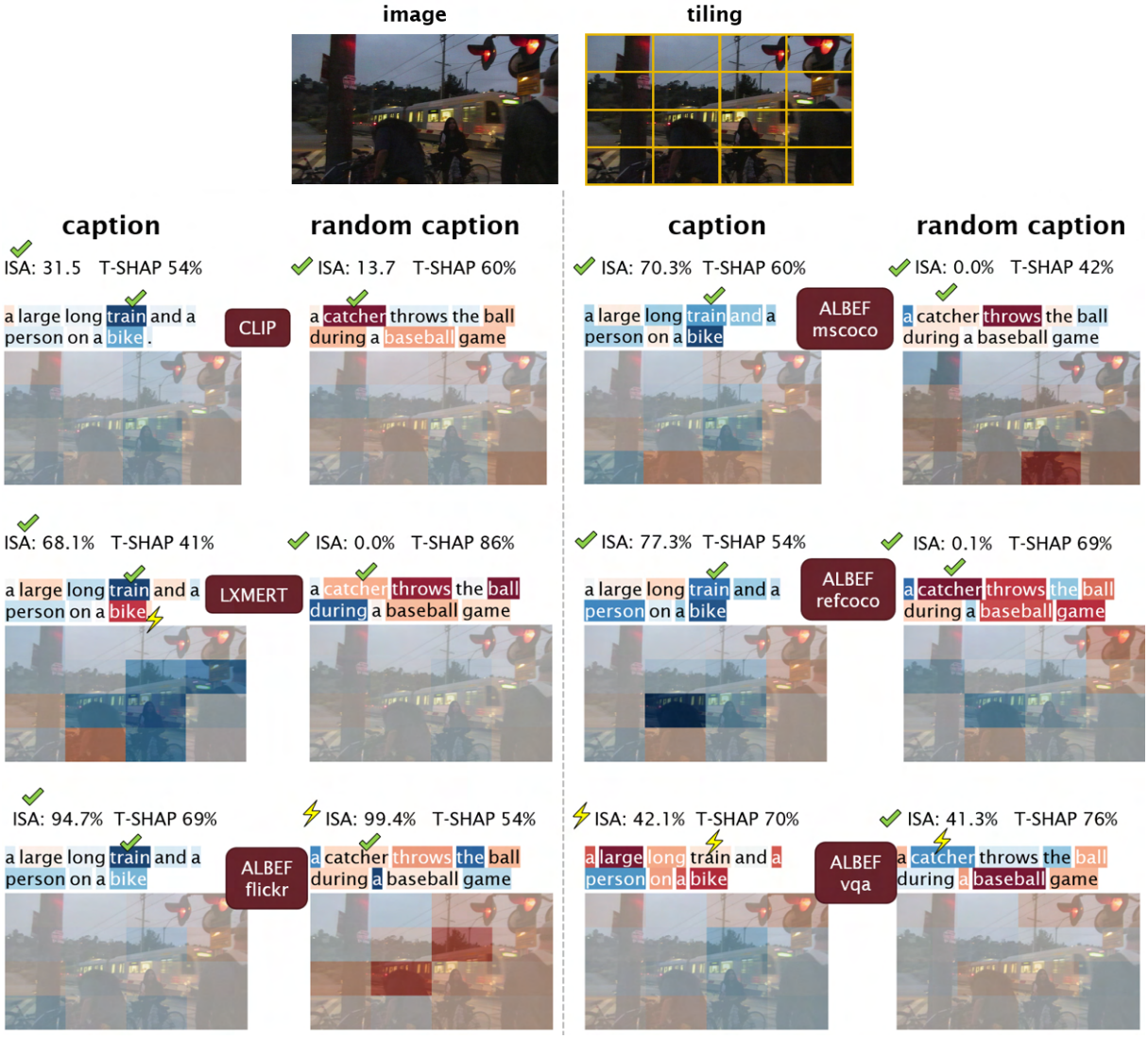

Figure 7: **High discrepancy** (MSCOCO): Image-sentence-alignment score (ISA) of the six VL models with their textual degree `T-SHAP` (in %). Each text and image token (image patch) is colour-coded: Blue tokens contribute to a high ISA, while red ones lower the ISA. The visual degree is $100 - \texttt{T-SHAP}$. Note that the ISA of CLIP is an absolute score, while ALBEF and LXMERT predict ISA probabilities. With ✓ we mark correct ISA and an highlight one important token that contributes in the right direction for aligning the image and the caption. With ⚡, we mark incorrect ISA and wrong contribution directions.

tan X GPU, computing MM-SHAP for one image-caption pair can take 2 seconds for ALBEF, 3 seconds for CLIP. LXMERT needing 15 seconds is the most expensive, because it computes image features with a CNN backbone for every masking configuration.

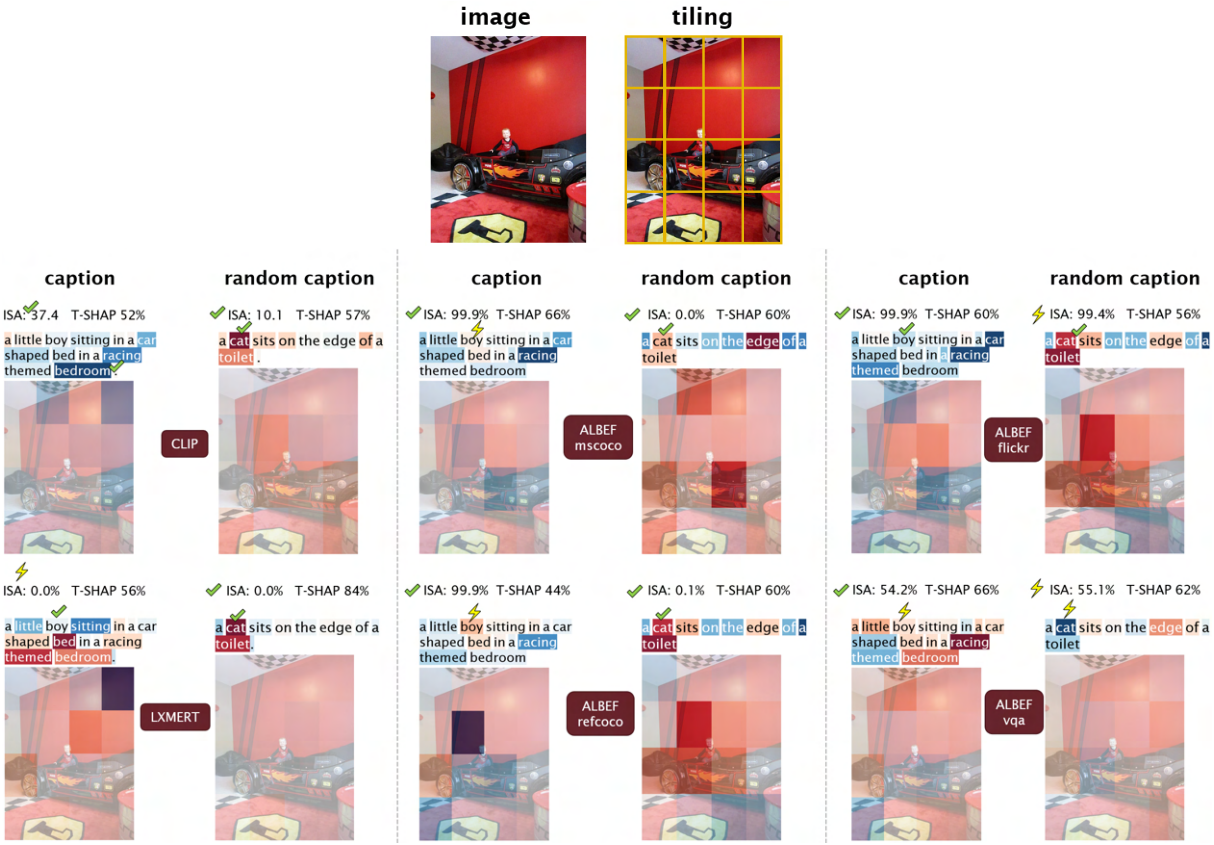

Figure 8: **High discrepancy** (MSCOCO) *hard example* where the models have trouble recognising the bed: Image-sentence-alignment score (ISA) of the six VL models with their textual degree `T-SHAP` (in %). Each text and image token (image patch) is colour-coded: Blue tokens contribute to a high ISA, while red ones lower the ISA. The visual degree is $100 - \texttt{T-SHAP}$. Note that the ISA of CLIP is an absolute score, while ALBEF and LXMERT predict ISA probabilities. With ✔ we mark correct ISA and highlight one important token that contributes in the right direction for aligning the image and the caption. With ⚡, we mark incorrect ISA and wrong contribution directions.

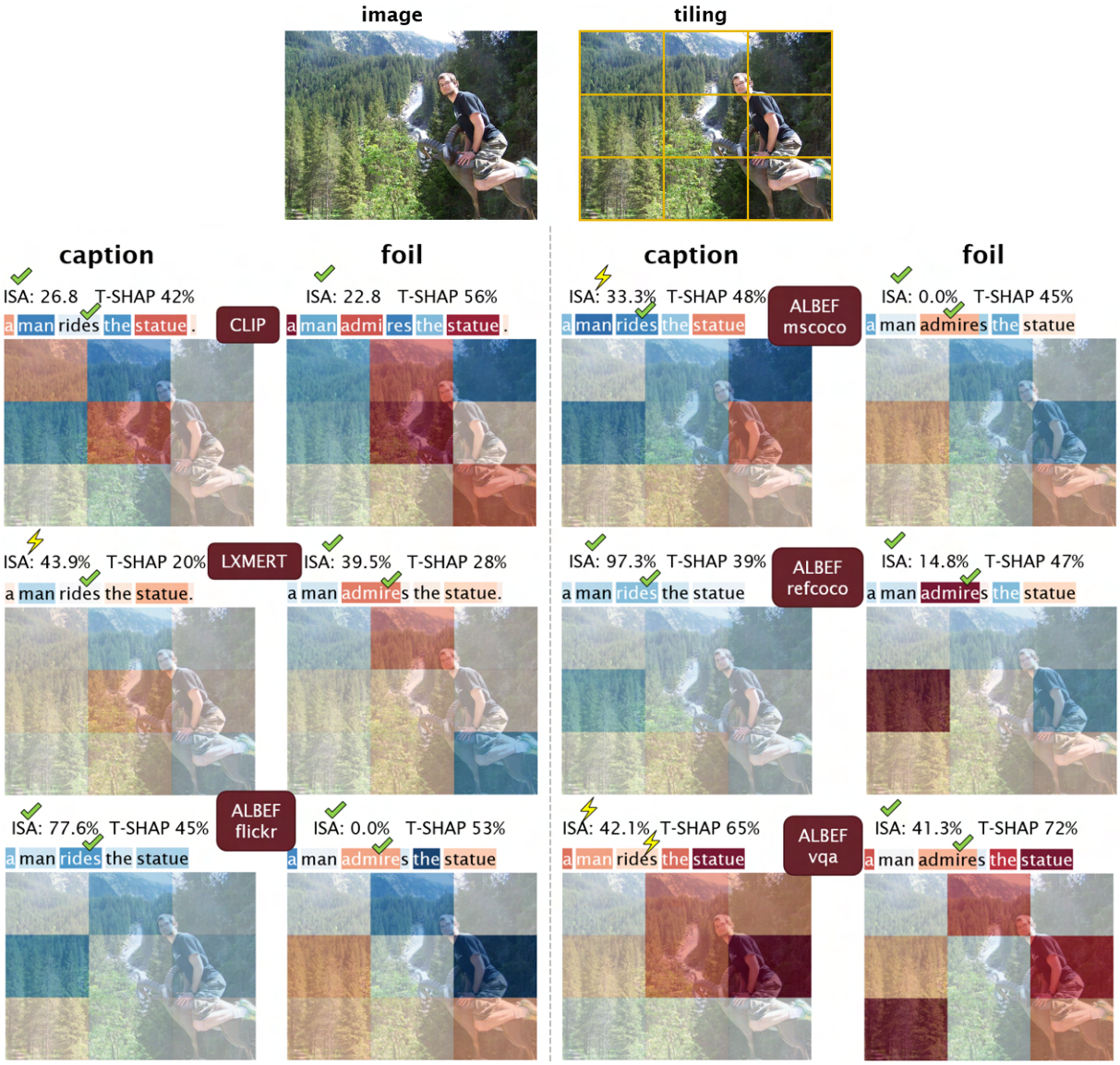

Figure 9: **Low discrepancy** (VALSE *action replacement*) – *hard example* where models and humans have trouble recognising the goat as a statue): Image-sentence-alignment score (ISA) of the six VL models with their textual degree `T-SHAP` (in %). Each text and image token (image patch) is colour-coded: Blue tokens contribute to a high ISA, while red ones lower the ISA. The visual degree is $100 - \texttt{T-SHAP}$. Note that the ISA of CLIP is an absolute score, while ALBEF and LXMERT predict ISA probabilities. With ✓ we mark correct ISA and highlight the correct / foil token that contributes in the right direction for aligning the image and the caption. With ⚡, we mark incorrect ISA and wrong contribution directions.

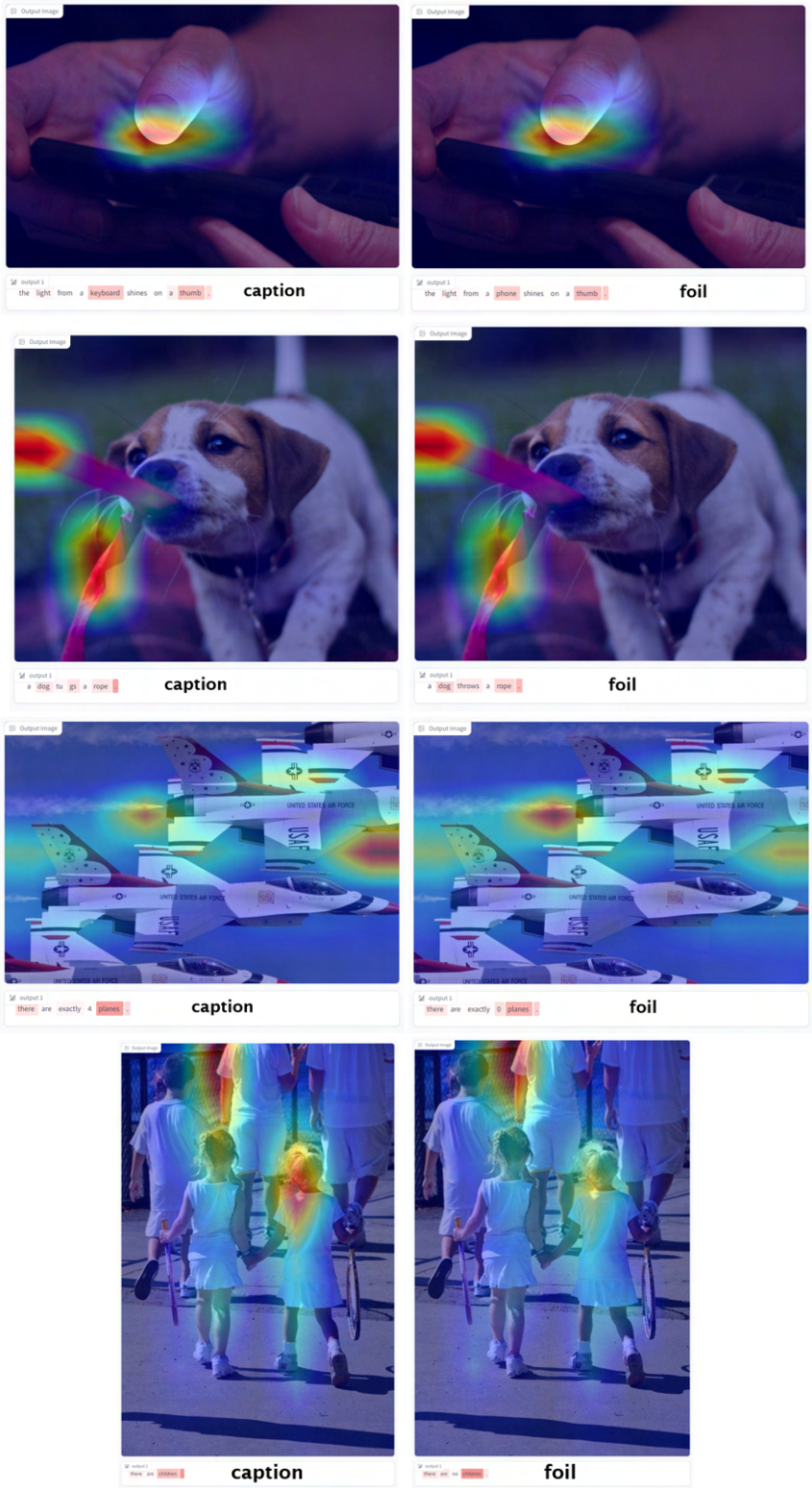

Figure 10: **Low discrepancy**. CLIP results of attention-based relevance visualisation, using the method of Chefer et al. (2021a) https://huggingface.co/spaces/PaulHilders/CLIPGroundingExplainability. Red means high relevancy, blue is zero relevancy and there is no negative relevancy (while Shapley values do allow for positive and negative contributions). Note that the heat-maps give the impression that the relevance irradiates from single spots. This is an artefact from the visualisation since the model works with 32x32 pixel patches and it is these patches that each have a relevance score. For reference: the images are around 500 pixels in height and width.

Figure 11: **High discrepancy**. CLIP results of attention-based relevance visualisation, using the method of Chefer et al. (2021a) https://huggingface.co/spaces/PaulHilders/CLIPGroundingExplainability. Red means high relevancy, blue is zero relevancy and there is no negative relevancy (while Shapley values do allow for positive and negative contributions). Note that the heat-maps give the impression that the relevance irradiates from single spots. This is an artefact from the visualisation since the model works with 32x32 pixel patches and it is these patches that each have a relevance score. For reference: the images are around 500 pixels in height and width.