UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

CV-Probes: Studying the interplay of lexical and world knowledge in visually grounded verb understanding

Permalink

https://escholarship.org/uc/item/3h83566r

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 47(0)

Authors

Benova, Ivana
Gregor, Michal
Gatt, Albert

Publication Date

2025

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

CV-Probes: Studying the interplay of visual grounding and social knowledge in verb understanding

Ivana Benova (ivana.benova@kinit.sk)
Faculty of Informatics, Brno University of Technology, Czech Republic
Kempelen Institute of Intelligent Technologies, Slovakia

Michal Gregor (michal.gregor@kinit.sk)
Kempelen Institute of Intelligent Technologies, Slovakia

Albert Gatt (a.gatt@uu.nl)
Utrecht University, Utrecht, Netherlands

Abstract

How do vision-language (VL) transformer models ground verb phrases and do they integrate contextual and world knowledge in this process? We introduce the CV-Probes dataset, containing image-caption pairs involving verb phrases that require both social knowledge and visual context to interpret (e.g., 'beg'), as well as pairs involving verb phrases that can be grounded based on information directly available in the image (e.g., "sit"). We show that VL models struggle to ground VPs that are strongly context-dependent. Further analysis using explainable AI techniques shows that such models may not pay sufficient attention to the verb token in the captions. Our results suggest a need for improved methodologies in VL model training and evaluation. The code and dataset will be available `https://github.com/ivana-13/CV-Probes`.

Keywords: multimodal models; grounding

Introduction

The meaning of natural language expressions has two important dimensions. On the one hand, meaning can be computed based on lexical and compositional semantics. On the other, meaning also relies on conceptual and world knowledge. This distinction broadly corresponds to that between *formal* linguistic competence—the ability to produce and understand expressions correctly based on their form—and *functional* competence, the ability to produce utterances that satisfy communicative goals and interpret them based on what we know about the world (Mahowald et al., 2024). When computing meaning, a natural or artificial system also connects linguistic symbols to the non-linguistic world, a challenge sometimes referred to as symbol grounding (Harnad, 1990). In AI, the extent to which Large Language Models (LLMs) are (or should be) able to connect natural language symbols to non-linguistic data remains a highly debated question (Bender & Koller, 2020; Bisk et al., 2020; Pavlick, 2023; Mandelkern & Linzen, 2024; Leivada, Marcus, Günther, & Murphy, 2024). Functional competence plays a role here, too. Consider determining whether a caption correctly describes one of the visual scenes in Figure 1. Both scenes involve a seated woman holding a cup in her hand. A description involving these details can be verified against an image by linking specific predicates to image regions. On the other hand, most English speakers would agree that only one of the images of seated women also involves the action of *begging*. Determining this involves reasoning about the woman's intentions, the social significance of her actions, and other visual cues in the image. As Mahowald et al. (2024) argue, reasoning with social and world knowledge is also part of our functional linguistic competence.

(a) A woman begs for money in the street. (A)
A woman sits in the street with a cup in her hand. (B)

(b) A woman sits in the street with a cup in her hand. (B)

Figure 1: Example from the CV-Probes dataset. While image 1a can be described by both a socially grounded (A) and physically grounded caption (B), only the latter (B) applies to image 1b.

In the field of multimodal learning, transformer-based models have significantly advanced the integration of visual and textual data in an attempt to partially address the grounding problem (Tan & Bansal, 2019; J. Li et al., 2021; Alayrac et al., 2022; Zeng, Zhang, & Li, 2022; J. Li, Li, Savarese, & Hoi, 2023; Liu, Li, Wu, & Lee, 2023; Dai et al., 2023; Deitke et al., 2024). Various benchmarks have been developed to assess such models' ability to ground fine-grained linguistic phenomena, including verbs (Hendricks & Nematzadeh, 2021; Beňová et al., 2024), counting (Parcalabescu, Gatt, Frank, & Calixto, 2020), spatial relations (Kamath, Hessel, & Chang, 2023), and word order (Thrush et al., 2022; Chen, Fernández, & Pezzelle, 2023; Ray et al., 2023; Ma et al., 2023).

For VLMs, verb phrases (VPs) pose a significant challenge. Studies have shown that they struggle with VP-centered tasks such as distinguishing cases involving swapping of verbs (*the woman fired/held the gun*) or arguments (*the woman shouted at the man* vs. *the man shouted at the woman*) (Parcalabescu et al., 2021; Pârcălăbescu, 2024). Similarly, video-language models have difficulty grounding expressions involving tem-

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

poral or situational elements, including VPs (Kesen et al., 2024; Xiao, Yao, Li, & Chua, 2024). Some of these results suggest a lack of formal competence (e.g., compositional interpretation of a verb and its arguments); others veer towards the functional (the difference between holding and firing in part relies on subtle visual cues and world knowledge).

This paper focuses on grounding VPs. We analyze VLM performance on image-text pairs like those in Figure 1. In each pair, the images have several visual features in common (e.g. the position of the subject in Figure 1, the extended arm holding a cup, the outdoor setting). These explicit, physical features support the use of a verb (like *sit*) for both images. On the other hand, certain other features (e.g. dress, type of cup, posture) warrant different inferences about the subject's intention and/or purpose and rely on social and world knowledge. For example, only the person in Figure 1a is likely to be *begging*. Our question is to what extent pretrained VLMs distinguish verbs like *sit* and *beg* by grounding them in the relevant features of the visual context, leveraging additional social knowledge to solve the grounding problem where relevant. To this end, we develop the **C**ontextual **V**erb-Probes (CV-Probes) dataset. We follow previous work that relies on the image-text matching capabilities of VLMs. However, this method has known limitations in revealing a model's grounding abilities (Beňová et al., 2024). Hence, we further rely on MM-SHAP (Parcalabescu & Frank, 2022), a performance-agnostic metric to quantify and interpret the contribution of individual tokens and modalities in VL models—to quantify the contribution of each input token (both visual and textual) to the model's predictions, with particular emphasis on the verbs in the captions.

Related Work

Vision-Language Models (VLMs) are typically trained on large datasets of image-text pairs (Changpinyo, Sharma, Ding, & Soricut, 2021; Schuhmann et al., 2021), and combine a visual backbone (e.g., vision transformers (Dosovitskiy et al., 2020)) with a textual encoder or decoder. Early VLMs, based on BERT (Devlin, Chang, Lee, & Toutanova, 2019), used single- or dual-stream architectures (Tan & Bansal, 2019; L. H. Li, Yatskar, Yin, Hsieh, & Chang, 2019; Lu, Batra, Parikh, & Lee, 2019; Bugliarello, Cotterell, Okazaki, & Elliott, 2021). Dual-encoders like CLIP (Radford et al., 2021) and ALIGN project visual and linguistic features into a common space. Recent models, such as Flamingo (Alayrac et al., 2022), BLIP2 (J. Li et al., 2023) and LlaVA-NeXT (Liu et al., 2024), use frozen visual and textual modules like LLMs linked by intermediate networks. Training objectives include masked or autoregressive language modeling, contrastive learning; and image-text matching. Instruction-tuned VLMs (Liu et al., 2023; Dai et al., 2023; OpenAI, 2024) extend these capabilities with interactive features.

Fine-grained benchmarks Several benchmarks study the abilities of VLMs to ground fine-grained linguistic phenomena and learn compositional language representations (Thrush et al., 2022; Ma et al., 2023; Kamath et al., 2023; Kamath, Hsieh, Chang, & Krishna, 2024). An important paradigm is foiling (Shekhar et al., 2017), whereby a dataset with images and corresponding texts ('positive' cases) is manipulated by changing essential parts of the caption, resulting in a *foil*, which is no longer true of the image. While CV-Probes is not strictly a foil-based benchmark, our experiments rely on the ability of a model to distinguish captions that differ in a specific linguistic phenomenon. In the same spirit, and closest to our work, are benchmarks with an explicit focus on verb phrase (VP) grounding, such as SVO-Probes (Hendricks & Nematzadeh, 2021) and parts of VALSE (Parcalabescu et al., 2021) and ViLMA (Kesen et al., 2024), which use a foiling-based method to probe VLM verb understanding, among other things. Focusing on fine-grained linguistic phenomena ultimately contributes to our broader understanding of VLMs' limits on grounded compositional reasoning tasks. Unlike previous work, CV-Probes explicitly focuses on VP interpretation in cases where there are both shared and distinct elements of the visual context.

Dataset

The CV-Probes dataset is designed with the following rationale. We start from images depicting actions that require reasoning about a person's intentions, based on a combination of visual context and social or world knowledge (caption A in Figure 1a). We refer to the verbs describing these images as 'socially grounded' (SG), since understanding these verbs requires reasoning about social factors and intentions, for which some visual features serve as cues. We contrast these verbs with alternatives that can describe the same images (caption B in Figure 1a) based on visual features, but which do not require reasoning about intentions beyond what is literally depicted. We refer to these as 'physically grounded' (PG). We further pair these image-caption pairs with additional images to which only the PG caption can apply (Figure 1b). Crucially, these images are visually similar to the ones depicting SG actions (both images in Figure 1 depict females seated outdoors). A VLM should assign a high image-text matching probability to both captions (A and B) for the image in Figure 1a, but only to the PG caption (B) for the image in Figure 1b. Abusing terminology somewhat, we will for the sake of brevity refer to images describable by captions containing socially grounded VPs as 'SG images' and to images not describable by such captions as 'PG images.'

Data selection and preprocessing

We leveraged the ImSitu dataset (Yatskar, Zettlemoyer, & Farhadi, 2016), a comprehensive resource for situation recognition tasks. ImSitu provided a foundation of 504 unique verbs and corresponding images, each encapsulating recognizable activities in their corresponding images. We selected 27 verbs that strongly rely on social and world knowledge to interpret them in a visual scene. The list of all SG verbs is in Table 1.

Table 1: Manually selected verbs for socially grounded (SG) captions from ImSitu dataset.

admire, apprehend, autograph, baptize, beg, brows, buy, celebrate, chase, cheerlead, coach, compete, congregate, count, educate, exercise, frisk, guard, hitchhike, hunt, interrogate, interview, pray, protest, race, study, vote

Table 2: CV-Probes dataset statistics

image-caption pairs	#
socially grounded pairs	117
physically grounded pairs	171

Each verb was paired with five different images from ImSitu. Verbs were further mapped to their FrameNet (Ruppenhofer, Ellsworth, Schwarzer-Petruck, Johnson, & Scheffczyk, 2016) frames. ImSitu specifies the roles or arguments of each verb; we crafted three SG descriptions per verb by mapping these arguments to the corresponding FrameNet template. These captions aimed to provide concise yet informative descriptions. We used ChatGPT 3.5 to make the captions more fluent. For consistency, all verbs are in the present tense and all articles are indefinite. We used the GRUEN pre-trained model (Zhu & Bhat, 2020) to further score the captions for grammaticality and selected the caption with the highest score as the SG caption for each image. The average score of all captions was 0.793 ± 0.017 (the top GRUEN score is 1), while the average score of the best-selected captions was 0.848 ± 0.021.

Physically grounded captions and images

Four independent annotators further annotated each image. The annotators were volunteers, non-native but fluent English speakers with a tertiary education. They were tasked with generating PG captions that reflected the depicted scenario. Annotators were instructed to describe what they saw, avoiding captions relying on world knowledge and inference. To facilitate this, they were explicitly provided with the SG captions and asked to avoid the SG verb in their new captions. Annotations were collected using the Doccano (Nakayama, Kubo, Kamura, Taniguchi, & Liang, 2018) platform; we had 20 volunteers. Subsequently, annotations underwent a review process by two of the authors to ensure descriptive fidelity. Annotations that used SG verbs in violation of the annotator instructions were removed. This iterative refinement process resulted in a curated dataset comprising one SG caption and one or more descriptive PG captions for each image.

We paired the PG captions with new images using the CLIP model (Radford et al., 2021) to perform image retrieval from the LAION dataset (Schuhmann et al., 2021). Given a PG caption, we retrieved image candidates and manually checked that the images corresponded only to the PG captions, but not the SG ones (Figure 1b).

Caption simplification

CV-Probes focuses on model's ability to match captions to images based on verb phrases. However, captions may contain additional information, often in syntactic arguments and adjuncts, which models may rely on during image-text matching. Thus, we manually created a simplified version of the captions, consisting exclusively of simple declarative sentences with a subject, verb, and object, but no additional modifiers (such as 'in the street' in Figure 1). Table 2 gives the statistics for the final CV-Probes dataset.

Human validation

We performed a second round of human validation on the simplified captions only, crowd-sourcing annotations via the Prolific platform,[1] with native speakers of English. Annotations were done over two separate sessions. In the first one, each annotator was presented with two images (SG and PG) and one caption (either SG or PG), and they were asked to choose whether the caption applies to both images, only one of them or neither. Three annotators evaluated each triplet. 6 annotators participated in this study. For triplets where at least 2 annotators chose different answers than the presumed true label, we manually improved the caption. The annotators were compensated by 9£ per hour of work. The resulting filtered dataset was validated in a second session to obtain human judgments against which we compare AI models. Each image-caption pair was evaluated by 3 annotators (24 in total). A further example of from the dataset can be seen in 2.

(a) Some people pray. (A) A group of men stand with crossed arms. (B)

(b) A group of men stand with crossed arms. (B)

Figure 2: Additional example from the CV-Probes dataset. While image 2a can be described by both a socially grounded (A) and a physically grounded caption (B), only the latter (B) applies to image 2b.

Model Performance on CV-Probes

To evaluate performance on the CV-Probes dataset, we leverage the capability of pretrained VLMs to perform image-text matching. For VLMs with an image-text matching head, we treat the task as a binary classification: given an image-text pair $\langle I, T \rangle$, the task is to determine the probability that T describes I. For a matching pair, a model response is considered correct if it assigns $p(1|I,T) > p(0|I,T)$, where 1 indicates a positive match. Similarly, for a non-matching pair, a model

[1] https://www.prolific.com/

is correct if $p(0|I,T) > p(1|I,T)$. We evaluate each image-caption pair separately. We do not explicitly compare matching and non-matching pairs, since (a) for SG images, both SG and PG captions apply; and (b) we consider our evaluation strategy as ultimately very lenient, since we only consider whether the probability of a correct match exceeds that of an incorrect match.

Our evaluation involved several models that are representative of different neural architectures. We introduce them briefly here. LXMERT (Tan & Bansal, 2019) is a dual-stream architecture that integrates visual features extracted using a Faster R-CNN visual backbone with linguistic features extracted using a BERT architecture. ALBEF (J. Li et al., 2021) encodes visual features with a vision transformer and linguistic features with the first six layers of BERT, using the last six BERT layers for multimodal fusion via cross-attention. BLIP (J. Li, Li, Xiong, & Hoi, 2022) employs a visual transformer and a BERT language model, integrating these through a Multimodal Mixture of Encoder-Decoder (MED) architecture. In contrast, BLIP2 (J. Li et al., 2023) introduces a Querying Transformer (Q-Former) to bridge the gap between a frozen image encoder and a frozen large language model (LLM). FLAVA (Singh et al., 2022) comprises a visual and textual transformer, each trained on modality-specific data before being integrated through a shared multimodal encoder. LLaVA-NeXT (Liu et al., 2024) uses a pretrained visual backbone with a large language model decoder, with pretraining stages including image-text matching and image-to-text generation. We also compare the results of these models with GPT-4o (OpenAI, 2024), a state-of-the-art model whose weights and architecture are not available. Both LLaVa-NeXT and GPT-4o were evaluated using a prompt asking whether a given caption matches an image. Hence, these models generate a response (match or non-match), while the other models explicitly compute the probability of an image-text match. Since GPT-4o is a closed model, we only report performance results for comparison purposes, reserving the in-depth analysis for other models.

Results

We report model accuracy on both the original (Table 3) and the simplified captions (Table 4) in CV-Probes.

In Table 3, the highest accuracy for matching pairs was obtained by FLAVA. For SG images, FLAVA is at or close to the maximum performance (98.29% and 87.13% for SG and PG captions respectively). For matching PG images and captions, it reaches an accuracy of 95.91%, surpassing all other models except the decoder-based LlaVA-NeXT in this regard. Conversely, LXMERT, ALBEF, and BLIP2 exhibited comparatively lower accuracy, and usually below chance (LXMERT at 26.50%) to 78.6% (BLIP) for images depicting context-dependent actions.

However, the most important results are in column 3, providing the accuracy for pairs of SG captions paired with PG images. By construction, these are non-matching cases. Here, FLAVA's performance drops significantly to below chance levels. This suggests a positive bias in the model and a tendency to assign high probabilities to image-text pairs, even in non-matching cases. Similar observations are made for specific models in SVO-Probes by (Hendricks & Nematzadeh, 2021). On the other hand, the best performance in this category is achieved by BLIP2 with an accuracy of 89.47%. We note that BLIP2 performance also exceeds that of LlaVA-NeXT and GPT-4o on these non-matching cases.

The same trends are broadly seen in Table 4, which contains results for the simplified captions. We observe that simplification sometimes causes a model to change its prediction. One possibility is that the process of simplification, which, for example, removes adjunct phrases in the captions (such as 'on the street' in Figure 1), leaves models with less signal in the textual modality to boost the probability of a match. This, in turn, may suggest that models rely on the non-verb parts of the captions more, an issue we return to in the following section.

A specific case is the evaluation of the LLaVA-NeXT and GPT-4o. LLaVA-NeXT performs very well on matching pairs, but its performance on non-matching pairs is close to random guessing and shows a significant drop for simplified captions (compare column 3, Tables 3 and 4). The performance of GPT-4o is however very close to human performance, lacking mostly for matching pairs. However, as these models are generative, we can not compare the results with the rest of the models that were evaluated using image-text matching and we also can not use MM-SHAP evaluation as we do in the next section with the rest of the models.

It is worth considering possible reasons why BLIP2 performs much better than other models, including FLAVA, on the key condition. One might be architectural: while all other models rely on a single token in the image-text matching head (usually, a `CLS` or `Encode` token), BLIP2 has 32 learnable query embeddings as input to the query transformer. The queries interact with each other through self-attention layers and interact with frozen image features through cross-attention layers. Queries are projected via a linear classifier, and the resulting logits across all queries are averaged to get the matching score. Another distinguishing feature of BLIP2 is that it is pretrained with image-grounded text generation. Finally, BLIP2 differs from other models with a vision transformer visual backbone in that it uses ViT-g as an image encoder, compared to ViT-B or ViT-L used in other models. As we noted, FLAVA exhibits an overall positive bias and fails in those cases where a SG caption does not match an image. FLAVA has unimodal backbones for vision and language, and incorporating unimodal pretraining (as well as other losses) improves the performance on vision tasks, NLP tasks, and multimodal tasks (Singh et al., 2022). On the other hand, image-text matching in FLAVA did not involve mining for (hard) negatives. This may underlie the positive bias: the model may predict a match based on, for example, only matching entities ('woman') or locations ('street').

In Table 5, we present a Pearson correlation between

Table 3: Image-text matching accuracy (%) on the original captions. SG: socially grounded; PG: physically grounded.

Images:	**SG (cf. Fig 1a)**		**PG (cf. Fig. 1b)**		
Captions:	*SG*	*PG*	*SG*	*PG*	Harmonic mean
LXMERT	26.50	37.79	83.63	41.52	39.90
ALBEF	63.25	70.18	74.85	92.98	73.84
BLIP	78.63	75.58	78.36	92.40	**80.75**
BLIP 2	62.39	55.23	**89.47**	80.70	69.32
FLAVA	**98.29**	**87.13**	22.22	**95.91**	51.89
LlaVA-NeXT	94.87	95.91	61.63	98.84	84.55
GPT-4o	88.89	86.55	85.38	87.13	86.97

Table 4: Image-text matching accuracy (%) on the simplified captions. SG: socially grounded; PG: physically grounded.

Images:	**SG (cf. Fig 1a)**		**PG (cf. Fig. 1b)**		
Captions:	*SG*	*PG*	*SG*	*PG*	Harmonic mean
LXMERT	29.91	41.86	77.78	38.60	41.63
ALBEF	51.28	53.80	71.93	78.36	61.77
BLIP	75.21	59.88	74.27	84.80	**72.39**
BLIP 2	47.01	36.63	**91.23**	58.48	52.20
FLAVA	**96.58**	**88.89**	12.87	**94.15**	36.39
LlaVA-NeXT	94.02	97.66	43.60	100.00	74.33
GPT-4o	91.45	91.81	78.36	94.15	88.48
Human	98.29	97.66	79.53	92.40	91.31

Table 5: Pearson correlation between human annotations and model classification in CV-Probes. Non-significant correlations (at $p \geq 0.05$) are in italics.

Images:	**SG (cf. Fig 1a)**		**PG (cf. Fig. 1b)**	
Captions:	*SG*	*PG*	*SG*	*PG*
LXMERT	0.248	*0.04*	0.209	*0.104*
ALBEF	0.187	0.2	*0.143*	*0.093*
BLIP	*0.116*	0.267	0.181	*0.144*
BLIP 2	*0.149*	0.201	*0.142*	0.167
FLAVA	*-0.006*	*0.044*	*0.049*	*0.036*
LlaVA-NeXT	*0.098*	*-0.086*	*0.023*	*-0.073*

the probabilities of image-caption matches calculated by the models and the votes obtained by human annotations in our second round (see *Dataset* above). While almost all correlations are positive, most have a low coefficient, and many are not significantly different from 0. Thus, VLM predictions do not correlate strongly with how humans evaluate image-caption matches in CV-Probes. Unfortunately we can not calculate correlation between human prediction score and model probability for GPT-4o as logits are not readily accessible via the GPT-4o multimodal API.

Do Models Rely on Verbs?

In this section, we focus on token-level explanations for the image-text matching results. In particular, we ask to what extent models rely on verb tokens to compute an image-text matching probability. For this purpose, we use MM-SHAP (Parcalabescu & Frank, 2022), an adaptation of SHAP (Lundberg & Lee, 2017), which is an approximation of Shapley values (Shapley et al., 1953). Shapley values are a game-theoretic formalism to compute the contributions of individual players in cooperative games, given a numerical outcome or score. For a given player, the idea is to compare the contribution of each possible coalition of players when that player is included versus when they are not. For AI model explanations, we consider input features to be the 'players'. Since the number of possible coalitions in a Shapley game is exponential, approximations such as SHAP provide a compute-optimal solution. Crucially, SHAP explanations are model-agnostic, making it possible to compare them across models.

MM-SHAP is designed for a VLM with n_T text and n_I image tokens.[2] The textual ϕ_T and the image contribution ϕ_I towards a prediction are defined as the sum of the absolute SHAP values of textual and visual tokens, respectively:

$$\phi_T = \sum_{j}^{n_T} |\phi_j| \quad ; \quad \phi_I = \sum_{j}^{n_I} |\phi_j| \tag{1}$$

MM-SHAP provides one score per modality, defined as the proportion of the overall SHAP total due to that modality:

$$T-SHAP = \frac{\phi_T}{\phi_T + \phi_I}; \quad V-SHAP = \frac{\phi_I}{\phi_T + \phi_I} \tag{2}$$

This formulation offers insight into potential unimodal collapse, whereby a model relies extensively on one modality to the detriment of the other (Hessel & Lee, 2020; Frank, Bugliarello, & Elliott, 2021; Parcalabescu & Frank, 2022).

[2]Image tokens are defined as superpixels or image patches.

Table 6: Average T-SHAP scores for the caption (overall) and average for the verb only, with proportion of overall SHAP attribution for the verb for 50 samples. Scores are shown separately for the case where the model predicts that the image and caption match ($p > 0.5$) and do not match. SG: socially grounded; PG: physically grounded.

		BLIP				BLIP2				FLAVA			
	Images:	**SG (cf. Fig 1a)**		**PG (cf. Fig 1b)**		**SG (cf. Fig 1a)**		**PG (cf. Fig 1b)**		**SG (cf. Fig 1a)**		**PG (cf. Fig 1b)**	
	Captions:	*SG*	*PG*	*SG*	*PG*	*SG*	*PG*	*SG*	*PG*	*SG*	*PG*	*SG*	*PG*
Match	*Overall (%)*	89.44	85.41	88.47	88.46	77.29	70.85	65.21	70.75	45.65	45.39	45.87	45.78
	Verb	0.3790	0.1596	**0.3047**	0.1586	0.4266	0.1681	**0.1294**	0.0877	0.1236	0.0871	**0.1521**	0.0999
	Verb (%)	47.10	21.26	42.18	20.84	60.31	22.50	32.14	15.96	32.28	21.33	36.22	25.93
Non-match	*Overall (%)*	75.20	77.39	72.64	81.58	60.99	56.46	56.10	52.24	46.66	47.73	45.42	52.83
	Verb	0.1228	0.0119	**-0.0323**	0.1586	0.1381	-0.0174	**-0.0131**	0.0322	0.0592	0.0273	**0.1394**	0.2555
	Verb (%)	24.92	16.47	35.83	20.48	24.44	20.07	20.21	13.07	18.12	11.38	21.88	24.66

Results

For reasons of space, we report results on a selection of models, focusing primarily on BLIP, which obtained the highest harmonic mean overall conditions in Tables 3 and 4, as well as BLIP2 and FLAVA. Table 6 reports results for the three models computed over 50 samples from CV-Probes. We use simplified captions to reduce the chance that attributions arise from text tokens that do not belong to the VP or its arguments. We report overall T-SHAP values and the contribution of the individual verb token towards match prediction. In addition, we estimate the percentage of the overall T-SHAP score that is due to the verb. We take this as an indicator of the overall importance of the verb token in accounting for the model predictions.

The T-SHAP values for BLIP suggest that this model relies heavily on the textual modality, far more than on the visual modality, raising the possibility of unimodal collapse (T-SHAP $\gg 50\%$). This suggests the model exploits text biases, to some extent reducing to a unimodal model for this task. Notably, the verb contribution for PG images with SG captions should lead to a non-match prediction. However, BLIP shows a high average verb attribution (0.3047) when incorrectly predicting a match, but the verb attribution is almost zero (-0.0323) when correctly predicting a non-match. This implies that correct predictions are not influenced by verb tokens, whereas incorrect predictions are.

The BLIP2 and FLAVA models are more evenly balanced between visual and textual modalities. For FLAVA in particular, T-SHAP overall is close to 50%, perhaps unsurprisingly, given its architecture. For BLIP2, verb attributions are high for match predictions (0.1294) and close to zero for non-match predictions (-0.0131) in the key condition, indicating that verb tokens do not play a correct role in image-text matching task. For FLAVA, the second highest average verb attributions occur with PG images with SG captions. This suggests the model has a poor grounding ability concerning SG verbs: these contribute *positively* overall to both match and non-match predictions in this non-matching context.

Based on these results, we conclude that these models struggle to ground SG and PG verb phrases effectively.

Conclusion

The ability to ground linguistic symbols in the non-linguistic world of perception and experience relies on both lexico-syntactic knowledge (part of our formal linguistic competence) and world knowledge (part of functional competence). World knowledge, in particular, mediates our ability to describe scenes and actions beyond their literal composition, by recognizing actions and intentions. In this sense, the ability to ground descriptions of actions and interpret verb phrases relies on social and world knowledge to different degrees.

In this study, we evaluated the ability of several vision-language (VL) models to ground socially grounded ('beg', 'pray') and physically grounded ('sit', 'stand') verb phrases within the context of image-text matching tasks. One of our primary contributions is a novel test dataset, CV-Probes, which pairs images with descriptions that rely on social and world knowledge to different degrees, with an emphasis on verb phrase grounding.

We find that VL models exhibit varying abilities to ground SG captions describing activities to visual data and tend to fail on the critical test case, where a caption with a SG verb phrase is paired with an image that does not depict such an action but is visually similar to its SG counterpart.

We further analyzed models in a performance-agnostic way using Shapley values, focusing specifically on the contribution of verb tokens to the model's predictions. The results indicated that models struggled with grounding SG verb phrases. For BLIP2, verb contributions were minimal for the non-match pairs (PG image with SG caption), indicating that the model does not consider verbs as significant predictors. The FLAVA model incorrectly attributed high importance to SG verbs in PG image scenarios. Even large decoder VLMs like LLaVA-NeXT show significant shortcomings. The closest performance to human annotation is obtained by GPT-4o, however we can not study the results of this model in depth as the model's weights are not publicly available.

Overall, our findings highlight that current publicly available VL models, including those analyzed in this study, have significant room for improvement in integrating social and world knowledge information in grounding verb phrases that describe scenarios of different kinds. These results underscore the need for advanced methodologies in training and evaluating VL models to enhance their ability to process and ground context accurately and integrate world knowledge, leading to more robust and reliable performance.

Acknowledgements

This research was partially supported by *DisAI - Improving scientific excellence and creativity in combating disinformation with artificial intelligence and language technologies*, a project funded by Horizon Europe under GA No. 101079164; by the *MIMEDIS*, a project funded by the Slovak Research and Development Agency under GA No. APVV-21-0114. The collaboration was facilitated by the Multi3Generation COST Action CA18231. The authors thank their anonymous reviewers for constructive comments on an earlier version of this paper.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... Simonyan, K. (2022). Flamingo: a Visual Language Model for Few-Shot Learning. In *Proceedings of the 6th Conference on Neural Information Processing Systems (NeurIPS 2022)*. Retrieved from http://arxiv.org/abs/2204.14198 (arXiv: 2204.14198)

Bender, E. M., & Koller, A. (2020). Climbing towards NLU : On Meaning , Form , and Understanding in the Age of Data. In *Proceedings ofthe 58th Annual Meeting ofthe Association for Computational Linguistics (ACL'20)* (pp. 5185–5198). Retrieved from https://www.business2community.com/s

Beňová, I., Košecká, J., Gregor, M., Tamajka, M., Veselỳ, M., & Šimko, M. (2024). Beyond image-text matching: Verb understanding in multimodal transformers using guided masking. *arXiv preprint arXiv:2401.16575*.

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., ... Turian, J. (2020). Experience Grounds Language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP'20)* (Vol. 2004.10151, pp. 8718–8735). Online: Association for Computational Linguistics. (arXiv: 2004.10151) doi: 10.18653/v1/2020.emnlp-main.703

Bugliarello, E., Cotterell, R., Okazaki, N., & Elliott, D. (2021). 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.

Changpinyo, S., Sharma, P., Ding, N., & Soricut, R. (2021). Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 3558–3568).

Chen, X., Fernández, R., & Pezzelle, S. (2023, December). The BLA Benchmark: Investigating Basic Language Abilities of Pre-Trained Multimodal Models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 5817–5830). Singapore: Association for Computational Linguistics. Retrieved 2024-03-25, from https://aclanthology.org/2023.emnlp-main.356 doi: 10.18653/v1/2023.emnlp-main.356

Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., ... Hoi, S. (2023, June). *InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning*. arXiv. Retrieved 2024-05-08, from http://arxiv.org/abs/2305.06500 (arXiv:2305.06500 [cs]) doi: 10.48550/arXiv.2305.06500

Deitke, M., Clark, C., Lee, S., Tripathi, R., Yang, Y., Park, J. S., ... Kembhavi, A. (2024, September). *Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Multimodal Models*. arXiv. Retrieved 2024-10-02, from http://arxiv.org/abs/2409.17146 (arXiv:2409.17146 [cs]) doi: 10.48550/arXiv.2409.17146

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings ofNAACL-HLT 2019,*. Minneapolis, MN: Association for Computational Linguistics. Retrieved from http://arxiv.org/abs/1810.04805 (arXiv: 1810.04805 ISSN: 0140-525X) doi: arXiv:1811.03600v2

Dosovitskiy, A., Beye, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*, *2010.11929*.

Frank, S., Bugliarello, E., & Elliott, D. (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 (EMNLP'21)*. Retrieved from http://arxiv.org/abs/2109.04448 (arXiv: 2109.04448)

Harnad, S. (1990). The symbol grounding problem. *Physica*, *D42*(1990), 335–346. Retrieved 2014-05-12, from http://www.sciencedirect.com/science/article/pii/0167278990900876

Hendricks, L. A., & Nematzadeh, A. (2021). Probing image-language transformers for verb understanding. *arXiv preprint arXiv:2106.09141*.

Hessel, J., & Lee, L. (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'20)* (pp. 861–877). Online: Association for Computational Linguistics. (arXiv: 2010.06572) doi: 10.18653/v1/2020.emnlp-main.62

Kamath, A., Hessel, J., & Chang, K.-W. (2023, October). *What's "up" with vision-language models? Investigating their struggle with spatial reasoning*. arXiv. Retrieved 2024-04-16, from http://arxiv.org/abs/2310.19785 (arXiv:2310.19785 [cs]) doi: 10.48550/arXiv.2310.19785

Kamath, A., Hsieh, C.-Y., Chang, K.-W., & Krishna, R. (2024). The Hard Positive Truth about Vision-Language Compositionality. In *Proceedings of the european conference on computer vision (eccv)*. Retrieved from http://arxiv.org/abs/2409.17958

Kesen, I., Pedrotti, A., Dogan, M., Cafagna, M., Acikgoz, E. C., Parcalabescu, L., ... Erdem, E. (2024). Vilma: A

zero-shot benchmark for linguistic and temporal grounding in video-language models. In *Proceedings of the 12th international conference on learning representations (iclr)*. Vienna, Austria.

Leivada, E., Marcus, G., Günther, F., & Murphy, E. (2024, September). *A Sentence is Worth a Thousand Pictures: Can Large Language Models Understand Hum4n L4ngu4ge and the W0rld behind W0rds?* arXiv. Retrieved 2024-10-24, from http://arxiv.org/abs/2308.00109 (arXiv:2308.00109) doi: 10.48550/arXiv.2308.00109

Li, J., Li, D., Savarese, S., & Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning* (pp. 19730–19742).

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning* (pp. 12888–12900).

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., & Hoi, S. C. H. (2021). Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, *34*, 9694–9705.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., & Chang, K.-w. (2019). VisualBERT: A simple and performant baseline for vision and language. *ArXiv preprint 1908.03557*. doi: 10.1007/s11159-020-09831-4

Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., & Lee, Y. J. (2024, January). *Llava-next: Improved reasoning, ocr, and world knowledge.* Retrieved from https://llava-vl.github.io/blog/2024-01-30-llava-next/

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023, April). Visual Instruction Tuning. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*. New Orleans, LA, USA: arXiv. Retrieved 2023-10-20, from http://arxiv.org/abs/2304.08485 (arXiv:2304.08485 [cs]) doi: 10.48550/arXiv.2304.08485

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)* (pp. 1–11). Vancouver, BC. Retrieved from http://arxiv.org/abs/1908.02265 (arXiv: 1908.02265)

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Ma, Z., Hong, J., Gul, M. O., Gandhi, M., Gao, I., & Krishna, R. (2023). CREPE: Can Vision-Language Foundation Models Reason Compositionally? In (pp. 10910–10921). Retrieved 2023-10-08, from https://openaccess.thecvf.com/content/CVPR2023/html/Ma_CREPE_Can_Vision-Language_Foundation_Models_Reason_Compositionally_CVPR_2023_paper.html

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024, March). *Dissociating language and thought in large language models.* arXiv. Retrieved 2024-10-31, from http://arxiv.org/abs/2301.06627 (arXiv:2301.06627) doi: 10.48550/arXiv.2301.06627

Mandelkern, M., & Linzen, T. (2024, September). Do Language Models' Words Refer? *Computational Linguistics*, *50*(3), 1191–1200. Retrieved 2024-10-24, from https://direct.mit.edu/coli/article/50/3/1191/121670/Do-Language-Models-Words-Refer doi: 10.1162/coli_a_00522

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., & Liang, X. (2018). *doccano: Text annotation tool for human.* Retrieved from https://github.com/doccano/doccano (Software available from https://github.com/doccano/doccano)

OpenAI. (2024). *Gpt-4o.* https://openai.com/index/gpt-4o. (Accessed: 2025-05-08)

Parcalabescu, L., Cafagna, M., Muradjan, L., Frank, A., Calixto, I., & Gatt, A. (2021). Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. *arXiv preprint arXiv:2112.07566*.

Parcalabescu, L., & Frank, A. (2022). Mm-shap: A performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks. *arXiv preprint arXiv:2212.08158*.

Parcalabescu, L., Gatt, A., Frank, A., & Calixto, I. (2020). Seeing past words: Testing the cross-modal capabilities of pretrained v&l models on counting tasks. *arXiv preprint arXiv:2012.12352*.

Pavlick, E. (2023, July). Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A*. Retrieved 2024-10-24, from https://royalsocietypublishing.org/doi/10.1098/rsta.2022.0041 (Publisher: The Royal Society) doi: 10.1098/rsta.2022.0041

Pârcălăbescu, L. (2024). *Measuring the Contributions of Vision and Text Modalities in Multimodal Transformers*. PhD Thesis, University of Heidelberg, Germany.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Ray, A., Radenovic, F., Dubey, A., Plummer, B. A., Krishna, R., & Saenko, K. (2023, November). *COLA: A Benchmark for Compositional Text-to-image Retrieval.* arXiv. Retrieved 2024-02-02, from http://arxiv.org/abs/2305.03689 (arXiv:2305.03689 [cs]) doi: 10.48550/arXiv.2305.03689

Ruppenhofer, J., Ellsworth, M., Schwarzer-Petruck, M., Johnson, C. R., & Scheffczyk, J. (2016). *Framenet ii: Extended theory and practice* (Tech. Rep.). International Computer Science Institute.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., ... Komatsuzaki, A. (2021). Laion-

400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Shapley, L. S., et al. (1953). A value for n-person games.

Shekhar, R., Pezzelle, S., Klimovich, Y., Herbelot, A., Nabi, M., Sangineto, E., & Bernardi, R. (2017). Foil it! find one mismatch between image and language caption. *arXiv preprint arXiv:1705.01359*.

Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., & Kiela, D. (2022). Flava: A foundational language and vision alignment model. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 15638–15650).

Tan, H., & Bansal, M. (2019). Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., & Ross, C. (2022). Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 5238–5248).

Xiao, J., Yao, A., Li, Y., & Chua, T.-S. (2024, June). Can I Trust Your Answer? Visually Grounded Video Question Answering. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 13204–13214). Retrieved 2025-04-09, from `https://ieeexplore.ieee.org/document/10655740` (ISSN: 2575-7075) doi: 10.1109/CVPR52733.2024.01254

Yatskar, M., Zettlemoyer, L., & Farhadi, A. (2016). Situation recognition: Visual semantic role labeling for image understanding. In *Conference on computer vision and pattern recognition.*

Zeng, Y., Zhang, X., & Li, H. (2022, June). Multi-Grained Vision Language Pre-Training: Aligning Texts with Visual Concepts. In *Proceedings of the 39th International Conference on Machine Learning* (pp. 25994–26009). PMLR. Retrieved 2024-02-01, from `https://proceedings.mlr.press/v162/zeng22c.html` (ISSN: 2640-3498)

Zhu, W., & Bhat, S. (2020). GRUEN for evaluating linguistic quality of generated text. *arXiv*, *2010.02498*. (arXiv: 2010.02498) doi: 10.18653/v1/2020.findings-emnlp.9