DEVBENCH: A multimodal developmental benchmark for language learning

Alvin W. M. Tan, Sunny Yu,
Bria Long, Wanjing Anya Ma, Tonya Murray,
Rebecca D. Silverman, Jason D. Yeatman, Michael C. Frank
Stanford University
{tanawm, syu03, bria, wanjingm, tonyamur,
rdsilver, jyeatman, mcfrank}@stanford.edu

Abstract

How (dis)similar are the learning trajectories of vision–language models and children? Recent modeling work has attempted to understand the gap between models' and humans' data efficiency by constructing models trained on less data, especially multimodal naturalistic data. However, such models are often evaluated on adult-level benchmarks, with limited breadth in language abilities tested, and without direct comparison to behavioral data. We introduce DEVBENCH, a multimodal benchmark comprising seven language evaluation tasks spanning the domains of lexical, syntactic, and semantic ability, with behavioral data from both children and adults. We evaluate a set of vision–language models on these tasks, comparing models and humans on their response patterns, not their absolute performance. Across tasks, models exhibit variation in their closeness to human response patterns, and models that perform better on a task also more closely resemble human behavioral responses. We also examine the developmental trajectory of OpenCLIP over training, finding that greater training results in closer approximations to adult response patterns. DEVBENCH thus provides a benchmark for comparing models to human language development. These comparisons highlight ways in which model and human language learning processes diverge, providing insight into entry points for improving language models.

1 Introduction

Humans are remarkably good language learners, acquiring language rapidly in the first few years of life without much explicit supervision. Recent machine learning models are also able to acquire some aspects of human language, although they require much larger quantities of language input than a typical human would receive [1, 2]. This "data gap" reflects differences in the learning processes and mechanisms employed by human language learners and language models; understanding the nature of these differences is pivotal for the joint goals of building better cognitive models of language learning and building more data-efficient language models.

Recent work has attempted to bridge this data gap by constructing models that are trained on less data, including on naturalistic data from children [2–5]. However, these models have typically been evaluated using benchmarks that reflect adult human performance, whether explicitly (i.e., with comparison data collected from adult participants) or implicitly (i.e., high accuracy on some recognition task). Such evaluation methods are not appropriate to the goal of understanding whether developmentally realistic data leads to human-like learning. We would not expect child performance to be similar to adult performance on these tasks for various reasons related to both language competence (i.e., children's vocabulary knowledge) and to their co-developing cognitive abilities

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

 https://doi.org/10.52202/079017-2462

(e.g., working memory limitations). Instead, it is crucial to evaluate models on benchmarks that can indicate whether the language ability gained by machine learning models matches the language ability gained by children when exposed to similar developmental data.

Research in child language acquisition has observed that children learn different aspects of language at different rates, and the acquisition of different levels of representation also interact in complex and nonlinear ways [6–10]. Thus, in order to characterise models' language learning performance, we should evaluate multiple levels of linguistic representation, including the lexicon, syntax, and semantics – ideally how these correspond to children's development at different ages.

Additionally, to conduct methodologically rigorous cognitive evaluation of machine learning models, it is important to compare human and model performance directly (rather than assuming that humans will perform well), and to ensure that evaluation conditions are as similar as possible for models and humans [11–13]. Notably, many language evaluations with infants and toddlers are multimodal in nature, because this method permits the measurement of language comprehension without the additional working memory and motor control demands of language production. Young children can use nonverbal response methods such as looking or pointing, which can nonetheless reflect their language abilities. The incorporation of multimodality has an additional advantage of introducing grounding, which has been suggested as a possible solution to the data gap [1].

These observations provide a set of desiderata for a developmentally appropriate evaluation of language models:

1. Wide dynamic range of difficulty
2. Multiple levels of linguistic representations
3. Corresponding data from children
4. High similarity in evaluation method between models and humans

To create a benchmark that satisfies these desiderata, we introduce **DEVBENCH**, a multimodal benchmark of language evaluation tasks. DEVBENCH includes a suite of seven tasks measuring lexical, syntactic, and semantic ability, with human data from both children and adults. Notably, the primary evaluation metric reflects models' similarity to human response patterns, rather than absolute performance levels, allowing us to capture finer-grained details about human–model similarity.

We evaluate a set of vision–language models on DEVBENCH, including not only state-of-the-art models but also smaller models and models trained on developmentally realistic data. We additionally investigate how DEVBENCH performance varies over model training by evaluating intermediate checkpoints of OpenCLIP [14]. Evaluation results suggest that current vision–language models exhibit variation in their human-likeness, suggesting further areas of research to develop language models that more closely approximate human language learning.

2 Related work

2.1 Multimodal benchmarks

Since the advent of multimodal models, various benchmarks have been developed to evaluate their performance [15]. A majority of these benchmarks primarily involve visual question answering [16–21]; other tasks include image captioning [18, 22], prediction [23], and retrieval [17, 22].

Most tasks in these multimodal benchmarks probe models' perception, reasoning, and knowledge, evaluating models on domains including action prediction, counting, relational reasoning, or optical character recognition. Correct responses to these tests require the conjunction of many skills – notably, all reasoning tasks also require perception, and perception tasks in turn broadly require object knowledge. Furthermore, most multimodal benchmarks focus on models' visual understanding, and no existing benchmark focuses specifically on the linguistic abilities of multimodal models.

2.2 Developmentally inspired evaluation

Some recent work in machine learning evaluation has used benchmarks inspired by developmental psychology and cognitive science. For example, the Large Language Model Response Score (LRS) [24] draws from key experiments in the child development literature to construct question answering tasks for large language models, though it does not involve comparison to data from

Table 1: Characteristics of multimodal and developmentally inspired benchmarks.

	Evaluation features				Domains			Human data	
Benchmark	Multi-modal	Zero-shot	Nonverbal response	Develop-mental	Lexicon	Syntax	Semantics	Children	Adult
LAMM [18]	✓	✓							
MultiBench [23]	✓	✓							
GEM [22]	✓	✓	✓						
MMBench [20]	✓	✓	✓						
SEED-Bench [28]	✓	✓	✓		✓		✓		
MME [19]	✓	✓	✓		✓		✓		
Perception Test [21]	✓	✓	✓		✓		✓		
M3Exam [29]	✓	✓		✓					✓
LRS [24]		✓		✓					
InfLevel [25]		✓	✓	✓					
Zorro [3]		✓	✓	✓		✓			
MEWL [26]	✓		✓	✓	✓		✓		✓
ModelVsBaby [27]	✓	✓	✓	✓	✓			✓	
DevBench	✓	✓	✓	✓	✓	✓	✓	✓	✓

children. In addition, most of its tasks were originally multimodal (with visual stimuli and verbal prompts), and it is not clear how the translation into a unimodal verbal task would affect human performance. In the visual domain, the Infant-Level Physical Reasoning Benchmark (InfLevel) [25] aimed to evaluate video models on physical reasoning, drawing from classic violation of expectation tasks in the infant cognition literature related to the principles of continuity, solidity, and gravity. However, this benchmark also did not have a direct comparison with human data. The Machine Word Learning (MEWL) benchmark [26] is a multimodal evaluation that builds upon hypothesised mechanisms of few-shot word learning in children, and requires the inference of the meaning of novel words from a set of visual scenes with referentially ambiguous labels. The benchmark contains data from adults, but their poor performance on some subsets of these trials (e.g., those requiring pragmatic implicature) suggests that children might find these tasks very difficult. More recently, the ModelVsBaby benchmark [27] has assessed out of distribution object recognition with corresponding data from 2-year-olds, providing a first step towards direct developmental model–human comparison.

Other work assessing developmental models has also developed ad-hoc evaluations. The most common method has been to design evaluation sets that resemble other machine learning model tasks (e.g., grammatical acceptability, image classification) while restricting the domain to a developmentally relevant domain (i.e., only including vocabulary familiar to the model, or choosing image categories that are child-relevant). This method has been applied both to language models (e.g., the Zorro benchmark on BabyBERTa models [3]), and to multimodal models (e.g., the Konkle Objects evaluation on CVCL models [4]). However, again in these cases there are typically no child data, and the assumption that children would perform well in such evaluations is not directly evaluated. We summarise the characteristics of a range of multimodal and developmental benchmarks in Table 1.

2.3 Model learning trajectory analyses

A few studies have used developmental approaches to analyse model learning trajectories. Chang and Bergen [30] examined the age of acquisition of different words by measuring the change in mean surprisal of a word over training epochs, finding dissimilarities in the predictors of age of acquisition in models and children. Evanson et al. [31] examined the age of acquisition of different syntactic structures by measuring the point at which models began preferring the grammatical sentence in a grammatical acceptability task. We extend these approaches here by examining training trajectories across multiple levels of linguistic representation.

3 DevBench description

DevBench contains seven tasks across lexical, syntactic, and semantic domains. Each task is accompanied by item-level human data so that full human response distributions can be compared to model scores. The lexical tasks measure vocabulary knowledge, operationalised as the ability to correctly pick out the visual referent of a noun label. The syntactic tasks measure grammatical knowledge, operationalised as the ability to correctly pick out a scene containing the correct relations

https://doi.org/10.52202/079017-2462

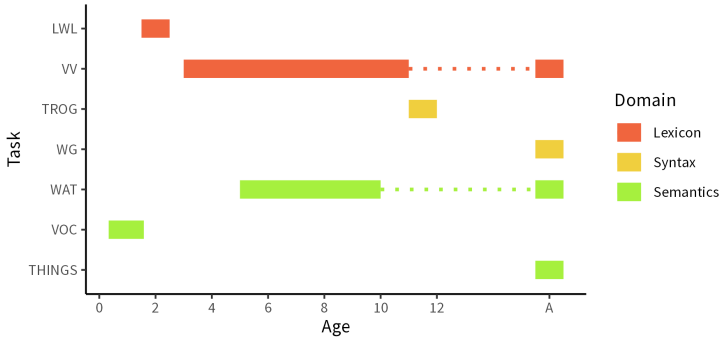

Figure 1: Tasks in DEVBENCH arranged by linguistic domain, along with the ages for which corresponding human data are available. A: Adult.

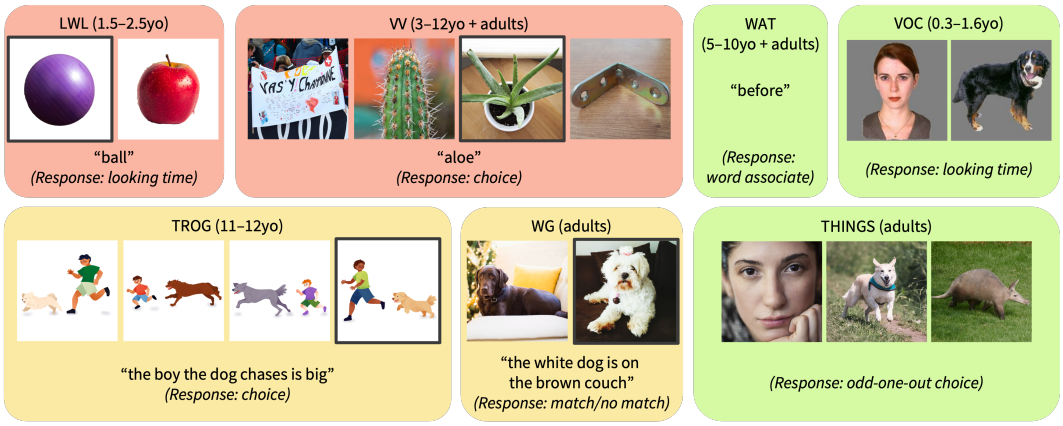

Figure 2: Sample trials for each task in DEVBENCH.

among its constituent members when given a sentential label. The semantic tasks measure conceptual (visual or linguistic) representation space via representational similarity. A schematic of the tasks and the ages of participants contributing data for each task is shown in Figure 1, and sample trials are shown in Figure 2. All code and data are available at `github.com/alvinwmtan/dev-bench`; all data are licensed under CC BY-NC-SA or more permissible licences, and have been anonymised to eliminate any personally identifiable information.

3.1 Tasks

Lexicon: Looking-while-listening (LWL) In this task, children are presented with two visual images (one target and one distractor) as well as a verbal cue (e.g., "Look at the dog!") [32]. Children's proportion looking time to each image is measured. We used data from three datasets to improve age coverage: 18-month-old data from Adams et al. [33], 24-month-old data from Frank et al. [34], and 30-month-old data from Donnelly and Kidd [35] (total $N = 294$). Some images in the original stimuli set were not shareable due to license restrictions; in these cases, we used replacement stimuli matched for visual and semantic properties.

Lexicon: Visual vocabulary task (VV) In this task, participants hear a word (e.g., "swordfish") and then see four visual images corresponding to the target and three distractor images that range in similarity to the target word (close, far, and distal) [36]. Participants respond by choosing the image that they think matches the verbal cue. We used an unpublished dataset from Long et al. [36] containing responses from 3- to 12-year-olds as well as adults (total $N = 1780$).

Syntax: Test of Receptive Grammar (TROG) In this task, participants are presented with four visual images (one target and three distractors) and then hear a phrase or sentence cue (e.g., "The

brown horse chases the white dog") [37]. In most trials, the distractors are constructed such that they would align with a label that contains the same words as the cue but in a different order (e.g., an image containing a brown dog chasing a white horse). Participants respond by choosing the image that they think matches the verbal cue. We used stimuli and unpublished data from Silverman and Yeatman [38], containing responses from children aged 11–12 (total N = 514).

Syntax: Winoground-NoTag (WG) This task contains trials with two images and two sentence labels, such that each image has one corresponding label, and the two labels only differ in word order [39]. Human ratings were collected by asking adult crowdworkers whether one particular label matched one particular image, and an image–caption score was calculated as the proportion of affirmative responses (total N = 171). For comparability to the other tasks in this benchmark, we considered the image–caption scores for the two images and one sentence of each trial, converted to proportions. We included only trials labelled "NoTag" from [40], which reflect vanilla Winoground trials that rely more narrowly on syntax (as opposed to also requiring pragmatics or other language abilities).

Semantics: Free word association task (WAT) In this task, participants were presented with a cue word, and asked to name the first word association that came to mind. We use association frequencies as a proxy for similarity scores. We combined a child dataset from Entwisle [41] with a subset of the adult dataset from Nelson et al. [42] comprising all cue words found in the child dataset (total $N >$ 7040), and thresholded the included responses to omit idiosyncratic responses.

Semantics: Visual object categorization (VOC) In this task, 4-, 10-, and 14-month-old infants saw pairs of images, and proportion looking times to each image was measured [43] (total N = 73). We used replacement stimuli for some images that were not shareable in the original stimuli set. Dissimilarities between images were calculated as the difference in proportion looking times to the two images.

Semantics: THINGS similarity ratings In this task, adult participants did a triplet odd-one-out task one triads constructed from the THINGS database [44, 45] (total N = 12340). These were used to generate a sparse positive similarity embedding [46], which we used to calculate pairwise similarities among images.

3.2 Human–model comparison

Because we were interested in response patterns, we conducted human–model comparison by examining the (dis)similarities in human and model distributions in responses.

Our goal on the lexicon and syntax tasks was to compare the distribution of human choices across response options to model choices. We obtained image–text matching logits for each response option, and calculated the *softmax-optimised Kullback–Leibler divergence* of human responses from model responses. This novel metric was operationalised as the minimum KL divergence between the human response probability distributions $\mathbf{h}$ from model logits $\mathbf{m}$, optimizing the softmax exponent β. We average this divergence across trials t, and calculate each distribution across images i:

$$D^*_{\mathrm{KL}}(h \parallel m) = \min_{\beta} \frac{1}{t} \sum_{t} D_{\mathrm{KL}} \left(\mathbf{h}_t \parallel \frac{e^{\beta \mathbf{m}_{it}}}{\sum_i e^{\beta \mathbf{m}_{it}}} \right)$$

For WAT, we calculated the softmax-optimised KL divergence of human association probabilities from softmaxed model text embedding similarities, averaged over all trials. For the visual semantic tasks, we instead conducted human–model comparison by applying representational similarity analysis (RSA) [47] on human and model representational similarity matrices, which represents correlations in the representational geometries of humans and models. All comparisons were conducted within each age bin when applicable.

Our method applies to models from which image–text matching scores could be directly extracted (i.e., similarity models). More recent models have alternatively integrated visual and language inputs via conditional text generation (i.e., generation models). There is as yet limited consensus for the best method to obtain image–text matching scores for generation models; we conducted two exploratory

 https://doi.org/10.52202/079017-2462

Table 2: Model characteristics and performance across all tasks, demonstrating variation across models. Arrows indicate the direction of better performance (i.e., lower is better vs. higher is better). Bolded results indicate most human-like result on a task.

			Lexicon		Syntax		Semantics		
Model	# params	# images	LWL (↓)	VV (↓)	TROG (↓)	WG (↓)	WAT (↓)	VOC (↑)	THINGS (↑)
CLIP-base [48]	149M	400M	0.014	0.205	0.732	0.256	0.495	-0.081	**0.397**
CLIP-large [48]	428M	400M	0.013	**0.179**	0.692	0.256	0.495	0.005	0.246
ViLT [49]	87M	4.1M	0.009	0.326	0.682	0.252	0.495	-0.053	0.127
FLAVA [50]	350M	70M	0.013	0.197	0.912	0.254	0.495	-0.042	0.189
BLIP [51]	252M	14M	0.010	0.193	**0.576**	0.259	0.495	-0.104	0.185
BridgeTower [52]	333M	4M	**0.008**	0.265	0.584	**0.250**	0.495	-0.095	0.345
OpenCLIP-H [53]	1.0B	32B	0.012	0.188	0.683	0.255	**0.495**	0.031	0.227
SigLIP [54]	800M	9B	0.067	0.612	0.888	0.258	0.495	-0.028	0.192
CVCL [4]	26M	600K	0.060	0.740	0.911	0.258	0.495	**0.138**	0.175
Human			0.010	0.091	0.028			0.251	
Random (OpenCLIP)	1.0B	0	0.087	0.740	0.908	0.258	0.495	0.246	0.054

evaluation methods relying on next-token prediction and on log-likelihood measurement. More details on the evaluation of generation models can be found in Appendix C.

3.3 Baselines

We also constructed two baselines for the benchmark to demonstrate the dynamic range that is possible for each task. First, we calculated a human baseline for tasks on which participant-level data were available (LWL, VV, TROG, VOC). To estimate this baseline, we randomly split the participants into two groups and calculated the between-group softmax-optimised KL divergence or RSA similarity as appropriate, repeating for 1000 random splits. We used the median result as a point estimate of the human baseline, which serves as a positive baseline for our tasks. Note that this baseline is an underestimate of the true values, because the results are not corrected upwards for the attenuation due to splitting the data in half.

We also added a random baseline for all tasks, generated using a random initialisation of the OpenCLIP model. The random baseline serves as a negative baseline for our tasks.

4 Benchmark

We evaluated a diverse set of vision–language models on our benchmark, using an NVIDIA T4 GPU, an NVIDIA A40 GPU, or CPUs (depending on resource availability). A summary of model performance for each task (averaged across all ages) is shown in Table 2, along with model characteristics (number of parameters and size of training set). For lexicon and syntax tasks as well as WAT, we report D^*_{KL}, for which *lower* scores indicate greater human-likeness in response patterns. For visual semantic tasks (VOC, THINGS), we report RSA similarity, for which *higher* scores indicate greater human-likeness in response patterns. Divergences are not comparable across datasets due to different features of each dataset. A description of all models and more details on the evaluation setup can be found in Appendix A.

Overall, CLIP-large and OpenCLIP-H performed relatively well on the lexicon and syntax tasks. CVCL, which was trained on a small set of annotated head-mounted camera from infants [4], was mostly dissimilar to humans in lexicon and syntax tasks, even on the looking-while-listening (LWL) task, which was administered to young children. However, CVCL outperformed other models on the visual object categorization task, suggesting that its visual representational space was more similar to infants'. SigLIP was also unusually poor-performing (especially given its general performance and accuracy), potentially suggesting that aspects of its training (e.g., its objective function) may have resulted in non-alignment with humans. We thus excluded it as an outlier from further analyses.

5 Analysis

To further understand the relationship between model and human responses, we conduct more fine-grained analyses, aiming to answer three specific research questions:

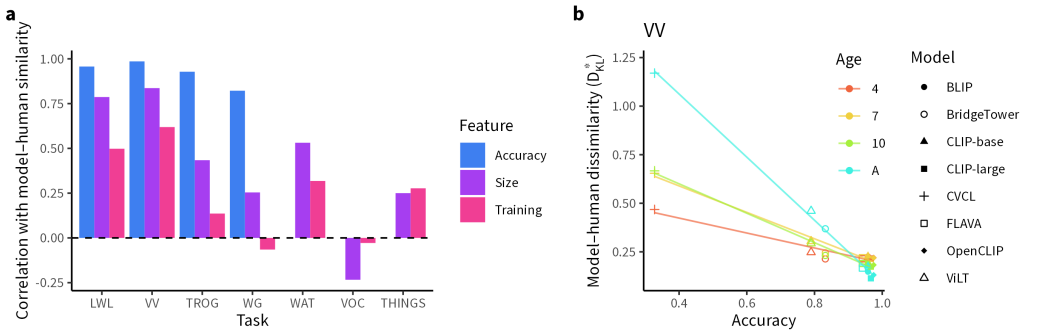

Figure 3: (a) Correlations between model–human similarity and task accuracy, log number of parameters (size), and log number of training images for each task (training), averaged across ages. Accuracy correlates the most strongly with model–human similarity, followed by size, then training. (b) Model–human dissimilarity as a function of task accuracy for each model on the Visual Vocabulary task. Higher performing models showed a closer correspondence to behavioral patterns from children and adults. A: Adult.

1. How does model–human similarity relate to other features of the models, namely their size (in terms of parameters), their training dataset size, and their overall accuracy on the tasks?
2. How does model–human similarity change over the course of model training, and in particular can we elucidate "developmental" trends in model training?
3. On which items are models and humans most dissimilar? On which items are they most similar?

5.1 Model feature analysis

To understand the variation in human-likeness exhibited by different models, we considered a range of model features that may affect response patterns, namely the number of parameters of the model, the number of examples in its training set, and its accuracy on the task (for lexicon and syntax tasks, which have a "correct" answer). For each task and each feature, we calculated Pearson's correlations between feature values and model–human similarities; number of parameters and number of training images were log-transformed. Similarity–feature correlations are shown in Figure 3a.

Overall, we found that model–human similarity was most correlated with task accuracy, with consistently high correlations across all ages and tasks. Model size also correlates relatively well with model–human similarity for most tasks except for VOC, which is reasonable given that VOC was conducted on the youngest participants (aged 4–14 months). In contrast, the number of training examples exhibited the poorest correlations with model–human similarity, suggesting that dataset size may not be as informative about human-likeness as accuracy or model size.

To illustrate the relationship between accuracy and model–human similarity, we plot these values for the Visual Vocabulary (VV) task in Figure 3b, which is also the task for which we have the largest coverage across age groups. In the VV task, we found a strong correlation between accuracy and model–human similarity for all age bins. In addition, we examined whether we would see differences in the strength of this correlation in data from children of different ages. We found that worse-performing models tended to show response patterns that were more similar to those from younger children, whereas better-performing models showed response patterns more similar to those from older children and adults – captured by an interaction effect between accuracy and age in a linear mixed-effect model (b = -0.057, SE = 0.003, $p < .001$). These results suggest that children's lexical representations are more similar to those instantiated in lower-performing multimodal models early in development, and gradually become more similar to higher-performing multimodal modals across middle childhood.

https://doi.org/10.52202/079017-2462

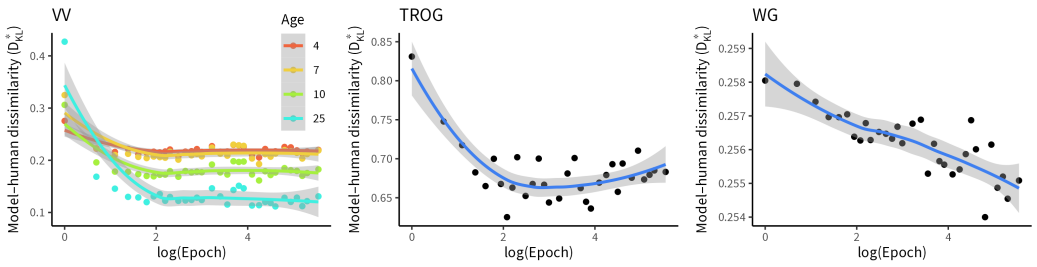

Figure 4: Trajectories of model–human similarity for VV, TROG, and WG. OpenCLIP-H becomes more human-like over training, and recovers developmental trends for VV.

5.2 Model training analysis

We next sought to understand whether human developmental trajectories were comparable to model training trajectories. We made use of OpenCLIP-H [53], an open model for which checkpoints were available on Hugging Face, allowing us to conduct "developmental" analyses. To do so, we sampled 32 checkpoints at approximately logarithmic intervals, and calculated model–human similarities for each task for each checkpoint. Training trajectory curves for three tasks (VV, TROG, and WG) are shown in Figure 4.

Overall, OpenCLIP increased in similarity to humans across training for these three tasks. Further, model performance on the Visual Vocabulary task also reflects a developmental trend: OpenCLIP exhibited greatest similarity to younger humans earlier in the training regime, and greatest similarity to older humans later in the training regime. This pattern matches the trend shown across models in Figure 3b – in both cases, better performing models show better correspondence to human data.

For brevity, we briefly describe the trajectory curves for OpenCLIP-H for the four remaining tasks; complete results are detailed in the SI. For LWL, we found that model–human similarity also improves with training, as expected. For the semantics tasks, the results were more complex, however. For WAT, model–human similarity largely remains constant over training, indicating that language representations do not significantly change. For VOC, the developmental trajectories were non-monotonic. For THINGS, model–human similarity *decreases* with training, perhaps indicating divergence with human visual representational space. Together, these results suggest that multimodal model representations may be less applicable to the semantics tasks we selected, and future work is needed to understand differences between semantic representations in uni- and multimodal representations.

5.3 Item-level analysis

Finally, we examined specific items to determine which items exhibited the greatest and least dissimilarity in response pattern between models and humans. For each model, we z-scored the D^*_{KL} values to control for inter-model variation in overall model–human similarity. We then averaged D^*_{KL} values across models for each trial, and extracted the top five items with the greatest mean divergence and the top five items with the least mean divergence for qualitative analysis. We illustrate this method for VV and WG, as shown in Table 3.

The most dissimilar items reveal certain features which may particularly drive model–human dissimilarity, particularly in comparison to the most similar items. For VV, some such features include polysemous targets (e.g., "horn" and "net"), targets which may have other labels (e.g., "hand plow" for "hoe", "pudding" for "flan"), or targets which could be labelled as one of the distractor categories (e.g., "lollipop" is a type of "candy"). For WG, several of the most dissimilar items have genuinely contentious captions – notably, the image for the caption "the dog is swimming and the person is standing" has the person hunched over rather than fully upright, the image for the caption "a person sits and a dog stands" has the dog mid-leap, and the image for the caption "green pants and blue top" has a top that could be described as grey. More broadly, across both VV and WG, humans appear to be better able to handle ambiguity and choose the most likely answer (perhaps through pragmatic reasoning [55]), whereas models are more likely to put less density on the true target.

Table 3: Top five most dissimilar items and top five most similar items between humans and all models for Visual Vocabulary and Winoground tasks.

Most dissimilar items	
VV	horn (distractors: bone, chin, ladybug) hoe (distractors: peg, dustpan, beaker) flan (distractors: fuse, amplifier, turnstile) net (distractors: tee, domino, hydrant) lollipop (distractors: candy, doorbell, crumb)
WG	a person whispering into a dog's ear / a dog whispering into a person's ear there are more ladybugs than flowers / there are more flowers than ladybugs the dog is swimming and the person is standing / the dog is standing and the person is swimming blue pants and green top / green pants and blue top a person sits and a dog stands / a person stands and a dog sits
Most similar items	
VV	foam (distractors: float, quilt, asparagus) saddle (distractors: handle, figurine, broccoli) stump (distractors: log, bookshelf, showerhead) sorbet (distractors: palette, tamale, chive) typewriter (distractors: printer, sunglasses, drumstick)
WG	a horse getting wet / getting a horse wet a large living thing in front of a large non-living thing / a large non-living thing in front of a large living thing a deer's nose is resting on a child's hand / a child's hand is resting on a deer's nose clothing on lines / lines on clothing soft shoes are on a smooth floor / smooth shoes are on a soft floor

6 Discussion

In this work, we introduced DEVBENCH, a multimodal benchmark for language learning consisting of seven tasks with corresponding data from both children and adults. Evaluating a set of vision–language models revealed variation across models in terms of model–human similarity across lexical, syntactic, and semantic domains. Furthermore, model–human similarity was strongly correlated with model accuracy, as well as model size to a lesser extent. Analysing OpenCLIP-H checkpoints across training also recovered developmental trends for some tasks, but not others.

More broadly, DEVBENCH provides a method for models trained on developmentally realistic data to be evaluated using a method that is comparable to how children are evaluated. Recent work has seen a plethora of new unimodal or multimodal learning models, including some that are evaluated here (e.g., [4]), but to date no models trained on developmentally realistic data (e.g., head-mounted camera data) have been evaluated actual developmental performance. We believe that benchmarks like DEVBENCH are essential for understanding the degree to which any given model can be used to approximate human learning.

Systematically comparing models and children's response patterns – rather than overall accuracy – is an essential piece of this puzzle. In doing so, our results already highlight ways in which model training trajectories are rather *unlike* human development trajectories, pointing towards new avenues for future work. For example, models appear to be worse than humans at handling ambiguous inputs; ambiguity resolution is thus a potential avenue for future multimodal model development.

6.1 Limitations and future work

The development of DEVBENCH was constrained by currently available human data; notably, this means that some tasks have relatively few items (in comparison with many other machine learning benchmarks), and some tasks have relatively few human participants. These limitations may result in uncertain reliabilities for the obtained results. More data are currently being collected for some of the tasks in DEVBENCH, which we anticipate will improve reliability, and simultaneously permit more expansive developmental analyses due to an extension of the included age ranges. However, we hope that this approach and standardized format will enable future researchers to contribute novel datasets to DEVBENCH, which we anticipate may grow in the coming years. Indeed, DEVBENCH

 https://doi.org/10.52202/079017-2462

also includes primarily tasks for English-speaking children and adults, due to dataset availability. This situation reflects inequalities in language acquisition research [56], and more data is needed to construct a multilingual version of DEVBENCH, which will be more comprehensive and generalisable.

Additionally, model performance in our evaluation setup may be affected by the domain gap between models' training data and the stimuli used in our benchmark; for example, TROG uses cartoon depictions of events, which are dissimilar to the more photorealistic training data of CLIP. Thus, our evaluation results likely represent a lower bound on model–human similarity. Children as young as two years of age are able to learn from and generalise to pictographic depictions of objects [57–59], however, suggesting that generalisation across representations is an early-acquired skill.

In our analyses of the relationship between model–human similarity and model features, we could not hold model architectures constant due to the limited availability of relevant model checkpoints – thus we cannot make strong claims about precisely what aspects of models lead to better fit to human responses. Systematic evaluation of the roles of training data and model size thus remains an important research question for future work (see [60, 61] for related work on scaling).

Finally, we chose one linking hypothesis between model logits and human distributions, namely optimised KL divergence. This hypothesis assumes a perfect calibration between model logits and true uncertainties, which is only valid to some extent. Further research in model calibration [62] may help to mitigate these effects.

6.2 Conclusion

We hope that DEVBENCH can serve as an encouragement for machine learning researchers and cognitive scientists to develop vision–language models that not only perform well, but also can more closely approximate human learning. In particular, DevBench highlights the need for more fully open models with training checkpoints [63], enabling the study of training trajectories, as well as the need for more human-realistic training [2, 60] to better characterise model–human correspondences across developmental change. It may also be possible to adopt a developmental, multimodal approach to studying domains other than language, such as mathematical, logical, and social reasoning; more intentional data collection across a wide range of ages, tasks, and contexts will help to provide an increasingly comprehensive set of comparison data to better understand model learning trajectories (e.g., [64]). These research directions, among others, will help us to better understand the processes underlying human development, and how we might transfer humans' learning efficiencies onto machine learning models.

Acknowledgments and Disclosure of Funding

Thanks to the Jacobs Foundation for support of stimulus and dataset creation. This work was funded in part by an NIH K99HD108386 award to BL and a Stanford Interdisciplinary Graduate Fellowship to WAM.

The authors made the following contributions. AWMT: Conceptualisation, Data curation, Methodology, Formal analysis, Software, Writing – original draft, Writing – review & editing. SY: Data curation, Software, Writing – original draft, Writing – review & editing. BL, WAM, TM, RDS, JDY: Data curation, Writing – review & editing. MCF: Conceptualisation, Methodology, Writing – original draft, Writing – review & editing, Supervision.

References

[1] Michael C. Frank. Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*, 27(11):990–992, 2023. ISSN 1364-6613. doi: 10.1016/j.tics.2023.08.007.

[2] Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora. In Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell, editors, *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.conll-babylm.1.

[3] Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. BabyBERTa: Learning More Grammar With Small-Scale Child-Directed Language. In Arianna Bisazza and Omri Abend, editors, *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.conll-1.49.

[4] Wai Keen Vong, Wentao Wang, A. Emin Orhan, and Brenden M. Lake. Grounded language acquisition through the eyes and ears of a single child. *Science*, 383(6682):504–511, 2024. doi: 10.1126/science.adi1374.

[5] Wentao Wang, Wai Keen Vong, Najoung Kim, and Brenden M. Lake. Finding Structure in One Child's Linguistic Experience. *Cognitive Science*, 47(6):e13305, 2023. ISSN 1551-6709. doi: 10.1111/cogs.13305.

[6] Arielle Borovsky. Lexico-semantic structure in vocabulary and its links to lexical processing in toddlerhood and language outcomes at age three. *Developmental Psychology*, 58(4):607–630, 2022. ISSN 1939-0599. doi: 10.1037/dev0001291.

[7] Ellen Irén Brinchmann, Johan Braeken, and Solveig-Alma Halaas Lyster. Is there a direct relation between the development of vocabulary and grammar? *Developmental Science*, 22(1):e12709, 2019. ISSN 1467-7687. doi: 10.1111/desc.12709.

[8] Laura Justice, Kate Cain, Hui Jiang, Jessica Logan, Rongfang Jia, Hui Jiang, Jessica A. Logan, and Rongfang Jia. Modeling the Nature of Grammar and Vocabulary Trajectories From Prekindergarten to Third Grade. *Journal of Speech, Language, and Hearing Research*, 61(4):910–923, 2018. doi: 10.1044/2018_JSLHR-L-17-0090.

[9] Margaux Keith and Elena Nicoladis. The role of within-language vocabulary size in children's semantic development: Evidence from bilingual children. *Journal of Child Language*, 40(4):873–884, 2013. ISSN 0305-0009, 1469-7602. doi: 10.1017/S0305000912000268.

[10] Jessie Ricketts, Nicola Dawson, and Robert Davies. The hidden depths of new word knowledge: Using graded measures of orthographic and semantic learning to measure vocabulary acquisition. *Learning and Instruction*, 74:101468, 2021. ISSN 0959-4752. doi: 10.1016/j.learninstruc.2021.101468.

[11] Michael C Frank. Baby steps in evaluating the capacities of large language models. *Nature Reviews Psychology*, 2(8):451–452, 2023.

[12] Anna A. Ivanova. Running cognitive evaluations on large language models: The do's and the don'ts, 2023.

[13] Jennifer Hu and Michael C Frank. Auxiliary task demands mask the capabilities of smaller language models. *arXiv preprint arXiv:2404.02418*, 2024.

[14] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible Scaling Laws for Contrastive Language-Image Learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2829. IEEE, 2023. ISBN 9798350301298. doi: 10.1109/CVPR52729.2023.00276.

[15] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019. doi: 10.1109/TPAMI.2018.2798607.

[16] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, 2016. URL `https://arxiv.org/abs/1612.06890`.

[17] Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. IGLUE: A benchmark for transfer learning across modalities, tasks, and languages. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2370–2392. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/bugliarello22a.html`.

[18] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, Jing Shao, and Wanli Ouyang. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark, 2023.

[19] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024.

 https://doi.org/10.52202/079017-2462

[20] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024.

[21] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, joseph heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and Joao Carreira. Perception test: A diagnostic benchmark for multimodal video models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 42748–42761. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/8540fba4abdc7f9f7a7b1cc6cd60e409-Paper-Datasets_and_Benchmarks.pdf.

[22] Lin Su, Nan Duan, Edward Cui, Lei Ji, Chenfei Wu, Huaishao Luo, Yongfei Liu, Ming Zhong, Taroon Bharti, and Arun Sacheti. GEM: A general evaluation benchmark for multimodal tasks. *CoRR*, abs/2106.09889, 2021. URL https://arxiv.org/abs/2106.09889.

[23] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A. Lee, Yuke Zhu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Multibench: Multiscale benchmarks for multimodal representation learning. *CoRR*, abs/2107.07502, 2021. URL https://arxiv.org/abs/2107.07502.

[24] Eliza Kosoy, Emily Rose Reagan, Leslie Lai, Alison Gopnik, and Danielle Krettek Cobb. Comparing Machines and Children: Using Developmental Psychology Experiments to Assess the Strengths and Weaknesses of LaMDA Responses. In *Proceedings of the First Workshop on AI Meets Moral Philosophy and Moral Psychology at NeurIPS 2023*, 2023. doi: 10.48550/arXiv.2305.11243.

[25] Luca Weihs, Amanda Rose Yuile, Renée Baillargeon, Cynthia Fisher, Gary Marcus, Roozbeh Mottaghi, and Aniruddha Kembhavi. Benchmarking progress to infant-level physical reasoning in ai. *TMLR*, 2022.

[26] Guangyuan Jiang, Manjie Xu, Shiji Xin, Wei Liang, Yujia Peng, Chi Zhang, and Yixin Zhu. MEWL: Few-shot multimodal word learning with referential uncertainty. In *Proceedings of the 40th International Conference on Machine Learning*, pages 15144–15169. PMLR, 2023. URL https://proceedings.mlr.press/v202/jiang23i.html.

[27] Saber Sheybani, Linda B. Smith, Zoran Tiganj, Sahaj Singh Maini, and Aravind Dendukuri. ModelVsBaby: A Developmentally Motivated Benchmark of Out-of-Distribution Object Recognition, 2024.

[28] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension, 2023.

[29] Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 5484–5505. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/117c5c8622b0d539f74f6d1fb082a2e9-Paper-Datasets_and_Benchmarks.pdf.

[30] Tyler A. Chang and Benjamin K. Bergen. Word Acquisition in Neural Language Models. *Transactions of the Association for Computational Linguistics*, 10:1–16, 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00444.

[31] Linnea Evanson, Yair Lakretz, and Jean Rémi King. Language acquisition: Do children and language models follow similar learning stages? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12205–12218. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.773.

[32] Anne E. Fernald, Renate Zangl, Ana Luz Portillo, and Virginia A. Marchman. *Using Eye Movements to Monitor Spoken Language Comprehension by Infants and Young Children: Looking While Listening*. Language Acquisition and Language Disorders. John Benjamins Publishing Company, 2008. ISBN 978-90-272-5304-0 978-90-272-5305-7 978-90-272-9150-9. doi: 10.1075/lald.44.06fer.

[33] Katherine A. Adams, Virginia A. Marchman, Elizabeth C. Loi, Melanie D. Ashland, Anne Fernald, and Heidi M. Feldman. Caregiver Talk and Medical Risk as Predictors of Language Outcomes in Full Term and Preterm Toddlers. *Child Development*, 89(5):1674–1690, 2018. ISSN 1467-8624. doi: 10.1111/cdev.12818.

[34] Michael C. Frank, Elise Sugarman, Alexandra C. Horowitz, Molly L. Lewis, and Daniel Yurovsky. Using Tablets to Collect Data From Young Children. *Journal of Cognition and Development*, 17(1):1–17, 2016. ISSN 1524-8372. doi: 10.1080/15248372.2015.1061528.

[35] Seamus Donnelly and Evan Kidd. Onset Neighborhood Density Slows Lexical Access in High Vocabulary 30-Month Olds. *Cognitive Science*, 45(9):e13022, 2021. ISSN 1551-6709. doi: 10.1111/cogs.13022.

[36] Bria Long, Wanjing Anya Ma, Rebecca D. Silverman, Jason D. Yeatman, and Michael C. Frank. Developmental changes in the precision of visual concept knowledge. In *Annual Meeting of the Vision Sciences Society*, 2024.

[37] Dorothy V. M. Bishop. *Test for the Reception of Grammar (TROG)*. Medical Research Council, 2nd ed edition, 1989.

[38] Rebecca D. Silverman and Jason D. Yeatman. Accelerating literacy with digital technology, 2023.

[39] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5228–5238, 2022. doi: 10.1109/CVPR52688.2022.00517.

[40] Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. Why is Winoground Hard? Investigating Failures in Visuolinguistic Compositionality. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2250. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.143.

[41] Doris R. Entwisle. *Word Associations of Young Children / by Doris R. Entwisle.* Johns Hopkins Press, 1966.

[42] Douglas L. Nelson, Cathy L. McEvoy, and Thomas A. Schreiber. The University of South Florida word association, rhyme, and word fragment norms., 1998. URL `http://w3.usf.edu/FreeAssociation/`.

[43] Céline Spriet, Etienne Abassi, Jean-Rémy Hochmann, and Liuba Papeo. Visual object categorization in infancy. *Proceedings of the National Academy of Sciences*, 119(8):e2105866119, 2022. doi: 10.1073/pnas.2105866119.

[44] Martin N. Hebart, Adam H. Dickter, Alexis Kidder, Wan Y. Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I. Baker. THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, 14(10):e0223792, 2019. ISSN 1932-6203. doi: 10.1371/journal.pone.0223792.

[45] Martin N Hebart, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife*, 12:e82580, 2023. ISSN 2050-084X. doi: 10.7554/eLife.82580.

[46] Charles Y. Zheng, Francisco Pereira, Chris I. Baker, and Martin N. Hebart. Revealing interpretable object representations from human behavior. In *Proceedings of the 7th International Conference on Learning Representations*, 2018. doi: 10.48550/arXiv.1901.02915.

[47] Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2, 2008. ISSN 1662-5137. doi: 10.3389/neuro.06.004.2008.

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, 2021.

[49] Wonjae Kim, Bokyung Son, and Ildoo Kim. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision, 2021.

[50] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A Foundational Language And Vision Alignment Model, 2022.

[51] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, 2022.

[52] Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, Wanxiang Che, and Nan Duan. BridgeTower: Building Bridges Between Encoders in Vision-Language Representation Learning, 2024.

[53] Romain Beaumont. Large scale openCLIP: L/14, H/14 and g/14 trained on LAION-2B, 2022. URL `https://laion.ai/blog/large-openclip`.

 https://doi.org/10.52202/079017-2462

[54] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. URL https://arxiv.org/abs/2303.15343.

[55] Noah D. Goodman and Michael C. Frank. Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, 20(11):818–829, 2016. ISSN 1364-6613. doi: 10.1016/j.tics.2016.08.005.

[56] Evan Kidd and Rowena Garcia. How diverse is child language acquisition research? *First Language*, 42 (6):703–735, 2022. ISSN 0142-7237. doi: 10.1177/01427237211066405.

[57] Patricia A. Ganea, Megan Bloom Pickard, and Judy S. DeLoache. Transfer between Picture Books and the Real World by Very Young Children. *Journal of Cognition and Development*, 9(1):46–66, 2008. ISSN 1524-8372. doi: 10.1080/15248370701836592.

[58] Gabrielle Simcock and Judy DeLoache. Get the picture? The effects of iconicity on toddlers' reenactment from picture books. *Developmental Psychology*, 42(6):1352–1357, 2006. ISSN 1939-0599, 0012-1649. doi: 10.1037/0012-1649.42.6.1352.

[59] Medha Tare, Cynthia Chiong, Patricia Ganea, and Judy DeLoache. Less is more: How manipulative features affect children's learning from picture books. *Journal of Applied Developmental Psychology*, 31 (5):395–400, 2010. ISSN 0193-3973. doi: 10.1016/j.appdev.2010.06.005.

[60] Bria Long, Violet Xiang, Stefan Stojanov, Robert Z. Sparks, Zi Yin, Grace E. Keene, Alvin W. M. Tan, Steven Y. Feng, Chengxu Zhuang, Virginia A. Marchman, Daniel L. K. Yamins, and Michael C. Frank. The BabyView dataset: High-resolution egocentric videos of infants' and young children's everyday experiences, 2024.

[61] Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. Observational Scaling Laws and the Predictability of Language Model Performance, 2024.

[62] Cheng Wang. Calibration in Deep Learning: A Survey of the State-of-the-Art, 2024.

[63] Michael C. Frank. Openly accessible LLMs can help us to understand human cognition. *Nature Human Behaviour*, 7(11):1825–1827, 2023. ISSN 2397-3374. doi: 10.1038/s41562-023-01732-4.

[64] Michael C Frank. Learning variability network exchange (levante): A global framework for measuring children's learning variability through collaborative data sharing, Sep 2024. URL osf.io/preprints/psyarxiv/namx2.

[65] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

[66] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 25278–25294. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/a1859debfb3b59d094f3504d5ebb6c25-Paper-Datasets_and_Benchmarks.pdf.

[67] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

[68] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[69] Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. Tinyllava: A framework of small-scale large multimodal models, 2024. URL https://arxiv.org/abs/2402.14289.

[70] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world, 2023. URL https://arxiv.org/abs/2306.14824.

[71] moondream AI team. moondream2, 2024. URL https://github.com/vikhyat/moondream.

[72] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models, 2024. URL https://arxiv.org/abs/2311.03079.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] Abstract and introduction reflect a summary of the benchmark and results.
   (b) Did you describe the limitations of your work? [Yes] See Section 6.1.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See the second paragraph of Section 6.1.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] All human data used here are secondary, and have been anonymised.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A] No theoretical results are given.
   (b) Did you include complete proofs of all theoretical results? [N/A] As above.
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] A link to the repository is provided in Section 3.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A] No training was conducted.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] As above.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 4.
   (b) Did you mention the license of the assets? [Yes] See Section 4.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See Section 3.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] All human data were secondary.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Section 4.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] No primary human data collection was conducted.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] As above.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] As above.

https://doi.org/10.52202/079017-2462

Table 4: Participant demographics for all tasks. A: Adult.

Task	Age bin	Mean age	N	Country of origin	Test language
LWL	1.5	1.42	166	USA	English
	2	2.49	28	USA	English
	2.5	2.56	100	Australia	English
VV	4	5.5	178	USA	English
	7	8.22	458	USA	English
	10	10.17	939	USA	English
	A		205	USA	English
TROG	11	11.77	514	USA	English
WG	A		171	Not specified	English
WAT	5	5.6	200	USA	English
	6	6.55	280	USA	English
	8	8.73	280	USA	English
	10	10.45	280	USA	English
	A		6000	USA	English
VOC	0.3	0.33	24	France	French
	0.8	0.9	24	France	French
	1.6	1.6	25	France	French
THINGS	A		12340	Not specified	English

A Additional method details

A.1 Additional task details

A breakdown of the demographics for each age bin for each task is shown in Table 4. An overview of key methodological dimensions for the sources of each task is shown in Table 5.

A.2 Evaluated models

CLIP-base & CLIP-large Contrastive Language–Image Pretraining [48] was one of the earliest families of vision–language models. It uses ViT for its vision encoder, and a GPT-style transformer for its text encoder, with further contrastive training on image–text pairs to increase the cosine similarity of matched image–caption pairs. We used the ViT-B/32 model for CLIP-base, and the ViT-L/14 model for CLIP-large.

ViLT Vision-and-Language Transformer [49] introduced a method for vision–language pretraining without convolutions. It uses word embeddings concatenated with a linear projection of image patches as inputs, which are passed through a unified transformer encoder trained on three objectives: image–text matching, masked language modelling, and word–patch alignment. We used the ViLT-B/32 model in our evaluation.

FLAVA Foundational Language and Vision Alignment Model [50] combines multimodal and unimodal pretraining objectives to broaden the types of usable data sources. The FLAVA model uses a ViT vision encoder, a ViT text encoder, and a ViT multimodal encoder that fused the vision and text hidden states. The vision encoder was trained on masked image modelling while the text encoder was trained on masked language modelling. The multimodal encoder was trained on masked multimodal modelling and image–text matching.

BLIP Bootstrapping Language–Image Pretraining [51] uses a multimodal mixture of encoder–decoder transformers. It incorporates unimodal language and text encoders trained on image–text contrastive learning, an image-grounded text encoder trained on image–text matching, and an image-grounded text decoder trained on language modelling. They also used a bootstrapping approach to produce synthetic captions for web images, and noisy image–caption pairs are filtered out. We used the BLIP-B/16 model with image–text matching training on the COCO dataset [65] for evaluation.

BridgeTower BridgeTower [52] adds more extensive cross-modal interaction to the "Two Tower" approach of image and text modelling. It includes a ViT vision encoder, a RoBERTa text encoder, and a cross-modal encoder connected to the unimodal encoders via bridge layers. The model is pretrained on masked language modelling and image–text matching.

OpenCLIP-H OpenCLIP [53] is an open-data, open-source implementation of the CLIP model. We used the OpenCLIP-H/14 model trained on LAION-2B [66] for evaluation, which additionally has most intermediate checkpoints uploaded to Hugging Face; these were used for our training trajectory analyses.

SigLIP SigLIP [54] is a modification of CLIP which does not rely on batch-wise normalisation, using a sigmoid loss instead of a softmax loss. It uses a ViT vision encoder and a transformer text encoder, and treats each image–text pair as a binary classification problem (match vs non-match).

CVCL Child's View for Contrastive Learning [4] is a family of models trained on naturalistic egocentric videos drawn from head-mounted camera footage from a single child aged 6–25 months. We used the main model of CVCL, which uses a ResNeXt-50 model as its vision encoder, and mean-pooled text embedding as its text encoder; these were trained with a contrastive learning objective on co-occurring utterance–frame pairs from the egocentric video dataset.

A.3 Evaluation setup

For lexicon and syntax tasks, models were evaluated by passing in each image–text pair for each trial as inputs, and obtaining model logits for each pair. For LWL and WG, there were two images in each trial, while for VV and TROG, there were four images in each trial. Model logits were then used to calculate the softmax-optimised KL divergence with human responses.

For VOC and THINGS, we obtained image embeddings for each stimulus, and obtained a representational similarity matrix (RSM) by calculating the pairwise cosine similarity for each pair of images. We then compared the model RSM with that obtained from human responses by calculating the Spearman's rank correlation coefficient for entries below the main diagonal in model and human RSMs.

For WAT, we obtained text embeddings for each stimulus, and calculated the pairwise cosine similarity for all cue–target pairs in the human response data. Model similarity values were then used to calculate the softmax-optimised KL divergence with human response distributions for each cue word.

Some models (e.g., BridgeTower) always required both image and text inputs. For these models, we used an empty string as the dummy text input when obtaining image embeddings for VOC and THINGS, and we used a neutral gray square as the dummy image input when obtaining text embeddings for WAT.

A.4 Softmax-optimised Kullback–Leibler divergence

We used the softmax-optimised KL divergence as our novel metric of model–human dissimilarity. (Ordinary) KL divergence reflects how different a target probability distribution is from a reference probability distribution, often considered the true distribution. In the case of DEVBENCH, the reference distribution is obtained from human responses, while the target distribution is obtained from model responses.

The typical method of deriving probabilities from model responses is by conducting a softmax over logits. However, we considered that model logits may not be calibrated to the same scale as human responses, and therefore included the temperature, β, as a free parameter. In other words, the resultant distribution after optimisation can be considered a one-parameter projection from logit space to probability space, and the best fitting projection is that which induces the minimum KL divergence to the human response distribution.

 https://doi.org/10.52202/079017-2462

B Additional detailed results

B.1 Detailed benchmark results

Benchmark results for all tasks, split by human age bins, are shown in Table 6. As with the summarised results, CLIP-large and OpenCLIP-H perform relatively well across tasks, along with BridgeTower. CVCL performs poorly on most tasks, but exhibits the best correlations to 10-month-olds and 19-month-olds on the visual object categorisation task.

Additionally, although it is not the key index for our benchmark, we also report the accuracies on all datasets in Table 7 for reference.

B.2 Detailed feature analyses

Correlations between model features and model–human similarities for all tasks, split by human age bins, are shown in Table 8. Note that semantics tasks have no "correct" answer, and thus no accuracies. Accuracy is consistently the most correlated feature even when results are broken down by age bins. The semantic tasks also show greater variability when considering age bins, such that some age bins have better correlations with size, and other age bins have better correlations with training. However, it is important to note that there was very little variability in model–human similarity for the word association task, which may have exaggerated the variability in feature–similarity correlations.

B.3 Detailed trajectory analyses

OpenCLIP-H training trajectories for LWL, WAT, VOC, and THINGS are shown in Figure 5. LWL shows the expected developmental trend, with human-likeness increasing over training, although it is important to note that the three age groups use different sets of stimuli and are thus not inter-comparable. WAT shows almost no change over training, suggesting that language representations are not significantly changing over contrastive training. VOC shows a more complex pattern, whereby OpenCLIP-H similarity to infants aged 10 and 19 months exhibits a U-shaped pattern with a minimum around epoch 16, whereas similarity to infants aged 4 months exhibits an inverse U-shaped pattern also peaking around epoch 16. Finally, THINGS shows an inverse pattern, whereby model–human similarity actually decreases over training, rather than increasing. The trajectories for VOC and THINGS suggest that multimodal contrastive learning may actually result in representations that are less human-like, suggesting that human visual representations may not be shaped by vision–language correspondences to the same extent as models like OpenCLIP-H.

C Evaluating generation models

C.1 Evaluation methods

As an exploratory analysis, we also evaluated generation models using two methods. The first method relied on next-token prediction. We passed in an input consisting of the image and a text prompt: "Caption: '<text>'. Does the caption match the image? Answer Yes or No." This prompt closely matched the actual human task for the Winoground dataset. We obtained next-token prediction logits for 'Yes' and 'No', and then subtracted the 'No' logits from the 'Yes' logits, which approximates the log odds ratio between 'Yes' and 'No' for each image; these logit differences were then treated as image–text matching scores.

The second method relied on log-likelihood measurement. We passed in an input consisting of the image and a text prompt: "Describe this image. <sep> <text>" We then measured the log-likelihood of the sequence; these logits were then treated as image–text matching scores. Note that even though we measured the log-likelihood of the whole sequence, the substring that did not directly correspond to the target text was constant across comparisons for each trial, and thus should not affect the relative likelihood values.

We further attempted a third evaluation method relying on explicit ratings. We passed in an input consisting of the image and a text prompt: "Caption: '<text>'. How much does the caption match the image? Give only a rating from 0 to 100." However, this method largely produced '0' or '100' responses, and thus we do not include the corresponding results.

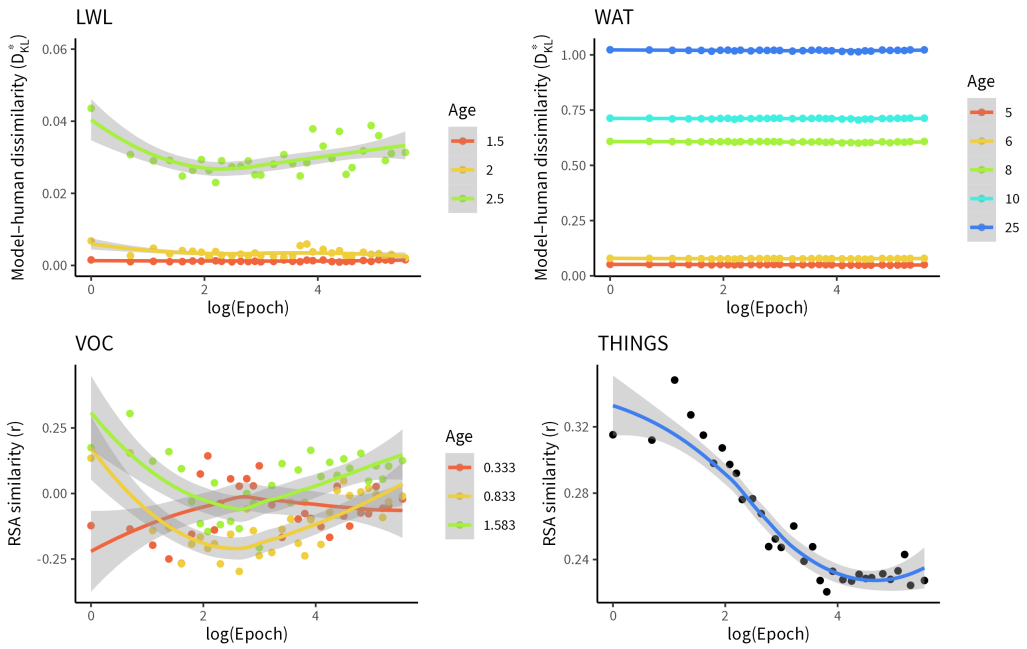

Figure 5: Trajectories of model–human similarity for LWL, WAT, VOC, and THINGS. Note that the three age groups for LWL are not comparable because they use different stimuli sets. For LWL and WAT, smaller values indicate greater model–human similarity, whereas for VOC and THINGS, larger values indicate greater model–human similarity.

Additionally, we only evaluated the lexical and syntactic tasks for these models, since it was not always clear that pure image or text features (needed for the semantic tasks) were extractable from the models.

C.2 Evaluated models

LLaVA-NeXT LLaVA-NeXT [67] is the 1.6 version of the Large Language and Vision Assistant (LLaVA) [68]. The model was trained on visual instruction data generated by GPT, using a visual encoder with Vicuna as the LLM component. LLaVA-NeXT improves upon LLaVA by increasing the image resolution and improving the mixture of the training data. We used the LLaVA-NeXT model with a Mistral 7B LLM.

TinyLLaVA TinyLLaVA [69] introduced a framework for generating small vision–language models, using the core idea from LLaVA of having an image encoder along with an LLM. We used their best performing model, which is the 3.1B model using Phi-2 as the LLM and SigLIP as the training objective.

Kosmos-2 Kosmos-2 [70] is a multimodal large language model that incorporates object semantics (using bounding boxes) as well as grounding for text. It includes an image encoder as well as a Magneto transformer as its LLM component. The model was trained on image–text pairs, and was also instruction-tuned on image–text instruct data along with grounded image–text data.

moondream2 moondream2 [71] is a tiny vision–language model designed to be performant even with a small number of parameters, such that it can be run even on mobile devices. It includes an image encoder as well as Phi-2 as the LLM, and is trained on visual question answering. There is limited documentation about this model and its training method.

CogVLM CogVLM [72] is a vision–language foundation model that allows for deep fusion of image and text representations by including a visual expert in the attention and feedforward layers of

https://doi.org/10.52202/079017-2462

the language model. It uses a ViT for its vision encoder, and a pretrained GPT as its LLM component. We used the CogVLM-17B model in our evaluation.

C.3 Results and discussion

Results from the next-token prediction and log-likelihood measurement methods are shown in Table 9. In both methods, TinyLLaVA outperforms other generation models in most tasks, often by a sizeable margin. We also report accuracies from both methods in Table 10. We note that both in terms of model–human similarity and accuracy, the results may be underestimations as the specific prompts may be out of distribution for these models.

As we only evaluated five generation models, it was not possible to do further analyses relating to model features; further work evaluating a larger range of models would enable such analyses. It is also important to note that our evaluation methods represent only two out of a range of possible approaches; a more systematic search might reveal other aspects of variation across models.

Table 5: Overview of selected methodological dimensions for all tasks. A: Adult.

Task	Source	Age bins	Recruitment method	Administration method	Data collection method	Experimental setup	Primary goal of study
LWL	Adams et al.	1.5	Hospital and others	In-person	Eyetracking	Preferential looking	Investigate associations between caregiver talk and language skills in full/preterm children
	Frank et al.	2	Museum	In-person	Eyetracking	Preferential looking	Investigate the reliability of using tablets to collect experimental data from young children
	Donnelly & Kidd	2.5	Not specified	In-person	Eyetracking	Preferential looking	Investigate the relationship between phonological onset density and lexical access
VV	Long et al.	4, 7, 11, A	Schools and online	In-person and online	4AFC	Best match	Investigate developmental changes in visual concept knowledge
TROG	Silverman & Yeatman	11	Schools	In-person	4AFC	Best match	Investigate developmental changes in grammatical knowledge
WG	Thrush et al.	A	Amazon MTurk	Online	2AFC	Match or non-match	Investigate visuo–linguistic compositionality in vision–language models
WAT	Entwisle	5, 6, 8, 10	Schools	In-person	Oral response	Word association	Investigate chidlren's verbal conceptual space and its relationship with demographic factors
	Nelson et al.	A	Not specified	In-person	Written response	Word association	Investigate adults' verbal conceptual space and cue-to-target association strengths
VOC	Spriet et al.	0.3, 0.8, 1.6	Not specified	In-person	Eyetracking	Preferential looking	Investigate infants' visual conceptual space and its relationship with adult brain activity
THINGS	Hebart et al.	A	Amazon MTurk	Online	3AFC	Odd one out	Investigate adults' visual conceptual space and its relationship with adult brain activity

https://doi.org/10.52202/079017-2462

Table 6: Model performance across all tasks, split by human age bins (in years). A: Adult.

Model	Lexicon							Syntax	
	LWL (↓)			VV (↓)				TROG (↓)	WG (↓)
	1.5	2	2.5	4	7	10	A	11	A
CLIP-base	0.002	0.007	0.032	0.220	0.228	0.195	0.177	0.732	0.256
CLIP-large	0.002	0.007	0.031	0.216	**0.214**	**0.174**	**0.113**	0.692	0.256
ViLT	0.003	0.005	0.018	0.248	0.304	0.293	0.460	0.682	0.252
FLAVA	0.001	0.006	0.031	0.214	0.220	0.190	0.166	0.912	0.254
BLIP	**0.001**	0.008	0.021	0.226	0.216	0.182	0.147	**0.576**	0.259
BridgeTower	0.001	0.005	**0.017**	**0.213**	0.245	0.231	0.369	0.584	**0.250**
OpenCLIP-H	0.002	**0.002**	0.031	0.220	0.219	0.183	0.131	0.683	0.255
SigLIP	0.051	0.020	0.131	0.426	0.578	0.587	0.857	0.888	0.258
CVCL	0.005	0.027	0.147	0.468	0.655	0.667	1.170	0.911	0.258

Model	Semantics								
	WAT (↓)					VOC (↑)			THINGS (↑)
	5	6	8	10	A	0.3	0.8	1.6	A
CLIP-base	0.052	0.079	0.608	0.714	1.023	-0.372	-0.077	0.207	**0.397**
CLIP-large	0.052	0.079	0.608	0.714	1.023	-0.250	-0.038	0.302	0.246
ViLT	0.052	0.079	0.608	0.714	1.023	-0.106	-0.020	-0.033	0.127
FLAVA	0.052	0.079	0.608	0.714	1.023	-0.426	0.094	0.207	0.189
BLIP	0.051	**0.079**	0.608	0.714	1.023	-0.078	-0.237	0.002	0.185
BridgeTower	0.052	0.079	**0.608**	**0.713**	1.023	-0.330	-0.064	0.108	0.345
OpenCLIP-H	**0.051**	0.079	0.608	0.714	**1.023**	**-0.021**	-0.010	0.125	0.227
SigLIP	0.052	0.079	0.608	0.714	1.023	-0.179	-0.068	0.161	0.192
CVCL	0.052	0.079	0.608	0.714	1.023	-0.198	**0.189**	**0.423**	0.175

Table 7: Model accuracies across all tasks, averaged across ages.

Model	Lexicon		Syntax	
	LWL	VV	TROG	WG
CLIP-base	0.987	0.958	0.462	0.608
CLIP-large	0.987	0.966	0.372	0.594
ViLT	1.000	0.790	0.449	0.655
FLAVA	0.987	0.941	0.308	0.605
BLIP	1.000	0.958	0.603	0.497
BridgeTower	1.000	0.832	0.628	0.731
OpenCLIP-H	1.000	0.975	0.487	0.579
SigLIP	1.000	0.924	0.423	0.541
CVCL	0.592	0.328	0.231	0.281

Table 8: Correlations between model features and model–human similarities across all tasks. Task performance is split by human age bins (in years). A: Adult.

Feature	Lexicon							Syntax	
	LWL			VV				TROG	WG
	1.5	2	2.5	4	7	10	A	11	A
Accuracy	0.907	0.971	0.993	0.965	0.987	0.993	0.998	0.928	0.822
Size	0.831	0.818	0.711	0.811	0.837	0.844	0.854	0.434	0.254
Training	0.486	0.598	0.410	0.553	0.602	0.627	0.694	0.136	-0.065

Model	Semantics								
	WAT					VOC			THINGS
	5	6	8	10	A	0.3	0.8	1.6	A
Size	0.631	0.334	0.564	0.391	0.737	0.045	-0.426	-0.320	0.250
Training	0.689	-0.064	0.172	-0.041	0.834	0.154	-0.194	-0.045	0.277

Table 9: Generation model performance across all tasks. Arrows indicate the direction of better performance (i.e., lower is better vs. higher is better). Bolded results indicate most human-like result on a task.

Model	# params	# images	Lexicon							Syntax	
			LWL (↓)			VV (↓)				TROG (↓)	WG (↓)
			1.5	2	2.5	4	7	10	A	11	A
Next-token prediction method											
LLaVA	7B	1.31M	0.052	0.030	0.162	0.468	0.656	0.671	1.176	0.910	0.258
TinyLLaVA	3.1B	102K	0.035	**0.002**	**0.064**	**0.246**	**0.257**	**0.229**	**0.188**	**0.410**	**0.202**
Kosmos-2	1.6B	90M	0.053	0.022	0.117	0.468	0.656	0.672	1.176	0.905	0.258
moondream2	1.9B	NA	**0.023**	0.007	0.155	0.450	0.634	0.643	1.126	0.757	0.254
CogVLM	17B	1.5B	0.059	0.029	0.149	0.447	0.612	0.619	1.063	0.868	0.237
Log-likelihood measurement method											
LLaVA	7B	1.31M	0.059	0.030	0.171	0.468	0.656	0.671	1.176	0.910	0.258
TinyLLaVA	3.1B	102K	**0.043**	0.017	**0.075**	**0.402**	0.560	0.562	**0.971**	**0.738**	**0.220**
Kosmos-2	1.6B	90M	0.051	**0.013**	0.146	0.415	**0.553**	**0.557**	0.975	0.781	0.225
moondream2	1.9B	NA	0.059	0.030	0.174	0.468	0.656	0.671	1.175	0.910	0.258
CogVLM	17B	1.5B	0.059	0.018	0.174	0.468	0.657	0.672	1.177	0.908	0.234

Table 10: Model accuracies across all tasks, averaged across ages.

Model	Lexicon		Syntax	
	LWL	VV	TROG	WG
Next-token prediction method				
LLaVA	0.579	0.176	0.205	0.509
TinyLLaVA	1.000	0.958	0.782	0.728
Kosmos-2	0.750	0.252	0.333	0.497
moondream2	0.816	0.345	0.436	0.529
CogVLM	0.553	0.437	0.256	0.538
Log-likelihood measurement method				
LLaVA	0.487	0.202	0.231	0.497
TinyLLaVA	0.855	0.462	0.474	0.614
Kosmos-2	0.632	0.479	0.436	0.547
moondream2	0.461	0.210	0.154	0.474
CogVLM	0.461	0.235	0.179	0.500

https://doi.org/10.52202/079017-2462