Grounded Concreteness: Human-Like Concreteness Sensitivity in Vision–Language Models

Aryan Roy, Zekun Wang and Christopher J. MacLellan
Georgia Institute of Technology
aroy389@gatech.edu, zekun@gatech.edu, cmaclell@gatech.edu

Abstract

Do vision–language models (VLMs) develop more human-like sensitivity to linguistic concreteness than text-only large language models (LLMs) when both are evaluated with text-only prompts? We study this question with a controlled comparison between matched Llama text backbones and their Llama Vision counterparts across multiple model scales, treating multimodal pretraining as an ablation on perceptual grounding rather than access to images at inference. We measure concreteness effects at three complementary levels: (i) output behavior, by relating question-level concreteness to QA accuracy; (ii) embedding geometry, by testing whether representations organize along a concreteness axis; and (iii) attention dynamics, by quantifying context reliance via attention-entropy measures. In addition, we elicit token-level concreteness ratings from models and evaluate alignment to human norm distributions, testing whether multimodal training yields more human-consistent judgments. Across benchmarks and scales, VLMs show larger gains on more concrete inputs, exhibit clearer concreteness-structured representations, produce ratings that better match human norms, and display systematically different attention patterns consistent with increased grounding.

1 Introduction

Human meaning is not uniformly "linguistic": some concepts are tightly linked to perception and action (e.g., *apple, run*), while others are largely relational and context-dependent (e.g., *stronger, justice*). A long tradition in cognitive science treats *concreteness* as a graded dimension of conceptual representation, with concrete words benefiting from richer sensory codes and exhibiting robust behavioral advantages over abstract words (Paivio, 1990; Barsalou, 2008). Concreteness therefore offers a rare bridge between cognitive theory (how humans represent meaning) and computational diagnostics (how models encode and use meaning),

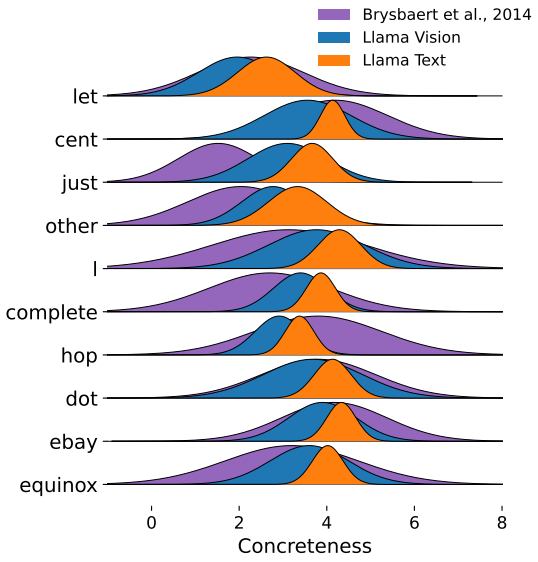


Figure 1: Comparison of concreteness rating distributions for selected words. For each word, we plot the empirical distribution of model-generated token ratings from Llama Vision (VLM) and Llama Text (LLM), alongside human norms from Brysbaert et al. (2014).

enabling measurable tests of *cognitive alignment* between humans and modern language systems (Coltheart, 1981; Brysbaert et al., 2014). More broadly, recent work in language acquisition argues that neural models can serve as hypothesis generators and testers for cognitive theories, provided we design analyses that connect internal mechanisms to behavioral signatures (Portelance, 2022a).

A central tension is that contemporary large language models (LLMs) learn from text alone, raising questions about whether "meaning" can be recovered from form without grounding (Harnad, 1990; Bender and Koller, 2020). While distributional learning can capture many semantic regularities, the absence of perceptual experience may be especially consequential for *concreteness*: in humans, concrete concepts are supported by sensorimotor simulations and imagery-like codes (Paivio, 1990;

Findings of the Association for Computational Linguistics: ACL 2026, pages 41934–41950

Barsalou, 2008). Vision-language models (VLMs) offer a natural testbed for this debate. By aligning text with visual representations (e.g., CLIP-style contrastive learning and cross-modal projector alignment), VLMs may develop more human-like, graded concreteness representations than comparably sized text-only LLMs (Radford et al., 2021; Alayrac et al., 2022; Liu et al., 2023; Touvron et al., 2023). Yet prior evidence connecting concreteness to model behavior and representations is difficult to interpret as a *vision effect*. On one hand, multimodal/distributional semantics work links concreteness to perceptual grounding and visual consistency (Hill et al., 2014; Hessel et al., 2018; Mickus et al., 2023); on the other hand, separate lines probe concreteness as an axis in embedding spaces using text-derived features (Charbonnier and Wartena, 2019; Wartena, 2024). However, these strands rarely provide a controlled ablation that isolates the contribution of visual input, and they typically analyze a single level (task performance *or* representations) rather than jointly linking behavior, geometry, and processing dynamics.

This motivates a controlled LLM–VLM ablation in which the language backbone is held as comparable as possible and the primary difference is access to vision, allowing the contribution of visual grounding to be isolated. In this work, concreteness awareness is evaluated under this ablation by triangulating evidence across three complementary levels of analysis. First, at the **output level**, we ask whether VLM accuracy on question answering increases with the concreteness of the queried concepts. We further elicit model-produced concreteness ratings (token-level) and measure their alignment with human norms (Figure 1) (Coltheart, 1981; Brysbaert et al., 2014). Second, at the **embedding level**, we test whether token representations organize by graded concreteness by projecting token-level embeddings into a low-dimensional space (t-SNE) and measuring within-bin compactness via intra-cluster dispersion across concreteness bins. Third, at the **attention level**, we quantify contextual dependence as the entropy of each token's self-attention distribution: if abstract meaning is more compositionally supported by surrounding context, abstract tokens should exhibit broader, higher-entropy attention, whereas concrete tokens should exhibit sharper, lower-entropy attention concentrated on fewer positions. This prediction aligns with classic context availability accounts of concreteness effects in psycholinguistics, which argue that abstract words benefit more from supportive context for comprehension than concrete words (Schwanenflugel and Shoben, 1983; Schwanenflugel et al., 1992).

Our analyses are designed as a vision ablation for semantic grounding: holding the base language model family and scaling regime as constant as possible, we ask what changes to the language models when vision is introduced. This yields a coherent story linking classic cognitive accounts of concreteness—dual coding and grounded cognition (Paivio, 1990; Barsalou, 2008)—to measurable signatures in modern foundation models: (i) *behavioral sensitivity* to concreteness in downstream QA, (ii) *representational geometry* that recovers a concreteness ordering, and (iii) *contextual dependence* reflected by attention entropy.

Contributions. We make following contributions: (1) a controlled LLM–VLM ablation study that isolates the effect of visual grounding on concreteness awareness across matched model families and scales; (2) output-level evidence that tests concreteness sensitivity in QA and quantifies alignment between model-elicited concreteness ratings and human norms; (3) internal diagnostics connecting grounding-based accounts to model representations and processing, including concreteness-conditioned clustering in embedding space (t-SNE with intra-cluster dispersion) and attention-entropy measures of contextual dependence motivated by context-availability theories of abstract meaning.

2 Related work

2.1 Measuring concreteness in language

Concreteness is a graded psycholinguistic property that captures how directly a concept can be experienced through the senses, and it has been extensively measured through human norming studies. Classic resources such as the MRC Psycholinguistic Database provide lexical attributes including concreteness/tangibility judgments (Coltheart, 1981), and later large-scale norms substantially expand coverage and improve reliability for modern evaluation settings (Brysbaert et al., 2014).

In parallel, computational work has proposed algorithmic approximations of concreteness. Early NLP approaches connected concreteness-related cues to figurative language phenomena, using concrete vs. abstract contextual signals for literal/metaphorical sense identification (Turney et al.,

2011). More recent methods treat concreteness prediction as a supervised estimation problem over distributional features and contextual representations (Charbonnier and Wartena, 2019). A particularly relevant line incorporates visual information: "visual concreteness" can be operationalized via cross-image consistency within multimodal datasets (Hessel et al., 2018), and visually grounded learning objectives can shape linguistic structure and representations (Shi et al., 2019). Given known context sensitivity (a word's perceived concreteness can shift with discourse and reference), human norms provide an external anchor for evaluation, while model-produced ratings can be treated as context-conditioned distributions rather than fixed type-level attributes.

2.2 Grounding and vision language models

Grounding-based accounts of meaning emphasize that linguistic symbols ultimately connect to perception and action, motivating the classic symbol-grounding problem (Harnad, 1990). In NLP, grounding has been studied through both definitional discussions and benchmark/task design, with critiques highlighting that "grounding" can mean different things depending on modality, interaction, and evaluation protocol. This motivates evaluating grounding beyond downstream success rates, using complementary diagnostics that probe internal representations and processing rather than relying on task behavior alone (Bisk et al., 2020a; Chandu et al., 2021; Mickus et al., 2023).

Modern VLMs provide a scalable route to grounding by learning joint representations of text and vision. Several models are explicitly designed to encourage fine-grained grounding via cross-modal alignment objectives (e.g., word/phrase–region alignment) (Tan and Bansal, 2019; Chen et al., 2020), and to evaluate grounded lexical acquisition beyond standard downstream transfer (Ma et al., 2023). Large-scale contrastive and generative pretraining has produced general-purpose models that align linguistic descriptions with visual features, supporting transfer to many multimodal tasks (Radford et al., 2021; Alayrac et al., 2022; Liu et al., 2023).

These strands motivate treating vision as a causal factor that can strengthen concreteness awareness. If concrete concepts are more consistently tied to perceptual regularities (i.e., they have more stable visual correlates), then adding visual supervision should preferentially benefit how models recognize and represent concreteness.

Building on this premise, prior work has established foundational links between multimodal pretraining and the improved processing of concrete language. Alper et al. (2023) demonstrate that vision-and-language pretraining enhances zero-shot visual language understanding in encoder models, showing a clear behavioral advantage over purely unimodal baselines. Similarly, Yanuka et al. (2024) leverage this phenomenon for multimodal dataset curation, demonstrating that visually grounded models strongly correlate with human concreteness judgments at both the lexical and sentence levels. While these studies provide critical behavioral evidence that multimodal grounding enriches the semantic representation of physical concepts, they primarily evaluate downstream task outputs. Our study extends this line of inquiry by transitioning the focus from external behavioral benchmarking to internal mechanistic interpretability, bridging the gap between multimodal grounding and the cognitive probing methodologies discussed next.

A critical consideration when evaluating vision-language models is the impact of multimodal alignment on the foundational text representations. Historically, several studies have observed an alignment tax, where extending a language model with visual supervision degrades its performance on purely text-based natural language understanding tasks (Iki and Aizawa, 2021; Madasu and Lal, 2023; Singh et al., 2022). However, recent methodological advances utilizing strictly controlled LM-VLM minimal pairs, such as the Molmo architecture (Deitke et al., 2025), have begun to challenge this consensus. For instance, Qin et al. (2025) explicitly compare matched language and vision-language backbones and demonstrate that multimodal training can actively improve text-only performance, specifically regarding the deployment of taxonomic knowledge. Our work aligns with and extends this minimal-pair paradigm. By evaluating matched architectures, we isolate the multimodal tuning phase to demonstrate that vision-language training does not generically harm textual representations; rather, it provides a highly targeted benefit to the processing and internal geometric organization of concrete, perceptually grounded concepts.

2.3 Neural models as cognitive probes of language learning and processing

A growing cognitive-science perspective treats neural networks as tools for generating and testing mechanistic hypotheses about human learning, rather than purely as engineering solutions (Portelance, 2022b). This includes using language models as "psycholinguistic subjects," evaluating whether model-based surprisal and state representations predict human processing difficulty and syntactic expectations (Goodkind and Bicknell, 2018; Futrell et al., 2019). Another line tests whether models acquire human-relevant grammatical generalizations via targeted syntactic evaluations and minimal-pair benchmarks (Linzen et al., 2016; Gulordava et al., 2018; Warstadt et al., 2020). More recent work pushes toward developmental plausibility by constraining data and supervision (e.g., BabyLM) or by studying interactive learning dynamics (Warstadt et al., 2023; Ma et al., 2025).

A highly relevant dimension of this cognitive probing literature examines the direct alignment between neural attention mechanisms and human visual attention during reading. Prior work establishes that neural attention weights correlate significantly with human eye-tracking metrics, such as fixation durations and gaze patterns, during machine reading comprehension tasks (Sood et al., 2020a). Furthermore, researchers have successfully integrated these human gaze signals as a supervisory signal in neural attention layers to improve performance across various downstream natural language processing tasks (Sood et al., 2020b). Building on this correspondence, recent research demonstrates that explicitly guiding or masking transformer attention using real-life or model-predicted eye-tracking data can actively enhance model capabilities in question-answering (Zhang and Hollenstein, 2024). Because our methodology utilizes layer-wise attention entropy to quantify how models distribute contextual reliance, this established neural-gaze correspondence directly supports our treatment of internal attention dynamics as a valid empirical proxy for human cognitive processing effort. Together, these efforts motivate treating representational properties of language models as empirical objects for studying human cognitive constructs.

3 Experiment Setup

Models The study compares matched pairs of text-only LLMs and vision-language models (VLMs) at two parameter scales. For the LLMs, the backbone is Meta's Llama 3.1 family (an 8B and a 70B Text-only Model) (Meta AI, 2024a). For VLMs, the corresponding vision models from the Llama Vision 3.2 family (an 11B and a 90B vision LLM) (Meta AI, 2024b) were chosen that utilize the Llama 3.1 text-only models as a backbone. The text-only models are fitted with a vision adapter and then trained on a multimodal dataset to create the vision models. We refer Appendix A for details. Unless otherwise stated, evaluation uses text-only prompts (no images), so the LLM–VLM comparison functions as an ablation on *access to visual supervision during training* rather than access to images at inference.

Measuring concreteness Token-level concreteness $C(w)$ is obtained from the human ratings in the 40k English words (Brysbaert et al., 2014) (40K). For each word that appears in 40K, its concreteness score is set to the corresponding 40K mean rating. For out-of-vocabulary proper nouns (e.g., named entities) that are not covered by 40K, the score is set to the maximum concreteness value on the 40K scale (5). For function words without a clear concreteness interpretation (e.g., articles and prepositions) that are also absent from 40K, the score is set to 0. Sentence-level concreteness for an input string x with word tokens $w_{1:n}$ is the mean of token scores:

$$C(x) = \frac{1}{n} \sum_{i=1}^{n} c(w_i). \quad (1)$$

For subword-tokenized model inputs, word-level scores are propagated to constituent sub-tokens to enable tokenwise analyses.

Text datasets Evaluation uses standard text-only QA benchmarks covering diverse reasoning demands and a broader range of question concreteness: ARC-Easy and ARC-Challenge for grade-school science multiple-choice questions (Clark et al., 2018); BoolQ for naturally occurring yes/no questions (Clark et al., 2019); WinoGrande for adversarial pronoun/coreference resolution (Sakaguchi et al., 2020); CommonsenseQA for commonsense multiple-choice QA (Talmor et al., 2019); Social IQA for reasoning about social interactions and implications (Sap et al., 2019) and

PIQA for physical commonsense reasoning (Bisk et al., 2020b). Performance is measured by accuracy under a unified prompting format. We refer Appendix B and C for details on datasets and prompts.

3.1 Research questions and methods

This section describes how each hypothesis is operationalized and tested. All analyses are conducted for both model scales to assess scaling effects.

Does the VLM outperform the LLM on QA questions as question concreteness increases? For each benchmark dataset, each question is scored as correct or incorrect under a unified prompting format. To summarize performance as a function of concreteness, sentence-level concreteness scores are pooled across all datasets and discretized into six equal-width bins of size 0.6, spanning $[1.8,\ 4.8]$ (the observed range is $[1.96,\ 4.67]$). For each bin, accuracy is computed as the mean correctness over questions whose sentence concreteness falls in that interval. In addition, the bin-wise accuracy gap between the VLM and its matched LLM is reported, $\Delta\mathrm{Acc} = \mathrm{Acc}_{\mathrm{VLM}} - \mathrm{Acc}_{\mathrm{LLM}}$, to quantify where vision provides an advantage. We hypothesize that ΔAcc is expected to be larger in higher-concreteness bins, indicating that the VLM is relatively more robust on concrete questions than the text-only model.

Do VLM token representations exhibit tighter within-concreteness clusters than LLM token representations? Each word (w) covered by 40K is rounded to a discrete concreteness bin $b(w) \in \{1, \ldots, 5\}$. For each model, we extract last-layer contextual representations and average over occurrences to obtain a type vector $\bar{\mathbf{h}}(w)$. To measure within-bin dispersion, we fit $\bar{\mathbf{h}}(w)$ with 2D t-SNE, yielding $\mathbf{z}(w)$. Within this t-SNE space, dispersion is measured as the mean pairwise cosine distance among tokens with the same label, where lower values indicate more compact clusters:

$$D \;=\; \mathbb{E}_{\ell}\, \mathbb{E}_{w \neq w' \sim \mathcal{W}_{\ell}} \big[1 - \cos\big(\mathbf{z}(w), \mathbf{z}(w')\big)\big] \quad (2)$$

where $\mathcal{W}_{\ell}$ represents the discretized concreteness bin. Lower D indicates tighter within-concreteness clusters.

Do abstract tokens exhibit higher-entropy attention distributions than concrete tokens, and is this abstract–concrete separation sharper in VLMs? For each layer ℓ and head h, self-attention weights follow the standard Transformer definition (Vaswani et al., 2017):

$$\mathbf{A}^{(\ell,h)} = \mathrm{softmax}\left(\frac{\mathbf{Q}^{(\ell,h)}\mathbf{K}^{(\ell,h)\top}}{\sqrt{d_k}}\right),$$

where $\mathbf{A}^{(\ell,h)}_{i,j}$ is the attention paid by token i to token j and forms a probability distribution over j due to softmax normalization. For each token i, attention entropy is computed as:

$$H^{(\ell,h)}(i) = -\sum_{j} \mathbf{A}^{(\ell,h)}_{i,j} \log \mathbf{A}^{(\ell,h)}_{i,j}. \quad (3)$$

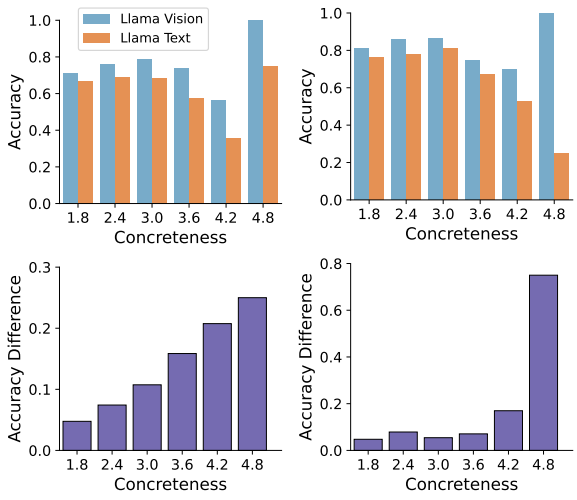
Entropy is then averaged across heads for each layer to obtain a per-token entropy score. At each layer, we test the association between token concreteness $c(w_i)$ and attention entropy via Pearson's r. We expect a *negative* correlation ($r < 0$): abstract tokens should exhibit *higher* attention entropy (more diffuse context integration), whereas concrete tokens should exhibit *lower* entropy (more focused attention), consistent with concreteness effects in comprehension (Schwanenflugel and Shoben, 1983; Schwanenflugel et al., 1992). Moreover, we predict this effect is stronger in VLMs than LLMs (i.e., more negative r in VLMs), reflecting a sharper abstract–concrete separation.

Do model generated concreteness judgments align better with human norms for VLMs? Each model is prompted to output a concreteness rating for every word token in each question on the 40K scale. We refer Appendix C for prompt details. To elicit reliable concreteness judgments and enforce a consistent output format, we use the `instruct` variants of the larger models (70B text-only and 90B vision-language). Because the same word type can appear in multiple contexts, each word w induces an empirical distribution over ratings under a model m:

$$p_m(r \mid w) \propto \sum_{x \in \mathcal{X}(w)} \mathbf{1}[\hat{r}_m(w, x) = r] \quad (4)$$

where $\mathcal{X}(w)$ are contexts containing w and $\hat{r}_m(w, x)$ is the model-produced rating for token w in question x. We construct an analogous human distribution $p_H(r \mid w)$ from the 40K annotations and quantify human–model agreement with the symmetric KL divergence:

$$D_{\mathrm{KL}}(w) = \tfrac{1}{2}[\mathrm{KL}(p_m \| p_H) + \mathrm{KL}(p_H \| p_m)] \quad (5)$$



(a) 8B model family. (b) 70B model family.

Figure 2: Top row: accuracy by question concreteness for Llama Text vs. Llama Vision. Bottom row: the VLM–LLM accuracy gap.

where smaller $D_{\mathrm{KL}}(w)$ indicates better alignment. We expect (i) $D_{\mathrm{KL}}(w)$ decreases as human concreteness increases, and (ii) this decrease is steeper for VLMs than for LLMs. To test the trend, we bin words by human concreteness in 0.5-wide bins and regress binned D_{SKL} on bin center, reporting slope, R^2, and p-value.

4 Analysis and discussion

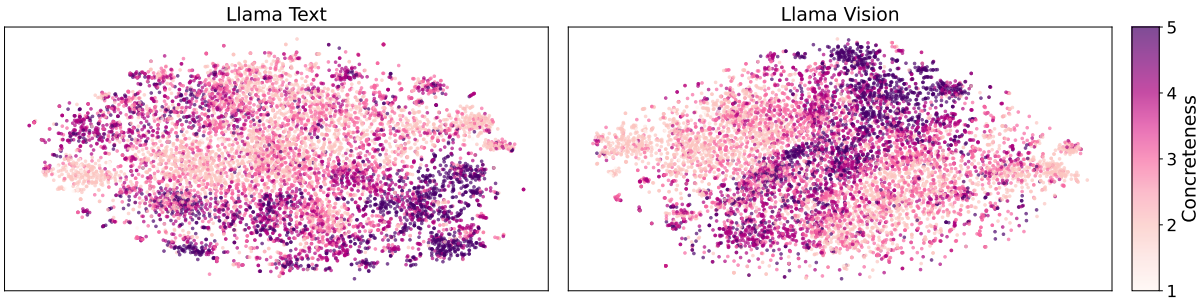
VLM outperform their LLM counterpart on QA questions and the gaps widen on more concrete questions. Figure 2 shows that, across all concreteness bins, VLMs consistently outperform their text-only counterparts at both scales. Across all datasets, for the smaller pair, accuracy increases from 68.0% (LLM) to 77.5% (VLM), and for the larger pair from 78.8% (LLM) to 85.5% (VLM). Since each VLM is trained from the same model family as its LLM counterpart, these gains indicate that multimodal training transfers to improved textual QA. We refer Appendix D for additional per-dataset results.

Crucially, the advantage is not uniform across question types: the bottom row of Figure 2 shows that the VLM–LLM gap increases with question concreteness in both scales, with the strongest separation in the most concrete bin. This pattern suggests that visual grounding disproportionately benefits questions whose successful resolution depends on perceptible entities, attributes, and events (e.g., shape, material, spatial relations), consistent with grounded accounts in which perceptual experience provides an additional scaffold for semantic representations (Harnad, 1990; Barsalou, 2008; Bisk et al., 2020a). A plausible mechanism is that vision–language training strengthens the association between concrete lexical items and perceptually anchored image features (e.g., object properties and spatial configurations), making the relevant evidence easier to retrieve and compose when answering concrete questions. In other words, the VLM's gains appear concentrated where the QA signal can be supported by grounded semantics rather than purely symbolic co-occurrence.

At the same time, the effect is smaller (and sometimes flatter) in lower-concreteness bins, where performance may depend more on abstract relations, discourse-level inference, or world knowledge not directly supported by perceptual grounding. Moreover, abstract language tends to be more polysemous and context-dependent, which can reduce the benefit of any single additional modality. Together, these results suggest that multimodal training provides an asymmetric benefit: it reliably improves QA overall, but disproportionately improves the processing and use of concrete concepts, which we further probe via representation geometry and attention diagnostics in subsequent experiments.

VLM token representations form tighter within-concreteness clusters than LLMs. Given that VLMs outperform their text-only counterparts on QA, we ask whether vision-text training also reshapes the *geometry* of token representations along a graded concreteness dimension. Figure 3a provides a qualitative view: in both the smaller and larger models, the VLM embeddings display a visibly tighter and less dispersed cluster for the highest concreteness category (bucket 5). This effect is more pronounced in the larger model. This pattern suggests that visual grounding encourages representations of perceptually grounded words to occupy a more coherent subregion of the space, consistent with grounded accounts of meaning and the symbol-grounding perspective (Harnad, 1990; Barsalou, 2008; Bisk et al., 2020a).

We quantify this effect using within-bin intra-cluster dispersion (Eq. 2) computed in the t-SNE space. Table 1 shows that VLMs achieve lower dispersion than LLMs at *every* concreteness level in both families. The effect is largest for the most concrete bin (c=5.0): dispersion drops from 0.76→0.66 in the 8B family and from 0.87→0.77



(a) 8B model family.

(b) 70B model family.

Figure 3: t-SNE of average last-layer token representations for Llama Text vs. Llama Vision, colored by human concreteness.

in the 70B family. Notably, dispersion is highest in the mid-concreteness range (roughly $c \in [2, 4]$) and drops sharply for the most concrete words, which is compatible with the idea that mid-range words are more heterogeneous (e.g., broader senses or mixed perceptual/abstract usage) while highly concrete words admit more stable, visually grounded semantics. The analysis and visualization in figure 3 and table 1 were done on t-SNE projections of the actual embeddings, for results on actual embeddings we refer to Appendix F.

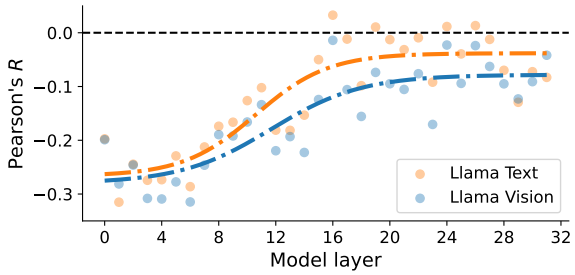
Taken together, the qualitative structure in Figure 3a and the consistent quantitative reductions in Table 1 support our hypothesis that VLM representations encode graded concreteness more cleanly than LLMs. This geometry offers a representational account of the concreteness-dependent QA gains. If highly concrete word types occupy a tighter region of the space, their representations are more consistent across contexts, reducing the need for context-dependent disambiguation and making grounded attributes easier to retrieve and compose. In turn, this should improve robustness on questions that hinge on perceptual properties (e.g., materials, shapes, spatial relations).

Concrete tokens show lower attention entropy, with a stronger effect in VLMs. Motivated by the representational results showing tighter within-concreteness clustering in VLMs, we next test whether models also differ in how much context they integrate for abstract versus concrete words. Prior psycholinguistic work suggests that concrete words tend to be easier to interpret and rely less on contextual support than abstract words, which are more context-dependent (Schwanenflugel and Shoben, 1983; Schwanenflugel et al., 1992). We operationalize contextual reliance using attention entropy (Eq. 3): higher entropy corresponds to more diffuse attention over many tokens, while lower entropy reflects more concentrated attention.

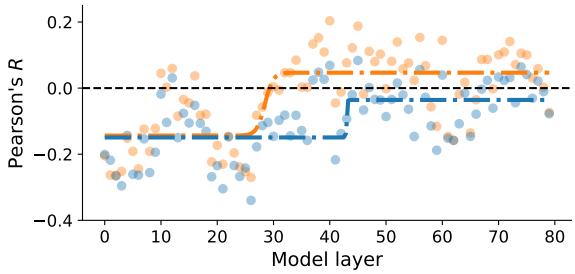
Figure 4 plots, for each layer, Pearson's r between token concreteness and token attention entropy (head-averaged), for both model scales. We report full Pearson's R values and their $p-$values in Appendix G. Across most layers, correlations are negative (or near zero), indicating that more

	8B model family		70B model family	
Conc.	**Text-only ↓**	**Vision ↓**	**Text-only ↓**	**Vision ↓**
1.0	0.87	**0.75**	0.93	**0.82**
2.0	0.94	**0.87**	0.98	**0.94**
3.0	0.98	**0.96**	1.00	**0.99**
4.0	0.99	**0.96**	1.00	**0.99**
5.0	0.76	**0.66**	0.87	**0.77**

Table 1: Within-concreteness cluster dispersion (mean pairwise cosine distance in 2D t-SNE; lower is tighter).



(a) 8B model family.



(b) 70B model family.

Figure 4: Layerwise Pearson's r between token concreteness and head-averaged attention entropy. Colored dash lines are sigmoid fitted curves.

concrete tokens exhibit lower attention entropy (i.e., they attend more selectively) whereas abstract tokens show higher entropy, consistent with the hypothesis that abstract meaning requires broader contextual integration. The effect is strongest in earlier-to-mid layers: both model families display more negative correlations in roughly the first half of the network, followed by a gradual attenuation toward later layers. This layerwise result suggests that context concreteness sensitivity is primarily expressed early in processing, while later layers may shift toward task-level integration that is less directly tied to lexical concreteness.

Importantly, VLMs show a consistently stronger negative correlation than their text-only counterparts. Averaging Pearson's r across layers yields $r = -0.12$ (LLM) vs. $r = -0.16$ (VLM) for the smaller models, and $r = -0.02$ (LLM) vs. $r = -0.10$ (VLM) for the larger models, indicating a sharper abstract–concrete separation in VLM attention behavior. One interpretation is that vision-text training provides an additional grounding signal that stabilizes the representations of concrete words, allowing the model to resolve them with more focused attention (lower entropy) and reducing the need to distribute attention broadly across other tokens. The results so far were obtained from the Llama model pairs. For additional discussion on Qwen model pairs, we refer to Appendix E.

VLM concreteness ratings align more closely with humans than LLM as concreteness increases. Finally, if the performance, representation geometry, and attention analyses reflect a human-like graded concreteness sensitivity, we should also observe human-aligned concreteness judgments from the models. Figure 5 plots human concreteness against human–model agreement measured by symmetric KL divergence (lower is better). Overall, the VLM exhibits lower divergence than the LLM (mean D_{KL}: 9.4 vs. 10.1), indicating closer alignment to human norms from (Brysbaert et al., 2014).

More importantly, agreement improves with concreteness: as human ratings increase, D_{KL} decreases for both models, but the trend is substantially sharper for the VLM. A linear fit on binned concreteness shows that the VLM's alignment increases reliably with concreteness (slope$= -1.810$, $R^2 = 0.857$, $p < 0.001$), whereas the LLM trend is weaker and not statistically reliable (slope$=$ -1.048, $R^2 = 0.328$, $p > 0.1$). This suggests that vision-text training does not merely shift ratings globally, but preferentially calibrates judgments for perceptually grounded words—precisely where vision provides an additional supervisory signal.

Because concreteness is an interpretable, human-normed semantic axis, improved human–model alignment makes model behavior easier to diagnose: it supports using concreteness as a principled factor for error analysis (e.g., when failures concentrate in abstract language) and as an interpretable control variable when comparing model families, aligning with calls for more rigorous, task-relevant interpretability evaluations (Doshi-Velez and Kim, 2017; Guidotti et al., 2019). Mechanistically, concreteness alignment provides a concrete target for circuit-level analysis: one can localize where concreteness enters computation (layers/heads/MLP features) and test causal interventions, complementing transformer reverse-engineering frameworks (Elhage et al., 2021; Geva et al., 2022).

Model concreteness behaviour carries through scale. Across all analyses, the concreteness effects observed in the 8B model family qualitatively mirror in the 70B family: VLMs show larger gains on concrete QA, tighter within-concreteness clustering, more negative concreteness–entropy correlations, and stronger human-aligned rating trends. This consistency suggests that concreteness organi-

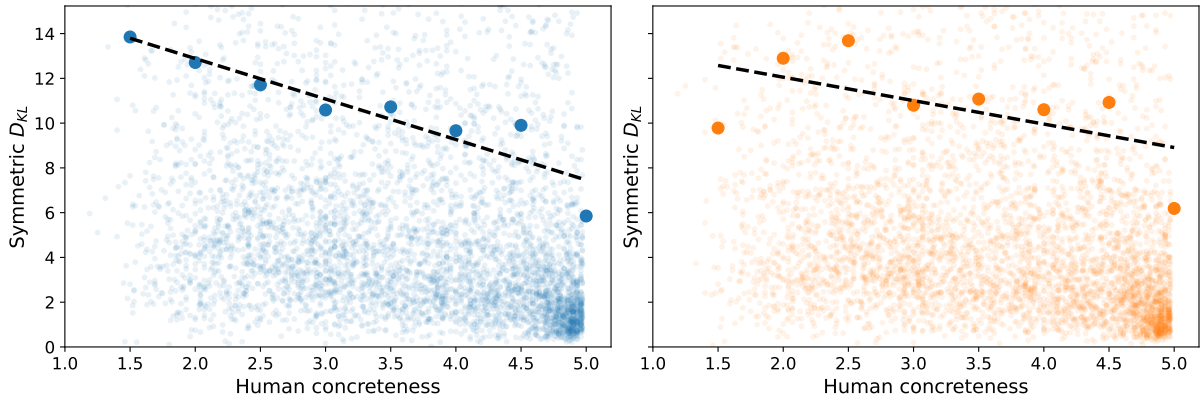


Figure 5: Human–model alignment of token-level concreteness judgments. Each point is a word with human concreteness on the x-axis and symmetric KL divergence between the model's rating distribution and the human (40K) distribution on the y-axis (lower is better). Larger dots show bin averages; dashed lines are linear fits over bins. The VLM (left) exhibits lower divergence and a steeper decrease in divergence with concreteness than the matched text-only LLM (right).

zation is not specific to small models, but a stable property that persists under scaling, with vision-text training providing an additive grounding signal rather than a scale-specific artifact.

5 Conclusion

We presented a controlled test of whether visual supervision during training induces more human-like concreteness sensitivity in foundation models. Holding the language backbone family and scaling regime as constant as possible and using text-only evaluation prompts, we compared matched Llama 3.1 LLMs against their Llama Vision 3.2 counterparts, treating the LLM–VLM contrast as an ablation on access to perceptual grounding rather than access to images at inference. Across diverse QA benchmarks, VLMs achieved higher accuracy overall and, crucially, showed larger improvements on questions with higher concreteness, consistent with grounded cognition and dual-coding accounts in which perceptual experience disproportionately supports concrete semantics (Paivio, 1990; Barsalou, 2008; Harnad, 1990). Internal analyses converged on a coherent mechanistic picture: VLM token representations formed tighter within-concreteness clusters in low-dimensional projections, suggesting more stable type-level semantics for highly concrete words; and attention-entropy diagnostics indicated a sharper abstract–concrete separation in contextual reliance, aligning with psycholinguistic theories that abstract meaning draws more heavily on supportive context (Schwanenflugel et al., 1992; Schwanenflugel and Shoben, 1983). Finally, elicited token-level concreteness ratings agreed more closely with human norms in VLMs, with a stronger improvement in human–model alignment as concreteness increased, indicating that vision–text training preferentially calibrates judgments precisely where perceptual supervision is informative.

Beyond the performance gap, our results position concreteness as a principled, interpretable axis for comparing model families and diagnosing grounding-related behavior. This provides a useful bridge between cognitive constructs and modern interpretability practice: concreteness offers a measurable target for localizing where grounded semantic information enters computation.

Limitations

Single model family. To provide initial evidence of cross-architecture generalizability, we conducted the attention entropy analysis on an additional minimal pair (Qwen2 and Qwen2-VL), which successfully replicated the stronger abstract-concrete separation observed in the vision model, as detailed in the Appendix. However, a full replication requires repeating the entire experimental pipeline, including downstream QA evaluation, embedding geometry, and human-alignment elicitation, on additional matched pairs. Evaluating architectures such as the Molmo and OLMo pair, or expanding on the multilingual Qwen family, will test whether these comprehensive concreteness effects persist across different model structures, languages, and training recipes (Bai et al., 2023; Wang et al., 2024).

Token frequency as a confound. Some apparent concreteness effects may be partly explained by lexical frequency/contextual diversity rather than grounding, but we cannot access true pretraining token counts. To reduce this confound, we can add frequency proxies (tokenizer-matched counts from large open corpora, plus average surprisal on held-out text) as covariates, or frequency-match concrete vs. abstract item sets before running the main regressions.

Developmental trajectory. We only analyze final checkpoints, so we cannot determine *when* concreteness sensitivity and human-alignment emerge or how this depends on scale. An immediate follow-up is a developmental-style study over intermediate checkpoints (or staged training) to track the concreteness–accuracy slope and internal separability over training time for each model scale.

References

Jean-Baptiste Alayrac, Jeffrey Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*.

Morris Alper, Michael Fiman, and Hadar Averbuch-Elor. 2023. Is bert blind? exploring the effect of vision-and-language pretraining on visual language understanding. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6778–6788.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Lawrence W. Barsalou. 2008. Grounded cognition. *Annual Review of Psychology*, 59:617–645.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020a. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020b. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911.

Khyathi Chandu, Siva Reddy, Alan W. Black, Yulia Tsvetkov, and Eric Nyberg. 2021. Grounding "grounding" in NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics.

Jean Charbonnier and Christian Wartena. 2019. Predicting word concreteness and imagery. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 176–187, Gothenburg, Sweden. Association for Computational Linguistics.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Max Coltheart. 1981. The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, and 31 others. 2025. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 91–104.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly,

Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):1–42.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1–3):335–346.

Jack Hessel, David Mimno, and Lillian Lee. 2018. Quantifying the visual concreteness of words and topics in multimodal datasets. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2194–2205, New Orleans, Louisiana. Association for Computational Linguistics.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014. Multi-modal models for concrete and abstract concept meaning. *Transactions of the Association for Computational Linguistics*, 2:285–296.

Taichi Iki and Akiko Aizawa. 2021. Effect of visual extensions on natural language understanding in vision-and-language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2189–2196, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Ziqiao Ma, Jiayi Pan, and Joyce Chai. 2023. World-to-words: Grounded open vocabulary acquisition through fast mapping in vision-language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 524–544, Toronto, Canada. Association for Computational Linguistics.

Ziqiao Ma, Zekun Wang, and Joyce Chai. 2025. Babysit a language model from scratch: Interactive language learning by trials and demonstrations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 991–1010, Albuquerque, New Mexico. Association for Computational Linguistics.

Avinash Madasu and Vasudev Lal. 2023. Is multimodal vision supervision beneficial to language? *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2637–2642.

Meta AI. 2024a. The llama 3.1 model family. Technical report and model cards for Llama 3.1 text-only models.

Meta AI. 2024b. The llama 3.2 model family. Includes Llama 3.2 Vision and Vision-Instruct model cards.

Timothée Mickus, Elaine Zosa, and Denis Paperno. 2023. Grounded and well-rounded: A methodological approach to the study of cross-modal and cross-lingual grounding. *arXiv preprint arXiv:2310.11938*.

Allan Paivio. 1990. *Mental Representations: A Dual Coding Approach*. Oxford University Press, New York, NY.

Eva Portelance. 2022a. *Neural Network Approaches to the Study of Word Learning*. Ph.D. thesis, Stanford University.

Eva Portelance. 2022b. Neural network approaches to the study of word learning. Ph.D. thesis, McGill University.

Yulu Qin, Dheeraj Varghese, Adam Dahlgren Lindström, Lucia Donatelli, Kanishka Misra, and Najoung Kim. 2025. Vision-and-language training helps deploy taxonomic knowledge but does not fundamentally alter it. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

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. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. WinoGrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Paula J. Schwanenflugel, Carolyn Akin, and Wei-Ming Luh. 1992. Context availability and the recall of abstract and concrete words. *Memory & Cognition*, 20(1):96–104.

Paula J. Schwanenflugel and Edward J. Shoben. 1983. Differential context effects in the comprehension of abstract and concrete verbal materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9(1):82–102.

Haoyue Shi, Jiayuan Mao, Kevin Gimpel, and Karen Livescu. 2019. Visually grounded neural syntax acquisition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1842–1861, Florence, Italy. Association for Computational Linguistics.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15617–15629.

Ekta Sood, Simon Tannert, Diego Frassinelli, Andreas Bulling, and Ngoc Thang Vu. 2020a. Interpreting attention models with human visual attention in machine reading comprehension. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 12–25, Online. Association for Computational Linguistics.

Ekta Sood, Simon Tannert, Philipp Müller, and Andreas Bulling. 2020b. Improving natural language processing tasks with human gaze-guided neural attention. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5099–5110, Hong Kong, China. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 680–690, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and 1 others. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at CoNLL 2023*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Christian Wartena. 2024. Estimating word concreteness from contextualized embeddings. In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, pages 81–88, Vienna, Austria. Association for Computational Linguistics.

Moran Yanuka, Morris Alper, Hadar Averbuch-Elor, and Raja Giryes. 2024. ICC : Quantifying image caption concreteness for multimodal dataset curation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11048–11064, Bangkok, Thailand. Association for Computational Linguistics.

Leran Zhang and Nora Hollenstein. 2024. Eye-tracking features masking transformer attention in question-answering tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7057–7070, Torino, Italia. ELRA and ICCL.

A Model and Compute Details

Model Architecture We evaluate models from the Llama 3.1 (text-only) and Llama 3.2 (vision-language) families. The vision-language models (VLMs) utilize the corresponding text-only models as their language backbone, augmented with a vision tower and cross-attention adapters. The architectural details for the text backbones are provided in Table 2.

For the vision models (Llama 3.2 11B and 90B), the vision tower is a ViT-H/14 based encoder. The 11B VLM utilizes the 8B text backbone, while the 90B VLM utilizes the 70B text backbone. All models use Grouped-Query Attention (GQA) and are trained with a context window of 128k tokens.

Compute Resources All inference and evaluation experiments were conducted on a cluster of 8 `NVIDIA A40` GPUs. Models were loaded in `bfloat16` precision to match the training dtype. The total compute budget for the evaluation of all benchmarks and concreteness scoring was approximately `2 Days` of GPU hours.

B Dataset Details

We utilize seven standard QA benchmarks covering distinct reasoning domains. Statistics for the evaluation splits used are detailed in Table 3. For datasets where the official test set is hidden (BoolQ, WinoGrande, CommonsenseQA, PIQA), we report results on the canonical validation/development split. For ARC, we use the test split. All datasets and model checkpoints are obtained from the public Hugging Face Hub (via the `datasets` and `transformers` libraries). The benchmarks consist of generic multiple-choice questions and do not require any user-provided inputs. We do not collect, store, or process personally identifying information. Accordingly, our experiments pose minimal risk of identity disclosure, and we report only aggregate accuracy metrics without releasing any per-example outputs that could contain sensitive content.

We calculated the exact lexical overlap with the Brysbaert et al. (2014) norming dataset. Across the core commonsense benchmarks (CQA, PIQA, SIQA, Winogrande), we observed a high coverage of 76% of unique content vocabulary. The specialized science (ARC) and encyclopedic (BoolQ) datasets showed lower coverage (65% and 45%, respectively), reflecting their higher density of abstract technical entities and proper nouns (for e.g. rutile, agonist, Rhimes) which appropriately lack general physical norms.

C Prompts

QA Evaluation (Non-Instruct Models) For the base (non-instruct) models, we utilized zero-shot prompting templates tailored to the task format. For multiple-choice datasets (ARC, CommonsenseQA, PIQA, WinoGrande), we used a standard completion format that lists options and prompts for an immediate answer:

```
You are a helpful assistant.
Answer the question immediately
with just the option letter (A,
B, C, D) or number.
Question: {question}
Options:
A. {choice_a}
B. {choice_b}
C. {choice_c}
D. {choice_d}
Answer:
```

For the reading comprehension dataset (BoolQ), we employed a template that conditions the binary response on the provided passage:

```
Read the following passage and
```

Model Family	Size	Layers	Hidden Dim	Attn Heads	KV Heads
Llama 3.1 (Text)	8B	32	4096	32	8
Llama 3.2 (Vision)	11B	32	4096	32	8
Llama 3.1 (Text)	70B	80	8192	64	8
Llama 3.2 (Vision)	90B	80	8192	64	8

Table 2: Architectural specifications for the language backbones used in this study. The VLM variants inherit these text specifications and add vision encoder parameters.

Dataset	Domain	# Questions	Avg. Length
ARC-Easy	Grade-school Science	2,376	∼39 words
ARC-Challenge	Grade-school Science (Hard)	1,172	∼47 words
BoolQ	Reading Comprehension (Yes/No)	3,270	∼120 words†
WinoGrande	Commonsense (Coreference)	1,267	∼32 words
CommonsenseQA	General Commonsense	1,221	∼26 words
PIQA	Physical Interaction	1,000	∼48 words
SIQA	Social Commonsense	1,954	∼36 words

Table 3: Summary of evaluation datasets. †Includes passage length.

```
answer the question with Yes or
No.

Passage: {passage}
Questions: {question}
Answer:
```

Concreteness Ratings To elicit concreteness ratings from the the large models (Llama 3.1 70B and Llama 3.2 90B Vision), we used the following prompt to ensure the output is aligned with the 1-7 MRC scale.

```
You are a psycholinguistics
expert. Your task is to rate the
'concreteness' of every content
word in the following text on a
scale from 1.0 (very abstract)
to 7.0 (very concrete/tangible).

Defintion:
- **Concrete words** refer to
things you can perceive directly
with your senses (touch, see,
hear, smell). Examples: 'apple',
'chair', 'scream'.
- **Abstract words** refer to
concepts, ideas, or emotions that
cannot be directly perceived.
Examples: 'freedom', 'justice',
'infinity'.

Input Text: "{text}"

Return your analysis strictly
as a JSON list of objects,
where each object has 'word' and
'score'. Ignore stop words (the,
a, is, etc.).
Example format:
[
{"word": "apple", "score": 6.2},
{"word": "freedom", "score":
2.77}
]
JSON Output:
```

D Detailed Results

Table 4 presents the raw accuracy scores for each model across all five datasets.

E Replication on Additional Model Pair

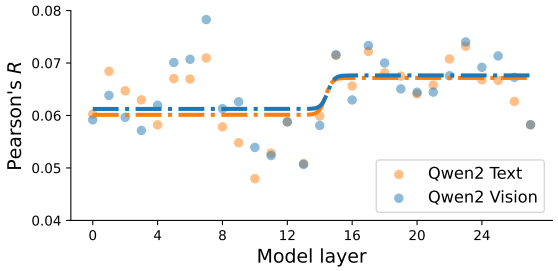


Figure 6: 7B Qwen model family.

To investigate the effects of architecture and training data on our observations, we ran the attention entropy experiment on an additional model pair. For this experiment we chose the Qwen model

Dataset	Small Category		Large Category	
	Llama 3.1 8B	Llama 3.2 11B	Llama 3.1 70B	Llama 3.2 90B
ARC-Easy	86.45%	**89.23%**	93.10%	**93.90%**
ARC-Challenge	65.70%	**78.41%**	89.08%	**91.64%**
BoolQ	76.18%	**82.02%**	85.02%	**90.31%**
WinoGrande	52.41%	**63.30%**	69.93%	**78.85%**
CommonsenseQA	61.51%	**71.58%**	76.17%	**77.15%**
PIQA	35.30%	**71.90%**	54.00%	**74.50%**
SIQA	64.43%	**71.19%**	65.10%	**78.71%**
Average				

Table 4: Per-dataset accuracy (%) for all evaluated models. The VLM variants generally perform comparable to or better than their text-only counterparts, despite receiving no visual input during this evaluation.

family, specifically the text-only and the multimodal version of Qwen2 7B. As the figure 6 indicates, the Qwen model pair does not exhibit a significant gap in correlation between concreteness and attention entropy across the models. We hypothesize that this difference occurs due to the different make-up of pre-training data used to train the model families. While the Llama family was trained on a largerly english based corpus, the Qwen models are trained on a pre-training mixture of Chinese and English. This means the qwen models are unable to represent the concreteness of tokens as well as the Llama family. While this is a negative result, it opens the door to future work that looks at concreteness in multilingual data.

F Intra-clustering Analysis on Raw Embeddings

We provide the intra-clustering analysis data on raw embeddings for the 8B Llama model pair in table 5. The analysis holds true for the raw embedding data as well.

G Attention Entropy Significance Analysis

To confirm that the observed divergence in attention entropy between the model pairs is systematic, we conducted paired t-tests on the layer-wise attention entropy differences. The evaluation compares the base LLM against its corresponding VLM counterpart across all transformer layers.

As detailed in Table 6, the mean layer-wise differences are highly statistically significant for both the 32-layer and 80-layer architectural scales ($p < 0.001$). These results formally validate that the multimodal alignment phase significantly alters the internal attention dynamics of the language backbone.

The following tables contains the raw correlation values from the Attention Entropy analysis for both model , Tables 7–9 contain the Pearson correlation (R) and Significance (p) with significance levels: $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 5: Intra-clustering Analysis on Raw Embeddings (11B vs 8B)

	Silhouette Mean		Intra Disp. (Cos)		Near. Cent. (Cos)		Count	
Concreteness	**11B**	**8B**	**11B**	**8B**	**11B**	**8B**	**11B**	**8B**
1.0	**0.0045**	-0.0108	**0.5522**	0.6697	**0.0148**	0.0246	118	116
2.0	**-0.0316**	-0.0373	**0.5707**	0.6807	**0.0099**	0.0146	2384	2392
3.0	**-0.0200**	-0.0286	**0.5514**	0.6532	**0.0099**	0.0146	2172	2194
4.0	**-0.0367**	-0.0442	**0.5482**	0.6417	**0.0110**	0.0165	1938	1928
5.0	0.0646	**0.0784**	**0.4973**	0.5690	**0.0156**	0.0237	1388	1370

Table 6: Paired t-test results comparing layer-wise attention entropy between base LLMs and VLMs. The divergence is statistically significant ($p < 0.001$) across both model scales.

	Llama 3.1 70B		Llama 3.2 90B	
Layer	R	p	R	p
Small (32 Layers)	32	0.038	6.63	2.08×10^{-7}
Large (80 Layers)	80	0.075	13.44	4.22×10^{-22}

Table 7: Large Scale Models (Part 1, Layers 0–31)

	Llama 3.1 70B		Llama 3.2 90B	
Layer	R	p	R	p
0	-0.20	***	-0.20	***
1	-0.26	***	-0.22	***
2	-0.26	***	-0.27	***
3	-0.25	***	-0.30	***
4	-0.15	***	-0.14	***
5	-0.19	***	-0.26	***
6	-0.24	***	-0.26	***
7	-0.12	***	-0.15	***
8	-0.19	***	-0.26	***
9	-0.12	***	-0.19	***
10	0.04	***	-0.02	0.09
11	0.00	0.82	-0.10	***
12	0.06	***	0.03	***
13	-0.08	***	-0.15	***
14	-0.03	***	-0.08	***
15	-0.04	***	-0.11	***
16	0.04	***	-0.05	***
17	-0.08	***	-0.11	***
18	-0.08	***	-0.13	***
19	-0.22	***	-0.27	***
20	-0.17	***	-0.23	***
21	-0.23	***	-0.30	***
22	-0.15	***	-0.15	***
23	-0.20	***	-0.23	***
24	-0.24	***	-0.27	***
25	-0.25	***	-0.24	***
26	-0.27	***	-0.34	***
27	-0.08	***	-0.18	***
28	-0.06	***	-0.11	***
29	-0.01	0.21	-0.10	***
30	0.00	0.74	-0.15	***
31	-0.01	0.58	-0.10	***

Table 8: Large Scale Models (Part 2, Layers 32–79)

	Llama 3.1 70B		Llama 3.2 90B	
Layer	R	p	R	p
32	0.05	***	-0.08	***
33	0.05	***	-0.14	***
34	0.08	***	-0.08	***
35	0.00	0.78	-0.13	***
36	0.00	0.84	-0.16	***
37	0.13	***	0.02	**
38	0.15	***	0.05	***
39	0.11	***	0.03	**
40	0.20	***	0.07	***
41	-0.05	***	-0.22	***
42	-0.01	0.27	-0.14	***
43	0.08	***	-0.09	***
44	0.12	***	-0.02	0.06
45	0.19	***	0.08	***
46	-0.02	0.09	-0.07	***
47	0.11	***	0.00	0.88
48	0.08	***	-0.03	***
49	0.10	***	-0.04	***
50	-0.00	0.87	-0.09	***
51	0.08	***	-0.03	***
52	0.14	***	0.02	0.15
53	0.06	***	-0.02	0.06
54	0.08	***	-0.06	***
55	0.02	**	-0.15	***
56	0.15	***	0.06	***
57	0.07	***	-0.03	***
58	-0.06	***	-0.11	***
59	-0.12	***	-0.19	***
60	0.14	***	0.04	***
61	-0.15	***	-0.15	***
62	-0.16	***	-0.16	***
63	-0.02	0.09	-0.11	***
64	0.02	0.11	-0.02	0.16
65	-0.14	***	-0.15	***
66	-0.04	***	-0.08	***
67	0.09	***	-0.00	0.78
68	0.09	***	0.02	**
69	-0.00	0.83	-0.06	***
70	0.10	***	0.03	***
71	0.11	***	0.02	**
72	0.14	***	0.03	***
73	0.07	***	-0.04	***
74	0.11	***	0.06	***
75	0.10	***	0.04	***
76	0.03	***	-0.03	***
77	0.06	***	0.02	**
78	0.00	0.82	-0.01	0.31
79	-0.07	***	-0.08	***

Table 9: Small Scale Models (Layers 0–31)

	Llama 3.1 8B		Llama 3.2 11B	
Layer	R	p	R	p
0	-0.20	***	-0.20	***
1	-0.32	***	-0.28	***
2	-0.24	***	-0.25	***
3	-0.27	***	-0.31	***
4	-0.27	***	-0.31	***
5	-0.23	***	-0.28	***
6	-0.29	***	-0.31	***
7	-0.21	***	-0.25	***
8	-0.17	***	-0.19	***
9	-0.17	***	-0.19	***
10	-0.13	***	-0.17	***
11	-0.10	***	-0.13	***
12	-0.18	***	-0.22	***
13	-0.18	***	-0.19	***
14	-0.15	***	-0.22	***
15	-0.05	***	-0.12	***
16	0.03	***	-0.01	0.19
17	-0.01	0.27	-0.11	***
18	-0.10	***	-0.16	***
19	0.01	0.32	-0.07	***
20	-0.01	0.24	-0.09	***
21	-0.03	***	-0.11	***
22	-0.01	0.38	-0.08	***
23	-0.09	***	-0.17	***
24	0.01	0.28	-0.02	**
25	-0.04	***	-0.09	***
26	0.01	0.22	-0.02	**
27	-0.01	0.25	-0.06	***
28	-0.07	***	-0.10	***
29	-0.13	***	-0.12	***
30	-0.07	***	-0.09	***
31	-0.08	***	-0.04	***