Does Visual Grounding Enhance the Understanding of Embodied Knowledge in Large Language Models?

Zhihui Yang[1] **Yupei Wang**[1] **Kaijie Mo**[1] **Zhe Zhao**[2] **Renfen Hu**[1*]
[1]Beijing Normal University [2]Tencent AI Lab
{yangzhihui, wangyupei, mokaijie, irishu}@mail.bnu.edu.cn
nlpzhezhao@tencent.com

Abstract

Despite significant progress in multimodal language models (LMs), it remains unclear whether visual grounding enhances their understanding of embodied knowledge compared to text-only models. To address this question, we propose a novel embodied knowledge understanding benchmark based on the perceptual theory from psychology, encompassing visual, auditory, tactile, gustatory, olfactory external senses, and interoception. The benchmark assesses the models' perceptual abilities across different sensory modalities through vector comparison and question-answering tasks with over 1,700 questions. By comparing 30 state-of-the-art LMs, we surprisingly find that vision-language models (VLMs) do not outperform text-only models in either task. Moreover, the models perform significantly worse in the visual dimension compared to other sensory dimensions. Further analysis reveals that the vector representations are easily influenced by word form and frequency, and the models struggle to answer questions involving spatial perception and reasoning. Our findings underscore the need for more effective integration of embodied knowledge in LMs to enhance their understanding of the physical world[1].

1 Introduction

Embodied knowledge is acquired through experience and contextualized in relation to the body (Castree et al., 2013). It is inherently sensory, encompassing sights, sounds, smells, touch, and taste (Ellingson, 2008). Humans rely on both sensory experiences and language experience to represent knowledge (Vinaya et al., 2024; Jones and Bergen, 2024; Jones et al., 2024; Andrews et al., 2014, 2009; Bi, 2021; Günther et al., 2019; Davis and Yee, 2021; Kim et al., 2019). Similarly, embodied knowledge is crucial for AI to bridge the gap between the digital and physical world (Lungarella et al., 2007; Liu et al., 2024d).

Prompt & Predicted Words
Hummingbirds are the [MASK] birds in the world. largest (0.47), smallest (0.13), fastest (0.11), only (0.04)
The trophy doesn't fit into the brown suitcase because the trophy is too [MASK]. small (0.25), big (0.23), heavy (0.17), large (0.17)
The trophy doesn't fit into the brown suitcase because the suitcase is too [MASK]. big (0.28), small (0.22), large (0.17), heavy (0.16)

Table 1: Word prediction examples by `BERT-base`.

However, LMs have traditionally relied on training on massive textual data to understand the world, following the Distributional Hypothesis (Harris, 1954). Although they can memorize knowledge from the data and learn to utilize statistical patterns to demonstrate language understanding and generation abilities, they often make mistakes on questions related to the real world due to a lack of grounding (Bender and Koller, 2020a; Merrill et al., 2021). As shown in Table 1, the text-only model BERT incorrectly identifies hummingbirds as the largest birds in the world and is confused about the relative sizes of objects, indicating that the model struggles to differentiate between antonyms with distinct sensory contrasts, such as *small-big*. Therefore, increasing work has advocated for grounded language learning through the integration of perceptual information and interaction with the physical and social world (Bisk et al., 2020; Bender and Koller, 2020b; Ma et al., 2023; Shi et al., 2025).

In recent years, large language models (LLMs) have demonstrated significantly enhanced intelligence (Du et al., 2022). Notably, multimodal LLMs that associate natural language with visual information exhibit robust visual question-answering capabilities (Yin et al., 2023). This observation naturally raises an intriguing question: **does visual**

[*]Corresponding author.
[1]Our dataset and code are available at `https://github.com/ererdewubudesi/embodied_knowledge_2025`.

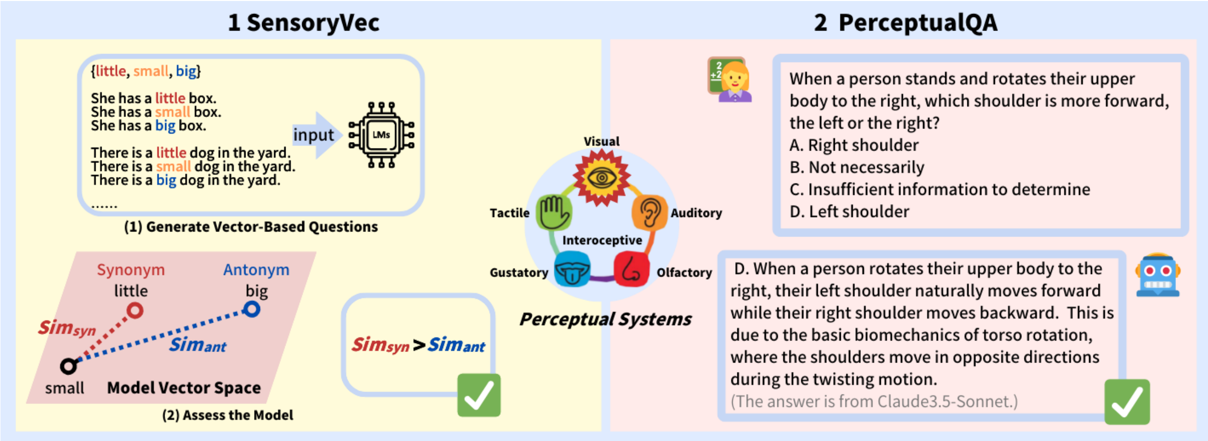

Figure 1: The two tasks in our embodied knowledge understanding benchmark.

grounding improve the models' understanding of embodied knowledge, thereby enabling them to better perceive and comprehend the real world?

To address this question, we propose an embodied knowledge understanding benchmark consisting of two tasks: `SensoryVec` and `PerceptualQA`, which evaluate the model's capability to represent sensory information in vector form and answer perception-related questions, respectively. As shown in Figure 1, both tasks are systematically designed based on the perceptual system framework from psychology (Gerrig et al., 2015), encompassing experiential knowledge from visual, auditory, tactile, gustatory, and olfactory modalities that complement language-derived representations (Bi, 2021). In `SensoryVec`, we concentrate on the five aforementioned external sensory modalities and interoception, constructing a sensory adjective lexicon with 349 word triples and 1,047 sentences reflecting natural contexts. By comparing the cosine similarities between sensory words and their synonyms versus antonyms, we assess the model's vector representations. For `PerceptualQA`, we construct a QA dataset comprising 9 tasks and 1,400 questions across five external sensory modalities. Considering the complexity of the visual modality (Gerrig et al., 2015), we design 5 visual subtasks: color attributes, colors in nature, geometry and transformations, symbols, and body.

We evaluate a wide range of state-of-the-art LMs, specifically selecting VLMs and text-only LMs with multiple comparable pairs, such as Qwen-VL and Qwen, LLaVA-Mistral and Mistral, where the former is directly built upon the latter. We have three main findings: (1) Current models exhibit suboptimal performance in understanding embodied knowledge, with the best-performing models achieving only around 70% accuracy on both tasks, far below human performance. (2) VLMs initialized with text model parameters demonstrate no clear advantage over their language model counterparts on either task. (3) Nearly all models perform significantly worse on visual tasks compared to other modalities, particularly on questions involving spatial understanding and reasoning in categories such as symbols, geometry, and body. These results suggest that multimodal learning has not yet fully capitalized on its potential to capture and leverage embodied knowledge. As embodied AI research actively integrates LLMs and vision large language models (VLLMs) into decision-making (Liu et al., 2024d), the models' limitations in understanding and reasoning about embodied knowledge may become a significant obstacle or concern. In summary, this paper presents a systematic, comprehensive benchmark for evaluating models' embodied knowledge across different modalities, which is valuable for model analysis and diagnosis. Moreover, the resource construction approach outlined in this paper can serve as a reference for more comprehensive model evaluation.

2 Related Work

2.1 Comparing VLMs and text-only LMs

The emergence of multimodal LMs has sparked interest in the differences between VLMs and text-only LMs. A series of studies have compared these two types of models based on vector representations, involving tasks such as word similarity, semantic probing, and measuring correlations between vectors and fMRI data.

Pezzelle et al. (2021) investigated the vector representation differences between multimodal models (e.g. ViLBERT and VisualBERT) and the text-only model BERT on existing word similarity benchmarks. They discovered that multimodal models surpass BERT on concrete words. Yun et al. (2021) compared the embedding differences between VLMs (VideoBERT and VisualBERT) and their text-only variants using a series of probing tasks covering physical commonsense QA, coreference resolution, semantic role labeling, and adjective-noun composition. They found that the multimodal models fail to significantly outperform the text-only variants. Bavaresco et al. (2024) compared word embedding alignment to brain activity using the fMRI dataset by Pereira et al. (2018). Representational similarity analysis revealed substantial differences among VLMs, with some exhibiting higher brain response correlations than unimodal models.

Tikhonov et al. (2023) further explored the factors influencing differences in word vector similarities. They calculated word similarity data for 13,000 word pairs based on multimodal LMs (CLIP, OpenCLIP, Multilingual CLIP) and text-only models (FastText, mBERT, XLM-RoBerta). Using 46 semantic features, they conducted regression analysis to predict vector similarity differences. The results showed that concreteness and taxonomic features from WordNet were significant predictors. However, a considerable portion of the embedding space differences remained unexplained.

In summary, existing research has investigated the differences between multimodal LMs and text-only LMs. However, these studies focused only on vector representations and relied on existing evaluation datasets, which prevented them from comprehensively and specifically diagnosing differences between the two types of models, leading to somewhat contradictory conclusions (Pezzelle et al., 2021; Yun et al., 2021; Bavaresco et al., 2024). Moreover, these works mainly focused on early multimodal and text-only LMs and have not yet considered the perceptual capabilities of LLMs.

2.2 Multimodal reasoning benchmarks

Numerous visual reasoning benchmarks, such as CLEVR (Johnson et al., 2016), MMMU (Yue et al., 2024), and MMBench (Liu et al., 2023c), have been proposed to evaluate VLMs. These benchmarks typically emphasize reasoning over domain knowledge, object attributes and relationships, as well as scene understanding and prediction. Structured as visual question answering tasks, their reliance on visual input at inference time makes it difficult to directly compare the performance of text-only and vision-language models.

2.3 Embodied AI

Embodied AI is essential for achieving Artificial General Intelligence (AGI) and connecting the digital and physical worlds (Liu et al., 2024d). For tasks such as 3D Visual Grounding and Embodied Question Answering, traditional methods often struggle with complex queries and require extensive labeled data. Recent studies (Yang et al., 2024c; Yuan et al., 2024; Majumdar et al., 2024; Patel et al., 2024) show that (multimodal) LLMs can effectively handle diverse queries while reducing data dependencies, often achieving strong performance even without task-specific fine-tuning.

Meanwhile, embodied data has been leveraged to enhance LLM capabilities. Recent studies (Yang et al., 2024b; Fu et al., 2024; Yu et al., 2024) have integrated tactile data with visual and linguistic modalities through contrastive learning and fine-tuning, enabling LLMs to better understand cross-modal embodied knowledge.

While multimodal LLMs are expected to serve as the brain of embodied agents (Liu et al., 2024d), bridging the gap to human-level performance remains challenging. Real-world applications such as autonomous driving and human-computer interaction require sophisticated object understanding, spatial reasoning, and geometric inference capabilities. Continuous assessment and improvement in these areas is essential for robust and generalizable embodied AI systems.

3 Method

This paper proposes an embodied knowledge understanding benchmark consisting of two tasks: `SensoryVec` and `PerceptualQA`, which evaluate the model's capability to represent sensory information in vector form and answer perception-related questions, respectively.

3.1 SensoryVec Task

Given LMs face difficulty in distinguishing antonyms with sensory contrasts (see examples in Table 1), we evaluated the models' vector representations specifically for sensory adjectives. By comparing the similarities between each word and its synonym versus its antonym, we examined whether models assign higher similarity to the synonymous term, as illustrated in Figure 1. A detailed motiva-

Task	Question Perspectives
Visual-Color Attributes	Hue, brightness, saturation
Visual-Colors in Nature	Hue, saturation, warm and cool tones
Visual-Geometry and Transformations	Quantity, shape, direction, size relation, position relation
Visual-Symbols	Quantity, shape, direction, size relation, position relation
Visual-Body	Shape, direction, distance, speed, position relation
Auditory	Volume, pitch
Tactile	Smoothness, elasticity, hardness, weight, thickness, stickiness, temperature
Gustatory	Sourness, sweetness, bitterness, saltiness, spiciness
Olfactory	Fragrance, stench, specific odors

Table 2: Overview of subtasks in `PerceptualQA`.

Task	Question
Visual-Color Attributes	Compared to red, what is the hue of burgundy? A. More orangish **B. More purplish** C. More bluish D. More greenish
Visual-Colors in Nature	Which of these foods has the highest color saturation? A. Jellyfish B. Raw eel C. Raw catfish **D. Cooked shrimp**
Visual-Geometry and Transformations	When a rectangle is folded so that its top and bottom edges coincide, what shape is formed? **A. Rectangle** B. Isosceles triangle C. Rhombus D. Square
Visual-Symbols	When the number [7] is rotated 45 degrees counterclockwise, which direction does its opening face? A. Left B. Up **C. Down** D. Right
Visual-Body	When standing with hands behind the back, are the elbows higher or lower than the hips? **A. Higher** B. Lower C. At the same height D. It varies
Auditory	Which sound is usually louder: the sound of frying fish or the sound of steaming fish? A. Steaming fish **B. Frying fish** C. Almost the same D. It depends
Tactile	Which fruit has the roughest skin? A. Persimmon B. Banana C. Pear **D. Pineapple**
Gustatory	Which food is least likely to have a bitter taste? A. Coffee B. Bitter gourd **C. Corn** D. Green tea
Olfactory	Which of the following things has the most distinct fragrance? A. Eggplant **B. Rose tea** C. Carrot D. Potato

Table 3: Sample questions from `PerceptualQA` (correct answers highlighted in **bold**).

tion for designing `SensoryVec` task can be seen in Appendix A.1.

First, we collected candidate sensory adjectives from three high-quality datasets (Lievers, 2015; Lynott and Connell, 2009; Lynott et al., 2019). Referring to the perceptual system framework from psychology (Gerrig et al., 2015), we selected sensory words related to visual, auditory, tactile, gustatory, olfactory, and interoceptive senses, and annotate their attributes. To ensure representativeness, we retained words with sensory ratings above 4 from the datasets constructed by Lynott and Connell (2009) and Lynott et al. (2019). Subsequently, we reviewed and cleaned the data (e.g., by removing duplicate words), as detailed in Appendix B.1.1.

Next, synonyms and antonyms are identified with reference to WordNet (Fellbaum, 1998) and a thesaurus (Random House, 2001). For words without strict antonyms, such as color terms, semantically distant counterparts are used. In this way, we constructed sensory triples like *small-little-big*. Furthermore, we curated three natural contextual sentences for each triple, generated by LLMs and then refined through human review (details in Appendix B.1.2). Finally, we obtained 349 groups of sensory adjectives and 1047 sentences, systematically covering a wide range of attributes across different sensory dimensions:

- **Visual**: color, shape, size, depth, distance, orientation, displacement, speed, etc.
- **Auditory**: pitch, loudness, rhythm, etc.
- **Tactile**: texture, cold, warm, etc.
- **Gustatory**: sour, sweet, bitter, spicy, salty, specific food flavors, etc.
- **Olfactory**: fragrant, odorous, pungent, etc.
- **Interoceptive**: hunger, satiety, etc.

In the evaluation, for contextualized models, we use sentences as input and obtain average token vectors from the last hidden layer, then compute the mean vector across the three sentences. For other models (such as Word2vec and GloVe), we directly utilize standalone word vectors.

3.2 PerceptualQA Task

Considering that question-answering is the most natural interaction mode for LLMs, we introduce the `PerceptualQA` task to comprehensively evaluate a model's embodied knowledge. Consistent with the SensoryVec task, we design questions referring to perceptual systems from psychological theory (Gerrig et al., 2015), covering visual, auditory, tactile, gustatory, and olfactory modalities. Given the complexity of vision and the emphasis on visual training in most existing multimodal models, we design five subtasks for the visual modality: color attributes, colors in nature, geometry and transformations, symbols, and body.

As shown in Table 2, questions for each modality are designed from multiple perspectives. They are intended to be easily answerable by humans through embodied imagination and reasoning.[2] For geometry, symbol, and body-related questions, we consider not only the original shapes of objects but also their transformed states (e.g., after rotation). Table 3 presents examples for each task. The full question frameworks are provided in Table 9 (see Appendix B.2.1).

We used an LLM (Claude-3.5-Sonnet) as a brainstorming tool to generate candidate questions for human annotators. Each question has four options with one correct answer. Most PerceptualQA items were manually filtered; the remaining items were further revised by annotators according to rigorous design criteria. Ultimately, the PerceptualQA task consists of 1,400 multiple-choice questions, with 200 questions per visual subtask and 100 questions per other perceptual understanding task. See details of dataset construction in Appendix B.2.2.

4 Experiments

4.1 Models and Settings

We systematically evaluated an extensive range of state-of-the-art VLMs and text-only models, spanning diverse architectures, model sizes, and availability (open-source vs. closed-source). Models assessed include CLIP, Word2Vec, GloVe, BERT, Mistral, Vicuna, LLaMA, Gemma, Qwen, LLaVA, GPT, Gemini, and Claude series. See Appendix C for details of models and parameters.

The selected models include 6 groups of comparable model pairs, where a vision-language model is built upon a text-only language model, enabling direct exploration of the impact of visual grounding. These include VisualBERT & BERT, LLaVA-1.6-Vicuna-7B & Vicuna-7B, LLaVA-1.6-Mistral-7B & Mistral-7B, Qwen-VL-Chat & Qwen-7B, Qwen2-VL-7B-Instruct & Qwen2-7B, and Qwen2-VL-72B-Instruct & Qwen2-72B (Li et al., 2019a; Liu et al., 2024c,a; Bai et al., 2023b; Wang et al., 2024a).

For SensoryVec, we evaluated the accuracy based on cosine similarity judgments for each triple.[3] For PerceptualQA, we reported average accuracy across two trials for reliability.[4]

Model	All Sensory	Visual	Non-Visual
Word2Vec	67.64	65.00	71.33
GloVe	62.50	57.81	69.12
BERT	**72.21**	70.44	**74.66**
<u>VisualBERT</u>	64.18	65.52	62.33
GPT-2	50.43	47.78	54.11
Qwen-7B	61.32	54.68	70.55
<u>Qwen-VL</u>	58.74	53.69	65.75
<u>Qwen-VL-Chat</u>	61.60	55.17	70.55
Qwen2-7B	63.32	58.62	69.86
Qwen2-7B-Instruct	66.19	61.58	72.60
<u>Qwen2-VL-7B-Instruct</u>	63.04	58.62	69.18
Mistral-7B	67.05	63.55	71.92
<u>LLaVA1.6-Mistral-7B</u>	66.76	65.02	69.18
Vicuna-7B	57.59	58.62	56.16
<u>LLaVA1.6-Vicuna-7B</u>	58.45	59.61	56.85
<u>CLIP</u>	71.06	**75.37**	65.07

Table 4: Results on SensoryVec for static embeddings, bidirectional, generative, and contrastive learning models (top to bottom). VLMs are <u>underlined</u>. Comparable models within each group share the same background color, with the LMs above the VLMs. **Bolded** values denote the highest accuracy per sensory modality.

4.2 Results

4.2.1 SensoryVec Task

Table 4 demonstrates the results on SensoryVec. All models perform suboptimally on this task, with accuracy ranging from 50% to 70%. Moreover, VLMs perform comparably or worse than their corresponding text-only models, consistent with Yun et al. (2021)'s findings. This reveals that perceiving sensory contrasts remains an important challenge for both text-only and multimodal models. Furthermore, across different modalities, we find that for most models, visual word representations are notably inferior to those of non-visual sensory words. Even VLLMs like Qwen2-VL-7B-Instruct and LLaVA1.6-Mistral-7B struggle to bridge this gap. Moreover, VLMs show no significant improvement over their text-only counterparts.

Surprisingly, the text-only language model BERT achieves the best overall accuracy at 72.21%.

[2] A detailed motivation for designing PerceptualQA task can be seen in Appendix A.2.

[3] Triples containing words that are not in Word2Vec and GloVe vocabulary are excluded from total count.

[4] Evaluation prompts are in Appendix C.2, with correct answer distributions shown in Table 10.

This indicates that even without explicit visual input, models can capture a significant amount of sensory-related information from large-scale textual data, although substantial gaps remain. These results echo findings from psychological studies showing that language can convey certain aspects of perceptual knowledge. For example, previous research has demonstrated that congenitally blind individuals can acquire some knowledge about visual properties such as shape, texture, or size through linguistic descriptions and conceptual inference, despite not having direct visual experience (Bi, 2021; Kim et al., 2019). It is important to note that, unlike language models, blind individuals have access to other sensory modalities (e.g., touch and hearing), providing them with richer embodied experiences. Nevertheless, the absence of sensory experience still hinders understanding of some vision-dependent attributes, such as color. In Kim et al. (2019), blind participants exhibited significant difficulty grouping and sorting animal colors compared to sighted controls.

Our detailed analysis of SensoryVec results suggests that the models' suboptimal performance is primarily due to an over-reliance on distributional semantics during training, as well as insufficient representation of form-similar and low-frequency terms. Section 5 and Table 6 illustrate these issues with detailed examples.

4.2.2 PerceptualQA Task

Table 5 presents the results for the PerceptualQA task. All models continue to perform poorly (see Appendix D.4 for complete results). The best model, Claude3.5-Sonnet, achieves an accuracy of 69.04%, significantly lower than human performance (86.00%).[5]

Across four comparable model groups, VLLMs (LLaVA1.6-Vicuna-7B, 41.64%; LLaVA1.6-Mistral-7B, 45.64%; Qwen2-VL-7B-Instruct, 51.00%; Qwen2-VL-72B-Instruct, 63.89%) show no clear advantage over their LLM counterparts (Vicuna-7B, 38.25%; Mistral-7B, 42.96%; Qwen2-7B, 49.36%; Qwen2-72B-Instruct, 62.32%), with an average accuracy gain of only 2.32%. This indicates that incorporating visual information does not substantially enhance model performance on these tasks. Furthermore, these findings exhibit cross-language generalizability (see Appendix D.2).

Model	All Sensory	Visual	Non-Visual
Human Baseline	86.00	85.20	88.00
Llama3.2-3B-Instruct	47.71	38.25	71.38
Vicuna-7B	38.25	32.60	52.38
LLaVA1.6-Vicuna-7B	41.64	35.25	57.63
Mistral-7B	42.96	31.40	71.88
LLaVA1.6-Mistral-7B	45.64	35.90	70.00
Qwen2-7B	49.36	39.85	73.13
Qwen2-7B-Instruct	47.82	35.80	77.88
Qwen2-VL-7B-Instruct	51.00	41.70	74.25
Llama3.1-8B	48.54	39.05	72.25
Gemma2-9B	51.89	40.65	80.00
Gemma2-27B	55.39	44.90	81.63
Llama3.1-70B	59.71	49.85	84.38
Qwen2-72B-Instruct	62.32	52.75	86.25
Qwen2-VL-72B-Instruct	63.89	54.45	87.50
Llama3.1-405B	63.46	54.55	85.75
GPT-3.5	50.46	39.50	77.88
Qwen-Max	68.71	**61.05**	87.88
GPT-4o-Mini	57.18	46.35	84.25
Gemini1.5-Flash-8B	54.39	44.55	79.00
Gemini1.5-Flash	56.07	45.20	83.25
GPT-4o	68.46	59.45	91.00
Gemini1.5-Pro	65.21	56.55	86.88
Claude3.5-Sonnet	**69.04**	60.00	**91.63**
Qwen-VL-Max	64.68	55.30	88.13

Table 5: Results on PerceptualQA for open-source models, closed-source LLMs, and closed-source VLLMs (top to bottom). VLLMs are underlined. Comparable models within each group share the same background color, with LLMs above VLLMs. **Bolded** values denote the highest accuracy per sensory modality.

Similar to SensoryVec, Table 5 reveals that all models perform substantially worse on visual tasks compared to other tasks. However, for humans, the difficulty gap between visual (85.20%) and non-visual (88.00%) tasks is minimal. This comparison suggests that models, unlike humans, face greater difficulty in certain sensory dimensions. The best model performance on visual tasks (61.05%) falls significantly short of human performance (85.20%). Additionally, we analyzed the influence of potential confounding variables (see Appendix D.3). In Section 5, we further discuss the performances of the different subtasks.

5 Discussion

Experimental results on both datasets demonstrate that existing models perform poorly in representing and applying embodied knowledge, with no clear or systematic improvements observed in VLMs over

[5]To establish a human baseline for model evaluation, we recruited seven native-speaking graduate students from the university community. Each question received two independent responses, yielding a Cohen's Kappa of 0.69. Participants completed the tasks individually using Wenjuanxing, an online survey platform with functionality comparable to Amazon Mechanical Turk. All participants reported good health and no relevant impairments. Further details are provided in Appendix D.6.

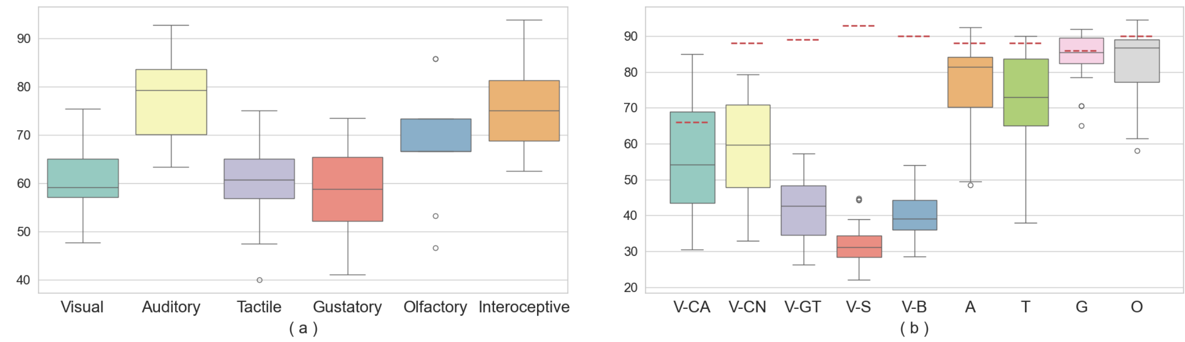

Figure 2: Box plots of accuracy on SensoryVec (a) and PerceptualQA (b) subtasks. PerceptualQA subtasks include Visual-Color Attributes(V-CA), Visual-Colors in Nature(V-CN), Visual-Geometry and Transformations(V-GT), Visual-Symbols(V-S), Visual-Body(V-B), Auditory(A), Tactile(T), Gustatory(G), and Olfactory(O). Red dashed lines indicate human performance.

Category	**Triple (word, synonym, antonym)**	**Context**
Tactile	(dry, waterless, wet)	The towel was completely (dry, waterless, wet) now.
Tactile	(sharp, honed, blunt)	The knife is (sharp, honed, blunt).
Gustatory	(sugarless, unsweetened, sugary)	She prefers (sugarless, unsweetened, sugary) drinks.
Gustatory	(salty, brackish, flavorless)	The soup tasted (salty, brackish, flavorless).
Visual-Shape	(concave, hollow, convex)	The sculpture's surface was (concave, hollow, convex).
Visual-Shape	(loose, baggy, tight)	She prefers her jeans to be (loose, baggy, tight).
Visual-Shape	(unwrinkled, smooth, wrinkled)	The surface feels (unwrinkled, smooth, wrinkled).
Visual-Shape	(open, unclosed, closed)	The gate was left (open, unclosed, closed).
Visual-Color	(black, sable, white)	The walls of the room were painted (black, sable, white).
Visual-Color	(brown, chestnut, blue)	Her hair is (brown, chestnut, blue).
Visual-Color	(green, verdant, red)	She wore a pair of (green, verdant, red) socks.
Visual-Color	(undimmed, bright, dim)	The stars in the night sky were (undimmed, bright, dim).

Table 6: Examples of models' prediction failures on SensoryVec, with $Sim_{antonym} > Sim_{synonym}$.

text-only LMs. We further analyzed the models' performance across different question categories to investigate the specific aspects where models struggle most.

As shown in Figure 2a, on SensoryVec, models perform poorly on tactile and gustatory categories in the non-visual domain. In the visual domain, where overall performance is relatively weaker, models particularly struggle with color and shape. Table 6 shows some failure examples.[6] It can be seen that sensory words judged more similar to their antonyms often exhibit greater similarity in form (e.g., *sugarless-sugary*, *unwrinkled-wrinkled*, *undimmed-dim*) or frequency (e.g., *dry-wet*, *sharp-blunt*, *open-closed*, *green-red*), highlighting the limitations of distributed semantic representations in capturing semantic similarity.

On PerceptualQA (Figure 2b), the symbol task poses the greatest challenge, followed by the body task, and the geometry and transformations task, while color-related tasks show comparatively better performance. We further examined the questions in the three most challenging visual understanding subtasks and list several difficult examples that no model could answer correctly in Table 7.[7] We observed that models' errors are not tied to specific shapes (e.g., triangles), symbol types (e.g., numbers), body parts (e.g., arms), transformation types (e.g., rotation), or question targets (e.g., positional relationships). This observation suggests a systemic deficiency in spatial reasoning, rather than a lack of specific visual or conceptual knowledge.

Additionally, a potential explanation for why some visual tasks are more difficult is that reasoning about an object's transformed states often requires more complex inference than reasoning over comparative attributes of multiple entities. The design intention behind PerceptualQA is that both

[6]CLIP answers half correctly; GloVe, Word2Vec, and Qwen-7B each solve one; others fail all.

[7]Results reflect only the models' first attempt out of two. Additional examples are provided in Appendix D.5.

Question & Options & Answers and Rationales from Claude3.5-Sonnet
[V-GT] If a rectangle has long sides on the top and bottom, and short sides on the left and right, and it is flipped around its left short side, where is the original bottom edge now positioned? **A. Bottom** B. Top C. Right zhe geD. Left B. When a rectangle is flipped around its left short side (like a door hinge), the bottom edge will rotate 180 degrees upward, ending up at the top position. This is similar to how a door swings open, with the hinge being the left side in this case.
[V-S] If the number 3 is horizontally cut in half, which direction does the opening of the bottom part face? A. Up **B. Left** C. Down D. Right A. When a number 3 is cut horizontally in half, the bottom part resembles a curved line with the opening facing upward, similar to the bottom half of a sideways heart shape.
[V-B] During the lower limb whipping motion, what is the trend of distance change between the knee and the hip? A. Decreases **B. Remains essentially unchanged** C. Increases D. Decreases first, then increases A. During a lower limb whipping motion, the knee moves closer to the hip as the leg flexes through the whipping action, resulting in a decrease in the distance between these two joints.

Table 7: Examples of models' `PerceptualQA` prediction failures and answers from Claude3.5-Sonnet. Task types are shown in brackets. Correct answers in **bold**.

	All Sensory	Visual	V-CA	V-CN	V-GT	V-S	V-B	Non-Visual	A	T	G	O
Human Baseline	86.00	85.20	66.00	88.00	89.00	93.00	90.00	88.00	88.00	88.00	86.00	90.00
GPT-4o-Mini	58.21	50.96	57.14	71.43	42.86	37.21	47.83	79.17	75.00	80.00	92.86	72.22
GPT-4o-Mini-FT	79.29	75.96	100.00	83.33	71.43	65.12	65.22	88.89	85.00	95.00	92.86	83.33

Table 8: Accuracy of GPT-4o-Mini on `PerceptualQA` subtasks before and after fine-tuning.

visual and non-visual questions require reasoning over a set of properties, as opposed to relying on shallow lexical associations. Specifically, when a question involves only a single target concept, we introduce transformations to elicit reasoning over derived attributes. For example, in the geometry and transformations task, we ask what shape is formed when a rectangle is folded along the line connecting the midpoints of its shorter sides. This requires understanding the geometric transformation of folding, as well as the shape of the object before and after the transformation. When the target concept involves multiple entities, the attribute set arises more directly. For instance, in the gustatory task, we ask how a set of foods should be grouped based on their taste. This involves identifying and comparing gustatory attributes and performing categorization.

To investigate whether fine-tuning could directly improve models' understanding of embodied knowledge, we utilized the `PerceptualQA` dataset to fine-tune GPT-4o-Mini and Qwen2 series models. The results of GPT-4o-Mini are presented in Table 8. After supervised fine-tuning (SFT), the performance of GPT-4o-Mini improved from 58.21% to 79.29%, but still lagged behind human performance (86.00%). This gap is primarily observed in the visual dimension, particularly in the three most challenging subtasks (V-GT, V-S, V-B) related to spatial perception and reasoning, where the fine-tuned model achieves an average accuracy of 67.26% compared to 90.67% of human performance. In addition, no significant performance improvements were observed in the Qwen2 series models. These results indicate that understanding embodied knowledge, especially spatial perception knowledge, remains a fundamental challenge for current models. Detailed experimental settings and results are presented in Appendix D.1.

VLLMs' limited improvement over LLMs in embodied knowledge understanding stems from their training mechanisms. Our questions are intentionally designed to require sensory experience rather than factual knowledge, making them challenging for models because answering them relies on embodied imagination and reasoning rarely conveyed in natural language. VLLMs are typically trained on static image-text pairs that capture surface-level visual features. Due to the high cost of grounding annotations, it remains impractical to comprehensively label the knowledge about the physical world (Ma et al., 2023). Consequently, even models like Qwen2-VL (Wang et al., 2024a), which adopt multi-stage training to improve image-text alignment, do not necessarily achieve better embodied understanding.[8] While multimodal LLMs are increasingly central to embodied AI research, their current limitations in perceiving and reason-

[8]The three stages include: (1) training the vision encoder on image-text pairs, (2) leveraging larger-scale fine-grained datasets to refine visual-textual alignment, and (3) instruction tuning with multimodal and text-based dialogue data.

ing about embodied knowledge pose significant challenges to their effectiveness as the cognitive core of embodied agents. Our experimental results and analysis suggest that existing training tasks and datasets may be insufficient for fostering meaningful advances in this area.

6 Conclusion

In this paper, we investigate how visual grounding influences models' understanding of embodied knowledge. To this end, we introduce a benchmark grounded in the psychological framework of the perceptual system, covering external senses (visual, auditory, tactile, gustatory, olfactory) as well as interoception. Our benchmark comprises two tasks—`SensoryVec` and `PerceptualQA`—with over 1,700 questions designed to systematically evaluate the models.

Our findings reveal that existing models perform suboptimally in embodied knowledge understanding, with vision-language models showing no significant advantage over text-only models. This suggests that current visual grounding approaches do not effectively enhance embodied knowledge comprehension. Further analysis shows that models' vector representations are susceptible to surface form and frequency bias, and they struggle with spatial perception and reasoning tasks. These insights underscore the importance of continued advancement in embodied knowledge understanding for AGI development.

Future research should explore developing diverse forms of training data, novel training tasks and architectures that better integrate multimodal perceptual information to advance embodied knowledge understanding. Specifically, in addition to text and image data, several categories of multimodal information may need to be incorporated, including dynamic sensory data such as video sequences, perception-related brain neural data, and feedback signals from non-human sensors such as haptic sensors and motion capture systems. These diverse data modalities may be essential to fully support the acquisition of comprehensive embodied knowledge. Furthermore, future models may require joint training with embodied agents on foundational tasks that involve interaction with simulated or real-world environments, which would enable them to actively explore, ground concepts through experiential learning, and acquire causal understanding beyond statistical correlations.

Limitations

The primary objective of our research was to diagnose the current models' understanding of embodied knowledge, rather than proposing methods to directly enhance model performance on this task. Although fine-tuning GPT-4o-Mini on `PerceptualQA` dataset yielded improved results, we have not yet fully explored the potential of purely textual data. Future work will aim to expand the size and diversity of the dataset, investigating its capacity to further boost model performance.

Ethics Statement

No sensitive or private information was used in this research. Benchmark construction underwent rigorous quality control, with careful measures taken to avoid data bias. To promote transparency and benefit the broader community, the benchmark has been made publicly available for future comparative evaluations and further research. All results are reported transparently and without manipulation.

Acknowledgment

The authors would like to thank Dr. Wang Yin and his lab members for their valuable discussions and suggestions. This research was supported by the Tencent Basic Platform Technology Rhino-Bird Focused Research Program.

References

Mark Andrews, S. Frank, and Gabriella Vigliocco. 2014. Reconciling embodied and distributional accounts of meaning in language. *Topics in cognitive science*, 6 3:359–70.

Mark Andrews, Gabriella Vigliocco, and David P. Vinson. 2009. Integrating experiential and distributional data to learn semantic representations. *Psychological review*, 116 3:463–98.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Anna Bavaresco, Marianne de Heer Kloots, Sandro Pezzelle, and Raquel Fernández. 2024. Modelling multimodal integration in human concept processing with vision-and-language models. *Preprint*, arXiv:2407.17914.

Emily M. Bender and Alexander Koller. 2020a. 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*, pages 5185–5198, Online. Association for Computational Linguistics.

Emily M Bender and Alexander Koller. 2020b. 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*, pages 5185–5198.

Yanchao Bi. 2021. Dual coding of knowledge in the human brain. *Trends in Cognitive Sciences*, 25(10):883–895.

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. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Noel Castree, Alisdair Rogers, and Rob Kitchin. 2013. *A dictionary of human geography*. Oxford University Press.

Charles P Davis and Eiling Yee. 2021. Building semantic memory from embodied and distributional language experience. *Wiley Interdisciplinary Reviews: Cognitive Science*, 12(5):e1555.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. 2022. A survey of vision-language pre-trained models. *Preprint*, arXiv:2202.10936.

Laura L Ellingson. 2008. Embodied knowledge. *The SAGE Encyclopedia of Qualitative Research Methods*, pages 244–245.

Christiane Fellbaum. 1998. Wordnet: An electronic lexical database. *MIT Press google schola*, 2:678–686.

Letian Fu, Gaurav Datta, Huang Huang, William Chung-Ho Panitch, Jaimyn Drake, Joseph Ortiz, Mustafa Mukadam, Mike Lambeta, Roberto Calandra, and Ken Goldberg. 2024. A touch, vision, and language dataset for multimodal alignment. *arXiv preprint arXiv:2402.13232*.

Richard J Gerrig, Philip G Zimbardo, Andrew J Campbell, Steven R Cumming, and Fiona J Wilkes. 2015. *Psychology and life*. Pearson Higher Education AU.

Fritz Günther, Luca Rinaldi, and Marco Marelli. 2019. Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, 14(6):1006–1033.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2016. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997.

Cameron R Jones and Benjamin Bergen. 2024. Does word knowledge account for the effect of world knowledge on pronoun interpretation? *Language and Cognition*, 16(4):1182–1213.

Cameron R. Jones, Benjamin Bergen, and Sean Trott. 2024. Do multimodal large language models and humans ground language similarly? *Computational Linguistics*, 50(3):1415–1440.

Judy S Kim, Giulia V Elli, and Marina Bedny. 2019. Knowledge of animal appearance among sighted and blind adults. *Proceedings of the National Academy of Sciences*, 116(23):11213–11222.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior research methods*, 44:978–990.

Jianquan Li, Renfen Hu, Xiaokang Liu, Prayag Tiwari, Hari Mohan Pandey, Wei Chen, Benyou Wang, Yaohong Jin, and Kaicheng Yang. 2020. A distant supervision method based on paradigmatic relations for learning word embeddings. *Neural Computing and Applications*, 32:7759–7768.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019a. Visualbert: A simple and performant baseline for vision and language. *ArXiv*, abs/1908.03557.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019b. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Francesca Strik Lievers. 2015. Synaesthesia: A corpus-based study of cross-modal directionality. *Functions of language*, 22(1):69–95.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024c. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. 2024d. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *arXiv preprint arXiv:2407.06886*.

Yuanzhan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023c. Mmbench: Is your multi-modal model an all-around player? *ArXiv*, abs/2307.06281.

Max Lungarella, Fumiya Iida, Josh Bongard, and Rolf Pfeifer. 2007. *50 Years of Artificial Intelligence: Essays Dedicated to the 50th Anniversary of Artificial Intelligence*, volume 4850. Springer.

Dermot Lynott and Louise Connell. 2009. Modality exclusivity norms for 423 object properties. *Behavior research methods*, 41(2):558–564.

Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. 2019. The lancaster sensorimotor norms: multidimensional measures of perceptual and action strength for 40,000 english words. *Behavior Research Methods*, 52:1271 – 1291.

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.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. 2024. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16488–16498.

William Merrill, Yoav Goldberg, Roy Schwartz, and Noah A Smith. 2021. Provable limitations of acquiring meaning from ungrounded form: What will future language models understand? *Transactions of the Association for Computational Linguistics*, 9:1047–1060.

Tomas Mikolov. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Bhrij Patel, Vishnu Sashank Dorbala, and Amrit Singh Bedi. 2024. Embodied question answering via multi-llm systems. *arXiv preprint arXiv:2406.10918*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, 9(1):963.

Sandro Pezzelle, Ece Takmaz, and Raquel Fernández. 2021. Word representation learning in multimodal pre-trained transformers: An intrinsic evaluation. *Transactions of the Association for Computational Linguistics*, 9:1563–1579.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Random House. 2001. *Random House Roget's Thesaurus*. Ballantine Books.

Igor Samenko, Alexey Tikhonov, and Ivan P. Yamshchikov. 2021. *Intuitive Contrasting Map for Antonym Embeddings*. IOS Press.

Freda Shi, Ziqiao Ma, Jiayuan Mao, Parisa Kordjamshidi, and Joyce Chai. 2025. Learning language through grounding. In *Proceedings of the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pages 38–43, Albuquerque, New Mexico. Association for Computational Linguistics.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Gemma Team. 2024. Gemma.

Alexey Tikhonov, Lisa Bylinina, and Denis Paperno. 2023. Leverage points in modality shifts: Comparing language-only and multimodal word representations. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 11–17, Toronto, Canada. 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, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Harshada Vinaya, Sean Trott, Diane Pecher, Rene Zeelenberg, and Seana Coulson. 2024. Context-dependent and dynamic effects of distributional and sensorimotor distance measures on eeg. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

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. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

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. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Fengyu Yang, Chao Feng, Ziyang Chen, Hyoungseob Park, Daniel Wang, Yiming Dou, Ziyao Zeng, Xien Chen, Rit Gangopadhyay, Andrew Owens, et al. 2024b. Binding touch to everything: Learning unified multimodal tactile representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26340–26353.

Jianing Yang, Xuweiyi Chen, Shengyi Qian, Nikhil Madaan, Madhavan Iyengar, David F Fouhey, and Joyce Chai. 2024c. Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7694–7701. IEEE.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Samson Yu, Kelvin Lin, Anxing Xiao, Jiafei Duan, and Harold Soh. 2024. Octopi: Object property reasoning with large tactile-language models. *arXiv preprint arXiv:2405.02794*.

Zhihao Yuan, Jinke Ren, Chun-Mei Feng, Hengshuang Zhao, Shuguang Cui, and Zhen Li. 2024. Visual programming for zero-shot open-vocabulary 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20623–20633.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567.

Tian Yun, Chen Sun, and Ellie Pavlick. 2021. Does vision-and-language pretraining improve lexical grounding? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4357–4366, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

A Motivation Behind Task Design

A.1 SensoryVec Task

For pretrained language models, word vectors are learned based on the Distributional Semantic Hypothesis: words that occur in similar contexts tend to have similar meanings (Harris, 1954). Since text-only models rely exclusively on contextual co-occurrences, they struggle to distinguish antonyms without incorporating additional mechanisms, such as contrastive mapping (Samenko et al., 2021) or semantic information from thesauri (Li et al., 2020).

Specifically, regarding antonyms, their frequent co-occurrence in similar contexts (e.g., "this is a big ball" versus "this is a small ball") naturally leads to similarity in vector space. Therefore, antonyms represent a special case where terms share numerous semantic properties while differing primarily along a single dimension. These key dimensions often involve sensory attributes like size (big–small) or temperature (cold–warm), which are precisely the types of distinctions that text-only models struggle to differentiate due to their fundamentally non-perceptual architecture.

Visual data potentially introduces sensory information such as color, shape, and size, which may help differentiate concepts that appear in similar linguistic contexts but represent opposite meanings.

By employing vector similarity, our approach provides a targeted evaluation of whether multimodal models can encode sensory knowledge that distinguishes antonyms, a form of knowledge that is largely inaccessible to text-only models. This task is motivated by the need to assess whether visual grounding can overcome the fundamental limitations of text-based representations, thereby offering a meaningful strategy for evaluating perceptual grounding in model representation spaces.

A.2 PerceptualQA Task

When designing questions for `SensoryVec` task, we aimed for the question-answer pairs to rely on human sensory experiences rather than factual background knowledge (e.g., "What color is an apple?"). Each question is intended to be easily answerable by humans through embodied imagination and reasoning, but is rarely expressed in natural language. This introduces significant challenges for both text-only and multimodal models as the answers cannot be directly learned from training corpora through statistical co-occurrence patterns.

B Data Construction Details

B.1 SensoryVec Task

B.1.1 Word Data Cleaning Principles

For data that satisfied both sensory categorization and scoring criteria across multiple datasets, we performed the following filtering steps:

1. Removed duplicate words within the same sensory category.
2. Excluded words lacking documented synonyms or antonyms.
3. Removed words absent from the established word embedding vocabulary.
4. Excluded words for which it was difficult to construct natural contextual sentences that align with the principles outlined in Section B.1.2.

B.1.2 Sentence Data Construction

Following the establishment of sensory triples (e.g., small–little–big), we curated three naturalistic contextual sentences for each triple according to the following procedure:

1. First, we prompted GPT-4o to generate candidate sentences based on three criteria: (1) grammatical correctness, (2) semantic appropriateness, and (3) contextual alignment, which refers to the correspondence between each polysemous word's meaning and its primary sensory modality.

   For sensory triples across diverse modalities, modality-specific prompts were developed. The following exemplifies a prompt employed for candidate generation:

   > Make English sentences so that any of the three adjectives {small, little, big} in the blank are grammatically correct and reasonable sentences. In addition, contextual information makes the meaning of these three words express what is perceived by the human eye. Give me ten sentences with spaces and no other information.

2. Then, we reviewed and modified the generated sentences according to the aforementioned criteria. In instances where a sensory triple lacked sufficient valid sentences, supplementary sentences were manually constructed adhering to identical criteria.

B.2 PerceptualQA Task

B.2.1 Question Frameworks

For the `PerceptualQA` task, we adopt a high-level question design framework that formulates questions from multiple perspectives targeting a specific sensory dimension of the concept being examined. Each subtask follows its own question construction methodology tailored to the characteristics of that sensory domain. Comprehensive question frameworks are presented in Table 9.

B.2.2 QA Data Construction

The construction of the `PerceptualQA` dataset adhered to the following procedures:

1. We first prompted Claude-3.5-Sonnet as a brainstorming aid to generate candidate questions based on the question framework shown in Section B.2.1 to support human annotators. The majority of `PerceptualQA` items are manually crafted to meet rigorous design criteria, as most model-generated questions are discarded due to lacking a correct or unique answer, or being simply infeasible to answer. A small subset of questions originating from the model are manually screened and refined according to the same criteria. The design criteria are as follows:

   (1) Questions and corresponding answers should rely primarily on human sensory experience rather than factual background knowledge.

   (2) Questions should be readily answerable by humans through embodied imagination and reasoning, yet rarely expressed in natural language.

   (3) Each question should be objective, with a singular correct response among four provided options.

   (4) Incorrect options should be plausible, not obviously incorrect, thereby ensuring appropriate task difficulty.

   (5) All questions and answers should be clear, natural, and grammatically well-formed.

   For each task, multiple prompts were developed based on the established question frame-

Task	Target Concepts	Question Perspectives
Visual-Color Attributes	Objectively and subjectively described colors (e.g., red, cherry red)	Hue, brightness, saturation
Visual-Colors in Nature	Edible animals (e.g., chicken), other animals (e.g., giraffe), ornamental plants (e.g., lucky bamboo), vegetables (e.g., carrot), fruits (e.g., banana), edible fungi (e.g., mushroom)	Hue, saturation, warm and cool tones
Visual-Geometry and Transformations	Lines, triangles, quadrilaterals, polygons, circles, composite shapes; may undergo translation, rotation, flipping, compression, stretching, folding, cutting, combination	Quantity, shape, direction, size relation, position relation
Visual-Symbols	Numbers (e.g., 6), letters (e.g., A), Chinese characters; may undergo translation, rotation, flipping, cutting, combination	Quantity, shape, direction, size relation, position relation
Visual-Body	External body parts (e.g., head) in static postures or dynamic movements	Shape, distance, direction, speed, position relation
Auditory	Sounds produced by objects or object interactions (e.g., baby crying, apple slicing)	Volume, pitch
Tactile	Food, objects (e.g., shuttlecock), human body parts (e.g., neck)	Smoothness, elasticity, hardness, weight, thickness, stickiness, temperature
Gustatory	Food	Sourness, sweetness, bitterness, saltiness, spiciness
Olfactory	Food (vegetables, fruits, dishes, seasonings), other items (e.g., lucky bamboo), environments (e.g., hospital)	Fragrance, stench, specific odors

Table 9: Question frameworks including target concepts and question perspectives in each `PerceptualQA` subtask.

work. The following exemplifies a prompt utilized for candidate generation:

> Construct 10 questions about the position relationships of body parts while lying.
> External body parts: head, shoulders, arms, hands, knees, feet, etc.
> Requirements: Answers should be objective choices. Return questions and options in JSON format.
> Example: {"posture dimension": "lying", "question target": "position relationship", "question": "When lying flat, is the knee typically higher or lower than the hip when viewed from the side?", "options": {"A": "Higher", "B": "Lower", "C": "Uncertain", "D": "Insufficient information to determine"}}
> Format: {"posture dimension": "lying", "question target": "position relationship", "question": "", "options": {"A": "", "B": "", "C": "", "D": ""}}

2. Then, we reviewed and modified the generated questions according to the aforementioned criteria. In instances where the task exhibited insufficient question quantity, supplementary questions were manually formulated adhering to identical criteria.

C Models and Experimental Settings Details

C.1 SensoryVec Task

We evaluated 16 models. For text-only models, we evaluated Word2Vec (Mikolov, 2013; Mikolov et al., 2013), GloVe (Pennington et al., 2014), Bert-base-uncased (BERT) (Devlin, 2018), Qwen-7B (Bai et al., 2023a), Qwen2-7B (Yang et al., 2024a), Qwen2-7B-Instruct (Yang et al., 2024a), Mistral-7B-Instruct-v0.2 (Mistral-7B) (Jiang et al., 2023), Vicuna-7b-v1.5 (Vicuna-7B) (Zheng et al., 2023), and GPT-2 (Radford et al., 2019). For

VLMs, we evaluated VisualBERT (Li et al., 2019b), Qwen-VL (Bai et al., 2023b), Qwen-VL-Chat (Bai et al., 2023b), Qwen2-VL-7B-Instruct (Wang et al., 2024b; Bai et al., 2023b), LLaVA-v1.6-Mistral-7B (LLaVA1.6-Mistral-7B) (Liu et al., 2024b, 2023a,b), LLaVA-v1.6-Vicuna-7B (LLaVA1.6-Vicuna-7B) (Liu et al., 2024b, 2023a,b), and CLIP (Radford et al., 2021).

C.2 PerceptualQA Task

We evaluated 24 models. For open-source LLMs, we evaluated Llama3.2-3b-instruct (Llama3.2-3B) (Touvron et al., 2023), Vicuna-7B, Mistral-7B, Qwen2-7B, Qwen2-7B-Instruct, Llama3.1-8b-instruct (Llama3.1-8B) (Touvron et al., 2023), Gemma2-9b-it (Gemma2-9B) (Team, 2024), Gemma2-27b-it (Gemma2-27B) (Team, 2024), Llama3.1-70b-instruct (Llama3.1-70B) (Touvron et al., 2023), Qwen2-72B-Instruct (Yang et al., 2024a), and Llama3.1-405b-instruct (Llama3.1-405B) (Touvron et al., 2023). For closed-source LLMs, we evaluated OpenAI's GPT-3.5-Turbo-0125 (GPT-3.5) and Alibaba Cloud's Qwen-Max-2024-09-19 (Qwen-Max). For open-source VLLMs, we evaluated LLaVA1.6-Vicuna-7B, LLaVA1.6-Mistral-7B, Qwen2-VL-7B-Instruct, and Qwen-VL-Max-2024-08-09 (Qwen2-VL-72B-Instruct) (Wang et al., 2024b; Bai et al., 2023b). For closed-source VLMs, we evaluated Alibaba Cloud's Qwen-VL-Max-2024-11-19 (Qwen-VL-Max), Google's Gemini-1.5-Flash-8B-001 (Gemini1.5-Flash-8B) (Team et al., 2024), OpenAI's GPT-4o-Mini-2024-07-18 (GPT-4o-Mini) (Hurst et al., 2024), Google's Gemini-1.5-Flash-002 (Gemini1.5-Flash) (Team et al., 2024), OpenAI's GPT-4o-2024-11-20 (GPT-4o) (Hurst et al., 2024), Google's Gemini-1.5-Pro-002 (Gemini1.5-Pro) (Team et al., 2024), and Anthropic's Claude-3.5-Sonnet-20241022 (Claude3.5-Sonnet).

In our evaluation, each question is asked independently with a corresponding prompt, eliminating the influence of historical context. The example of the prompt and question used in the evaluation is as follows:

Prompt: Based on the example provided, answer the question by selecting the most appropriate choice. Return your answer and rationale strictly in JSON format.

###Example Input: { "index": 000001, "question": "What color is the Fuji apple?", "options": { "A": "Yellow", "B": "Green", "C": "Red", "D": "Blue" } }
###Example Output: { "index": 000001, "answer": "C", "rationale": "Different apple varieties come in different colors, and Fuji apples are typically red." }

###Question: { "index": 4064, "question": "How many triangles are there in the uppercase letter [A]?", "options": { "A": "2", "B": "3", "C": "0", "D": "1" } }
Return only the JSON.

Correct Option	**A**	**B**	**C**	**D**
Trial 1	338	371	325	366
Trial 2	323	353	357	367

Table 10: Distribution of correct options across trials.

To ensure experiment reproducibility, the decoding temperature was set to 0.

To enhance the reliability of evaluation results and mitigate potential positional biases in model responses, we conducted two independent experiments for each model across the entire dataset. Each question was presented twice with randomly shuffled answer choices. Different random seeds were employed in each experiment to determine the position of the correct answer. The distribution of correct answers across these two experiments is presented in Table 10.

D Result Details

D.1 Fine-Tuning on PerceptualQA Dataset

We further investigated the feasibility of directly utilizing the `PerceptualQA` dataset as training data for fine-tuning models, with the explicit aim of enhancing their embodied knowledge comprehension. Table 11 presents comparative accuracies before and after the fine-tuning process.

Model	Method	Before	After
GPT-4o-Mini	SFT	58.21	**79.29**
Qwen2-7B	LoRA	47.50	45.00
Qwen2-7B-Instruct	LoRA	46.43	46.43
Qwen2-VL-7B	LoRA	47.86	48.21

Table 11: Accuracy of GPT-4o-Mini and Qwen2 series models on `PerceptualQA` before and after fine-tuning.

For our experimental design, we randomly shuffled the `PerceptualQA` dataset and allocated 80% for training and the remaining 20% for testing. The training data comprised 1,120 question-answer pairs, stored in JSON format. Notably, since the Qwen2-7B series models demonstrated suboptimal performance in generating structured outputs after fine-tuning, we configured these models to output only the correct answer option. In contrast, for GPT-4o-Mini, we implemented a more comprehensive output format including both the correct option and its specific content. This distinction in output requirements may partially account for performance differentials between models. We conducted supervised fine-tuning (SFT) on the GPT-4o-Mini model, while employing Low-Rank Adaptation (LoRA) tuning (Hu et al., 2021) for the Qwen2 series models due to computational resource constraints.

The precise experimental configurations were as follows: The GPT-4o-Mini model underwent fine-tuning for 3 epochs with a batch size of 2 and a learning rate multiplier of 1.8. To ensure reproducibility, we established a seed value of 656770515 for the fine-tuning procedure. We fine-tuned three models from the Qwen2 series using the LLaMA Factory framework (Zheng et al., 2024): Qwen2-7B, Qwen2-7B-Instruct, and Qwen2-VL-7B-Instruct. We employ the Low-Rank Adaptation (LoRA) (Hu et al., 2021) method for parameter-efficient fine-tuning, targeting the query and value projection matrices in the attention layers. For Qwen2-7B, we implemented the Alpaca template for instruction tuning, while Qwen2-7B-Instruct and Qwen2-VL-7B-Instruct utilized the Qwen and Qwen2-VL templates, respectively. All models underwent fine-tuning for 3 epochs with a batch size of 4 and a learning rate of 5e-6, utilizing the cosine learning rate scheduler. We applied gradient accumulation every 4 steps and conducted evaluations at 1,000-step intervals. The maximum sequence length was configured at 4,096 tokens, with weight decay set to 0.1 and a warm-up period of 100 steps. Training was conducted using 16-bit bfloat16 precision, with each model allocated to a separate NVIDIA GeForce RTX 4090 GPU.

Fine-tuning data examples for both GPT-4o-Mini and Qwen2 series models are presented herein. During the inference phase, we employed identical data formatting as used in the fine-tuning process, with the "answer" field information removed to serve as the prompt.

Fine-tuning Data Sample for GPT-4o-Mini:

{"messages": [{"role": "system", "content": "Based on the example provided, please answer the question by selecting the most appropriate choice. Provide your answer in JSON format, and return only the JSON.###Example Input{"question": "What color is the Fuji apple?","options": {"A": "Yellow","B": "Green","C": "Red","D": "Blue"}}###Example Output{"answer": "C.Red"}###Question"}, {"role": "user", "content": "{"question": "How many triangles are there in the uppercase letter [A]?", "options": {"A": "2", "B": "3", "C": "0", "D": "1"}}Return only the JSON."}, {"role": "assistant", "content": "{"answer": "D.1"}"}]}

Fine-tuning Data Sample for Qwen2 series:

{"instruction": "Based on the example provided, please answer the question by selecting the most appropriate choice. Provide your answer in JSON format, and return only the JSON.###Example Input{"question": "What color is the Fuji apple?","options": {"A": "Yellow","B": "Green","C": "Red","D": "Blue"}}###Example Output{"answer": "C"}###Question{"question": "How many triangles are there in the uppercase letter [A]?", "options": {"A": "2", "B": "3", "C": "0", "D": "1"}}Return only the JSON.", "input": "", "output": "D"}

D.2 Cross-Language Generalization on PerceptualQA Task

In order to validate the cross-language generalizability of our findings, we assessed the performance of the Qwen2 series models on the Chinese `PerceptualQA`. As illustrated in Table 12, the results indicate that VLMs do not exhibit a significant improvement compared to the text-only models. Furthermore, comparative analysis of model performance across both Chinese and English versions of `PerceptualQA` reveals consistent overall performance, confirming the reliability of our evaluation approach across these two languages. Both English and Chinese versions of the dataset are made available.

Model	All Sensory	Visual	V-CA	V-CN	V-GT	V-S	V-B	Non-Visual	A	T	G	O
English												
Qwen2-7B	49.36	39.85	42.00	58.75	35.00	28.75	34.75	73.13	68.50	62.00	83.00	79.00
Qwen2-7B-Instruct	47.82	35.80	41.00	59.75	26.75	22.00	29.50	77.88	73.50	68.00	82.50	87.50
Qwen2-VL-7B-Instruct	51.00	41.70	43.75	59.75	38.25	31.00	35.75	74.25	69.50	64.00	86.00	77.50
Chinese												
Qwen2-7B	51.72	41.95	47.50	64.75	34.75	22.50	40.25	76.13	62.50	70.50	87.00	84.50
Qwen2-7B-Instruct	51.54	39.90	46.50	60.00	33.25	24.00	35.75	80.63	72.50	73.00	86.50	90.50
Qwen2-VL-7B-Instruct	52.54	41.65	52.75	58.00	37.50	25.75	34.25	79.75	71.00	71.50	88.50	88.00

Table 12: Accuracy of `PerceptualQA` in English and Chinese.The full task names are Visual-Color Attributes(V-CA), Visual-Colors in Nature(V-CN), Visual-Geometry and Transformations(V-GT), Visual-Symbols(V-S), Visual-Body(V-B), Auditory(A), Tactile(T), Gustatory(G), and Olfactory(O).

D.3 Potential Confounding Variables in PerceptualQA Task

We analyzed all questions in the `PerceptualQA` dataset for length, average word frequency, and average Age of Acquisition (AoA) scores, and then conducted several analyses.

D.3.1 Method

We computed question length based on token count including all options. Word frequency was calculated using the `wordfreq` Python library. Average AoA scores were derived for each question and its four answer options using Kuperman et al.'s (2012) established ratings, excluding words absent from the Kuperman dataset.

D.3.2 Results across tasks

We calculated the mean and standard deviation of three metrics for each task. Our analysis revealed that the V-S, V-GT, and V-B questions, which demonstrated the lowest performance, did not exhibit significantly higher or lower values in sentence length or AoA scores. However, these question types displayed slightly higher word frequencies compared to other types. This likely stems from their focus on basic geometric shapes, numbers, and body-related nouns, in contrast to other types that involved animal and plant names. Based on these three linguistic metrics, we concluded that question type itself exerts a more substantial impact on model performance than these potential confounding variables.

D.3.3 Comparison between correctly and incorrectly answered questions

We further conducted statistical analyses comparing questions answered correctly versus incorrectly by 24 models.

Regarding question length, most models showed similar length distributions between both question sets, with typical differences of only 1-3 words. However, Mann-Whitney U tests revealed incorrectly answered questions were significantly longer. For most models, incorrectly answered questions had significantly higher average word frequencies. No significant differences in AoA metrics were observed between the two question sets for most models. This analysis suggests that question length and average word frequency may influence model performance if we don't take task type into consideration.

Task	Length	Average word frequency	Average AoA scores
V-CA	18.50 (2.38)	5.34 (0.17)	5.55 (0.71)
V-CN	54.58 (24.87)	4.76 (0.22)	6.07 (0.75)
V-GT	33.45 (11.74)	5.50 (0.31)	5.77 (0.58)
V-S	23.02 (7.45)	5.60 (0.31)	5.30 (0.45)
V-B	30.23 (5.54)	5.70 (0.18)	5.25 (0.40)
A	30.29 (10.62)	5.47 (0.32)	4.79 (0.41)
T	13.96 (3.84)	4.85 (0.29)	5.81 (0.85)
G	26.75 (19.20)	5.22 (0.20)	5.29 (0.52)
O	16.84 (5.43)	5.07 (0.38)	5.89 (0.64)

Table 13: Statistical properties of questions across tasks. Values are presented as mean (standard deviation).

D.4 Detailed Results on PerceptualQA Task

We present the results of all models across each subtask, with detailed results shown in Table 14.

D.5 More Examples of models' PerceptualQA prediction failures

We present additional examples of models' `PerceptualQA` prediction failures in Table 15.

D.6 Human Baseline Details for PerceptualQA Task

For `PerceptualQA`, we established a human baseline with native speakers to provide a reference for evaluating model performance.

Specifically, we recruited seven graduate students, with each question receiving responses from two different participants. We randomly selected 25% of the questions from each of the nine tasks, shuffled them, and divided them into seven sets of

Model	All Sensory	Visual	V-CA	V-CN	V-GT	V-S	V-B	Non-Visual	A	T	G	O
Human Baseline	86.00	85.20	66.00	88.00	89.00	93.00	90.00	88.00	88.00	88.00	86.00	90.00
Llama3.2-3B-Instruct	47.71	38.25	47.50	46.75	31.75	27.75	37.50	71.38	67.00	61.50	82.00	75.00
Vicuna-7B	38.25	32.60	30.50	33.00	35.25	30.25	34.00	52.38	48.50	38.00	65.00	58.00
LlaVa1.6-Vicuna-7B	41.64	35.25	37.50	37.25	30.50	31.75	39.25	57.63	49.50	49.00	70.50	61.50
Mistral-7B	42.96	31.40	42.50	35.50	26.25	24.25	28.50	71.88	67.00	65.50	78.50	76.50
Llava1.6-Mistral-7B	45.64	35.90	45.00	40.00	30.75	25.00	38.75	70.00	71.00	66.50	70.50	72.00
Qwen2-7B	49.36	39.85	42.00	58.75	35.00	28.75	34.75	73.13	68.50	62.00	83.00	79.00
Qwen2-7B-Instruct	47.82	35.80	41.00	59.75	26.75	22.00	29.50	77.88	73.50	68.00	82.50	87.50
Qwen2-VL-7B-Instruct	51.00	41.70	43.75	59.75	38.25	31.00	35.75	74.25	69.50	64.00	86.00	77.50
Llama3.1-8B	48.54	39.05	46.00	49.25	33.50	27.50	39.00	72.25	70.50	61.00	82.00	75.50
Gemma2-9B	51.89	40.65	45.50	55.50	38.00	28.25	36.00	80.00	81.50	72.50	83.00	83.00
Gemma2-27B	55.39	44.90	51.50	61.25	43.75	30.75	37.25	81.63	81.50	73.50	85.00	86.50
Llama3.1-70B	59.71	49.85	65.50	63.25	47.75	32.25	40.50	84.38	84.00	77.50	89.00	87.00
Qwen2-72B-Instruct	62.32	52.75	69.75	69.75	45.25	35.00	44.00	86.25	84.50	84.00	88.00	88.50
Qwen2-VL-72B-Instruct	63.89	54.45	68.75	73.00	51.00	32.75	46.75	87.50	83.00	86.50	89.50	91.00
Llama3.1-405B	63.46	54.55	69.75	72.00	47.50	37.00	46.50	85.75	83.50	80.50	90.00	89.00
GPT-3.5	50.46	39.50	40.50	42.50	43.75	29.25	41.50	77.88	73.00	69.50	84.50	84.50
Qwen-Max	68.71	61.05	85.00	74.50	57.25	44.25	44.25	87.88	87.00	84.50	91.00	89.00
GPT-4o-Mini	57.18	46.35	57.25	62.25	41.25	31.50	39.50	84.25	84.50	78.50	86.00	88.00
Gemini1.5-Flash-8B	54.39	44.55	56.75	59.50	42.50	28.50	35.50	79.00	82.00	67.50	85.00	81.50
Gemini1.5-Flash	56.07	45.20	65.75	48.25	42.75	31.25	38.00	83.25	81.00	76.50	88.50	87.00
GPT-4o	68.46	59.45	70.00	77.50	51.25	44.50	54.00	91.00	92.50	86.00	91.00	94.50
Gemini1.5-Pro	65.21	56.55	70.00	70.50	54.00	39.00	49.25	86.88	82.50	83.50	89.50	92.00
Claude3.5-Sonnet	69.04	60.00	70.25	75.25	57.00	44.75	52.75	91.63	92.00	90.00	92.00	92.50
Qwen-VL-Max	64.68	55.30	68.75	79.25	50.00	34.25	44.25	88.13	84.50	87.00	90.50	90.50

Table 14: Detailed Results on `PerceptualQA`. The full task names are Visual-Color Attributes(V-CA), Visual-Colors in Nature(V-CN), Visual-Geometry and Transformations(V-GT), Visual-Symbols(V-S), Visual-Body(V-B), Auditory(A), Tactile(T), Gustatory(G), and Olfactory(O).

Question & Options & Answers and Rationales from Claude3.5-Sonnet
[V-GT] If a rectangle has long sides on the top and bottom, and short sides on the left and right, and it is flipped around its left short side, where is the original bottom edge now positioned? **A. Bottom** B. Top C. Right D. Left B. When a rectangle is flipped around its left short side (like a door hinge), the bottom edge will rotate 180 degrees upward, ending up at the top position. This is similar to how a door swings open, with the hinge being the left side in this case.
[V-S] If the number 3 is horizontally cut in half, which direction does the opening of the bottom part face? A. Up **B. Left** C. Down D. Right A. When a number 3 is cut horizontally in half, the bottom part resembles a curved line with the opening facing upward, similar to the bottom half of a sideways heart shape.
[V-B] When standing and bending one leg backward at the knee, which is higher: the ankle of the raised leg or the knee of the supporting leg? **A. It depends** B. They are at the same height C. The knee of the supporting leg D. The ankle of the raised leg D. When bending one leg backward at the knee while standing, the ankle of the raised leg typically rises above the knee level of the supporting leg. This is due to the natural biomechanics of the leg and the way the knee joint bends, allowing the foot and ankle to be lifted higher than the stationary knee of the supporting leg.

Table 15: More Examples of models' `PerceptualQA` prediction failures and answers from Claude3.5-Sonnet. Task types are shown in brackets. The correct answers are shown in **bold**.

50 questions each, ensuring proportional representation of all nine subtasks across sets. Each participant completed two sets (100 questions total), following a fixed task order within each set. This design minimizes individual participant influence on task-level accuracy and establishes a reliable human performance benchmark. For each subtask, the human baseline was computed as the average accuracy of the two responses.

All participants reported good physical health with no relevant impairments (e.g., color blindness). Screening was conducted during recruitment based on self-reported information to confirm participants met the health criteria and could complete the assessment within the designated timeframe. Participants completed the tasks independently via an online survey platform. The use of LLM tools was prohibited. On average, participants required 37.32 minutes to complete the assigned questions. Recruitment occurred through announcements within graduate student communities at our institution. Participants were explicitly informed that their responses would be used for research purposes. No personally identifiable information was retained in association with participant responses. Each participant received compensation of 50 RMB, equivalent to 0.5 RMB per question.