Bridging Perception and Language: A Systematic Benchmark for LVLMs' Understanding of Amodal Completion Reports

Amane Watahiki[1] **Tomoki Doi**[1] **Taiga Shinozaki**[2,1]
Satoshi Nishida[3,4,5,6] **Takuya Niikawa**[7] **Katsunori Miyahara**[5] **Hitomi Yanaka**[1]
[1]The University of Tokyo [2]Keio University [3]NICT
[4]Osaka University [5]Hokkaido University [6]CiNET [7]Kobe University
amanew@g.ecc.u-tokyo.ac.jp, snzktig@keio.jp, s-nishida@nict.go.jp
{doi-tomoki701, hyanaka}@g.ecc.u-tokyo.ac.jp, niitaku11@gmail.com, kmiyahara@chain.hokudai.ac.jp

Abstract

One of the main objectives in developing large vision-language models (LVLMs) is to engineer systems that can assist humans with multimodal tasks, including interpreting descriptions of perceptual experiences. A central phenomenon in this context is amodal completion, in which people perceive objects even when parts of those objects are hidden. Although numerous studies have assessed whether computer-vision algorithms can detect or reconstruct occluded regions, the inferential abilities of LVLMs on texts related to amodal completion remain unexplored. To address this gap, we constructed a benchmark grounded in Basic Formal Ontology to achieve a systematic classification of amodal completion. Our results indicate that while many LVLMs achieve human-comparable performance overall, their accuracy diverges for certain types of objects being completed. Notably, in certain categories, some LLaVA-NeXT variants and Claude 3.5 Sonnet exhibit lower accuracy on original images compared to blank stimuli lacking visual content. Intriguingly, this disparity emerges only under Japanese prompting, suggesting a deficiency in Japanese-specific linguistic competence among these models.

Keywords: amodal completion; large vision–language model; evidentials; Basic Formal Ontology

Introduction

A key goal in developing large vision–language models (LVLMs) is to enable them to support humans in multimodal tasks, which requires accurate interpretation of texts describing perceptual experiences of images. A central phenomenon in human perception is *amodal completion* (AC), in which individuals perceive whole objects even when parts are occluded. For example, a subject might report "seeing two rectangles" in Figure 1, despite one being partially hidden. This perceptual "filling in" occurs without direct sensory input and so is called "amodal" completion.

AC is essential in daily life (Nanay, 2018) and appears frequently in human narratives. Therefore, for LVLMs to handle these narratives, they must interpret AC-related descriptions accurately. However, AC poses unique challenges: although the occluded parts are not physically visible, they appear to be there, leading subjects to use perceptual verbs such as "looks like." These verbs often coexist with the object's invisibility, creating seemingly paradoxical expressions. Humans resolve this paradox naturally, but it remains unclear whether LVLMs can do the same.

Previous studies have examined how models reconstruct occluded object properties such as shape or color, but the focus was mostly on visual processing (Ao et al., 2023), leaving textual inference in AC underexplored. Furthermore, AC was typically classified into only two or three coarse categories, therefore lacking clarity and nuance. Researchers often assume that the type of completed object is obvious, and few provide criteria for consistent tagging. This reliance on intuition hinders standardization and interoperability.

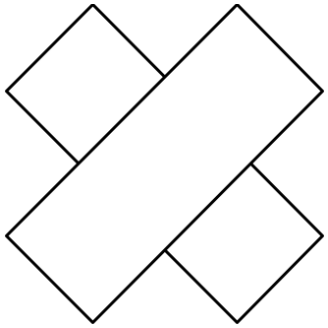

Figure 1: Example of amodal completion (Michotte, 1991). Although this figure is drawn on a single plane, it appears as if one rectangle is partially hidden under another rectangle, and the perceiver completes the occluded part "amodally."

To address the aforementioned issues, we introduce **VACT** (**V**isual **A**modal **C**ompletion with **T**exts), a benchmark using a two-choice question–answer format, with categories tagged based on **Basic Formal Ontology (BFO)**, which provides a structured, interoperable framework for categorizing entities. By comparing model and human performances across these fine-grained categories, we identify where LVLMs struggle. This also serves as a case study demonstrating the utility of ontologies such as BFO in the evaluation of artificial intelligence (AI).

Interestingly, some LLaVA-NeXT models and Claude 3.5 Sonnet perform less well on original images and better on blank ones for some categories, but only in the Japanese setting. We hypothesize that this is due to difficulty in distinguishing the typical versus evidential use of perceptual verbs in Japanese. Evidential use of a perceptual verb indicates the source of the speaker's information, often weakening its original lexical meaning. This allows native Japanese speakers to interpret seemingly paradoxical AC descriptions. While English also marks evidentiality, it does so through different syntactic strategies. This cross-linguistic variation may limit the ability of LVLMs to transfer linguistic knowledge from English to Japanese. Although still a hypothesis, our findings

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

point to a promising direction for future work: bridging perception and language by exploring how evidentiality shapes the expression and interpretation of perceptual experiences across languages.

Related Work

In previous research, various datasets were constructed to assess the AC capabilities of models in tasks involving *image processing* (Ao et al., 2023). However, none of these tasks assess the inference capabilities of LVLMs with respect to *textual descriptions of AC*, which is the focus of the present study.

Different forms of AC may reveal specific strengths and limitations in LVLM performance. However, the classifications used commonly in previous research are (i) insufficiently fine-grained to yield insights into the specific characteristics of each model and (ii) lack clear definitions for each category. For instance, Ao et al. (2023) identified shape completion, appearance completion, and order perception as key categories in recent AC tasks in image processing. Similarly, other studies classified AC based on whether it involves shape or color completion (Gerbino, 2020; Pessoa et al., 2001). Van Lier & Gerbino (2015) divided AC into two categories, two-dimensional (2D) and three-dimensional (3D), while Tse (1999) argued that all types of completion can be understood as volume completion. However, as will be shown, the range of objects subject to AC extends beyond these categories. For example, when we see a die, its back is occluded by its front surface, but we still perceive it as a cube. In this case, one could argue that what is being completed is not merely the color or shape but the die as a whole.

Turning to less-artificial real-world images, it becomes evident that existing categories are too vague and coarse-grained to support a systematic and comprehensive classification scheme. Is the back surface of a drinking glass 2D or 3D? On one hand, one might think that it is a 2D object because what is at stake here is something's "surface." On the other hand, one can also believe that no 2D object can exist because nothing can exist without thickness in the real world. There is nothing wrong with both ways of thinking. In the absence of clear criteria, annotating everyday instances of AC becomes challenging.

Proposed Dataset

Overview

In this study, we construct the VACT benchmark dataset, which presents a two-alternative forced choice (2AFC) task requiring models and participants to identify the correct description of an amodally complemted object in an image. Sample images and associated questions from the dataset are shown in Table 1. This benchmark is used to evaluate the inference capabilities of LVLMs on text-based AC tasks and to compare their performance against human judgments.

To assess this ability with a finer-grained and more formally defined taxonomy of AC, VACT incorporates a classification framework based on BFO (Arp et al., 2015), which is a top-level ontology designed to structure scientific knowledge systematically across disciplines in a consistent and interoperable fashion. As a top-level ontology, BFO offers high-level categories applicable across all domains of scientific inquiry.

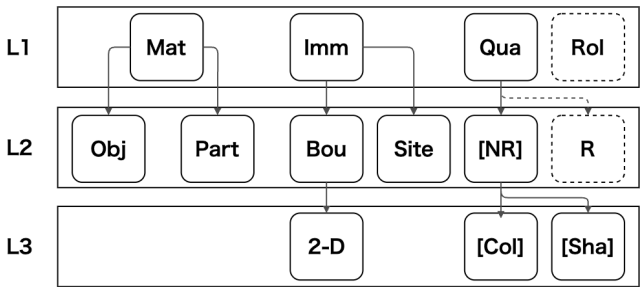

Figure 2: Partial taxonomy of BFO continuant categories. Dotted lines denote categories outside the scope of this study, and those not defined in BFO are enclosed in square brackets ([ ]).

Two key advantages motivate the adoption of BFO in this study: (i) BFO is designed to classify all entities into one of its categories, enabling it to accommodate objects and phenomena that do not fit into existing classifications, including those identified by Ao et al. (2023); (ii) BFO provides a clear definition or elucidation of each BFO category, facilitating systematic classification and reducing ambiguity.

Ontologies such as BFO classify recurring features of the real world and represent their interrelations through hierarchical taxonomies. Such taxonomies can be formalized as graphs, with nodes denoting categories and edges indicating the subtype or subclass (*is_a*) relation between these categories. With BFO, we obtain a structured classification of AC as shown in Figure 2.

Details of BFO Categories

Here, we briefly outline the categories relevant to VACT, with Figure 2 showing the part of the BFO taxonomy that is relevant to our study; see Arp et al. (2015) for a comprehensive treatment of BFO. VACT uses subtypes of both independent and dependent continuants as defined in BFO: independent continuants are entities that exist on their own, while dependent continuants "inhere" in or depend upon independent continuants for their existence. This study refers to categories directly under this level as **first-layer (L1)** categories in our framework. Linked via *is_a* relationships, their subcategories are designated as **second-layer (L2)** or **third-layer (L3)** categories, depending on their depth in the hierarchy. Figure 2 shows the parent–child relationships among these categories.

Classified as an L1 category, the **material entity (Mat)** refers to an independent continuant that includes matter as a constituent. Material entities are three-dimensionally extended and temporally persistent, regardless of the duration of persistence. **Object (Obj)** and **object part (Part)** at L2 are subtypes of material entity. A die is an example of an

Table 1: Example questions. Although this table shows three texts for one image, VACT consists of two-choice QA tasks; for each image and gold answer, two question pairs are prepared: (gold, miscompletion) and (gold, no-completion) pairs.

Question	A	B	C	D
Image				
Gold Answer	Although it is not visible in the photo, the glass seems to have a back with a shape similar to its front.	Although it is not visible in the picture, the brown horse's head appears to be connected to the brown horse's torso.	Although the illustration is drawn on a flat surface, what is depicted appears to be a cube rather than a hexagon.	Although the photo is flat, it appears that a vast space extends behind the central bird.
Miscompletion	Although it is not visible in the photo, the end of the straw appears to be shaped like a spoon.	Although the photo is flat, the white horse appears to be behind the brown horse.	Although it is not depicted in the illustration, the opposite side of the 1 face appears to have a 6 face.	Although they are not visible in the photo, the black spots on the back of the bird in the center appear to be present.
No-completion	Although it is not shown in the picture, the end of the straw doesn't appear to continue all the way into the glass.	Although the photo is flat, the white horse does not appear to be in front of the brown horse.	Although it is not depicted in the illustration, it doesn't look like there is a white square surface on the opposite side of the eye of 1.	Although it is not visible in the photo, the bird in the center does not appear to have white feathers on its back.
BFO Category	2-D, Shape	Part, Color	Object, Shape	Site

object, whereas the junction between a brown horse's head and its torso serves as an example of an object part. In contrast, the **immaterial entity (Imm)** is an independent continuant that lacks material constituents. The site and object fiat boundaries are the subtypes of immaterial entities. **Object fiat boundary (Bou)** is a non-material entity (0D, 1D, or 2D) that does not include a spatial region as a part. Intuitively, it denotes the boundary of a material entity at the interface with its surroundings, and this study considers only **2D boundaries (2-D)**. Consider question A in Table 3. The object referenced in the gold-standard answer is the back of the glass. The back of the glass is the surface that exists precisely where the glass meets its surroundings, and the surface itself does not contain any material part. Based on the 2D boundary definition and the correct description, we conclude that the amodally completed object—the back of the glass—is a 2D boundary. Another immaterial entity, **site (Site)**, is a 3D non-material entity that is either (i) bounded partially or wholly by a material entity or (ii) a 3D non-material part that satisfies condition (i). For example, in question D of Table 1, the amodally completed entity is a site. Here is our rationale. Predicting the correct answer requires AC of the space behind the bird. Because the space is a 3D non-material entity that is partially bounded by a material entity (i.e., the bird), what is completed in the correct answer is site.[1]

We now turn from independent continuants to dependent continuants. Dependent continuants are categorized into two primary subtypes: qualities and roles. Because it is controversial to think that **roles (Rol)** and **relational qualities (R)** are amodally completed, this study focuses exclusively on qualities. **Non-relational qualities (NR)** are the typical entities that are considered to be amodally completed. **Quality (Qua)** is an individual dependent continuant that when it inheres in a continuant is fully realized and instantiated within it. Representative examples of non-relational qualities include **color (Col)** and **shape (Sha)**.

Consider again question A in Table 1. If the correct description holds, then it implies that the back of the glass has a specific shape, presumably similar to its front. Thus, by selecting this as the correct answer, we engage in AC of the shape of the glass's back.[2]

Color and shape are subtypes of qualities but are not formally part of BFO because they are domain-specific categories. For example, there is no room for color in the field of physics.[3] Nevertheless, they are mentioned frequently in existing AC classification studies in the field of computer vision, so we include them in our classification framework.

[1] In fact, the "space behind the bird" would move in tandem with the bird. This mobility feature differentiates sites from other immaterial entities, such as spatial regions.

[2] When viewing this image, you might think that the surface continuing to the back of the glass has some specific color. However, the correct answer does not imply that the back of the glass has a specific color. Accordingly, we classify this instance as 2D boundary completion rather than color completion.

[3] Color here should not be conflated with "color charge" discussed in physics.

Dataset Construction

Data Collection

This study builds on a dataset from Nishida et al. (2024) (comprising 69 questions), which was modified to align with our research objectives and extended with additional images and questions. Nishida et al. (2024) introduced two-choice tasks to assess human understanding of experiences known as horizon consciousness (Gallagher & Zahavi, 2008), and because certain aspects of horizon consciousness are related conceptually to AC, selected questions and images were adapted for this study. Image data were mainly sourced from Pixabay,[4] a royalty-free image repository. We added the images and crafted additional answer choices based on three key criteria: (i) part or the whole of the object(s) in the image is occluded in a way that evokes AC, (ii) balanced representation of the categories of completed objects, and (iii) even distribution of image types (drawings vs. photographs).

Answer Choices

For each question, one of the two answer choices was constructed to be easily identifiable as correct by human participants. The other incorrect answer choice was categorized into two types: miscompletion (a description of incorrect completion) and no-completion (a description that fails to involve any completion). Table 1 shows examples of correct and incorrect answers. Although Table 1 presents three textual options per image, both model and human evaluations were conducted using two-choice question-answering tasks. Accordingly, two alternative choice pairs were constructed: (gold, miscompletion) and (gold, non-completion). Each question was annotated with up to two BFO categories corresponding to the entity that must be amodally completed to select the correct answer. The number of questions corresponding to each category is indicated in Table 1.

The primary experiment was conducted in Japanese, with an additional model evaluation performed in English. Prompts and answer choices were translated from Japanese into English.

Human Evaluation

Human evaluation was conducted to ensure that the correct answers were easily identifiable by human participants. Only the Japanese version of the prompts (used in model evaluation) was presented during the human evaluation phase.

For each image, we presented participants with either a (gold, miscompletion) or (gold, no-completion) pair of choices and asked them to select the appropriate description based on the given prompt. To avoid bias, the questions were divided into two sets to ensure that each participant viewed only one version of each image–question pair.

In total, 101 responses were collected online during December 6–26, 2024 (63 males, 38 females; age distribution: eight in their teens, 86 in their 20s, six in their 30s, and one in their 40s). Questions with correct response rates below 50% were excluded as unreliable. As a result, 122 questions were retained as valid: half featured miscompletion choices, and the other half no-completion ones.

[4] https://pixabay.com/

Verification

Models

We evaluated GPT-4o[5] and Claude 3.5 Sonnet[6] as representative commercial LVLMs.[7] Additionally, we assessed three scales of the LLaVA-NeXT series from the Hugging Face Hub—LLaVA-NeXT-34B, LLaVA-NeXT-72B, and LLaVA-NeXT-110B[8]—to examine whether and how model performance scales with parameter size. Their respective base LLMs are Yi-34B, Qwen1.5-72B, and Qwen1.5-110B. For brevity, we refer to Claude 3.5 Sonnet as Claude, and the LLaVA-NeXT variants as LLaVA-34B, LLaVA-72B, and LLaVA-110B.

Evaluation Methods and Metrics

Zero-shot evaluations were conducted using accuracy as the primary performance metric. We also evaluated the accuracy when blank dummy images were presented in place of the original images to explore whether the models might select plausible answers based solely on text.

Prompts

Below is an example of the prompts used in the experiment.

Prompts used in experiment

> Instructions: You will see an image and two sentences describing the experience of looking at the image. One of the sentences is correct, and the other is incorrect. Please answer which of the two sentences is appropriate by writing the number of the sentence and only the number. Explanation 1: [...] Explanation 2: [...]

The complete prompt is obtained by replacing the placeholders ([...]) with textual descriptions, such as those shown in Table 1. To avoid biases caused by the order of the descriptions or the labeling of the choices, the answer choices were shuffled randomly. This ensured a roughly equal distribution of correct answers between Explanation 1 and Explanation 2. The same prompts were used for both original and dummy image conditions. This prompt was slightly more detailed than the version used for human evaluation.

Results

First, Table 2 shows each model's accuracy for the entire dataset. The English result is also displayed. Commercial models such as GPT-4o and Claude approached human-level performance, whereas the LLaVA models showed room for

[5] https://openai.com/index/GPT-4o-system-card/

[6] https://assets.anthropic.com/m/61e7d27f8c8f5919/original/Claude-3-Model-Card.pdf

[7] We conducted our experiment from December 2024 to January 2025.

[8] Liu et al. (2024); Li et al. (2024)

Table 2: Accuracy in total. Total accuracy is calculated by the number of questions models could answer correctly divided by the total number of questions. Numbers in parentheses indicate the accuracy in dummy questions, where blank images were provided instead of the original images. Accuracy for humans represents an average value. We did not conduct human experiments with English instructions.

Model	GPT -4o	Claude	LLaVA -34B	LLaVA -72B	LLaVA -110B	Human
Accuracy (Japanese)	89 (77)	93 (91)	56 (50)	67 (54)	60 (56)	94
Accuracy (English)	88 (80)	90 (88)	70 (52)	79 (62)	90 (78)	-

improvement. In English, however, the performance gap between commercial and open models was reduced.

Next, Table 3 shows the accuracy for LVLMs and human participants, grouped by the type of amodally completed object. By employing BFO-based classifications, this categorization revealed specific areas of weakness for each model. For example, GPT-4o showed reduced performance in understanding AC involving object parts (77%), while Claude performed relatively poorly on color completion tasks (83%). All LLaVA variants struggled with boundary (56-57%) and color (53-56%) completions. Accuracy in color is also lower in GPT-4o (86%) compared to other categories, suggesting that color-based AC remains a persistent challenge for LVLMs.

Additional findings from Table 3 are summarized below. First, the models achieved relatively high accuracy rates for dummy questions. Interestingly, in some categories and models, the latter achieved higher accuracy with dummy images than with the original images. Claude outperformed its original-image results in two categories when evaluated with dummy images. Second, LLaVA-110B performed worse than the smaller LLaVA-72B. Across all categories, there were instances where LLaVA-72B answered correctly but LLaVA-110B did not. Finally, we analyzed error patterns based on whether the incorrect answer involved miscompletion or no-completion (Table 4). A particularly notable finding is that for all the models, the proportion of no-completion errors increased when the models were shown original images compared to dummy images.

Discussion

The results indicate that each model exhibits specific weaknesses in interpreting texts related to AC, with color completion emerging as a consistent challenge across all models. This finding is critical for deploying LVLM-based systems in real-world scenarios involving diverse types of AC, and it highlights the value of ontological frameworks such as BFO in constructing meaningful and diagnostic benchmarks.

Furthermore, the findings presented in the previous section raise the following questions.

Table 3: Accuracy by BFO categories in Japanese. Numbers in parentheses indicate the accuracy in dummy questions, where blank images were provided instead of the original images. Cells are shaded gray when the accuracy for dummy questions is higher than for the original questions. For each model, the lowest score is made bold when the original image is presented. The value for the human performance represents the average accuracy.

L1	Mat		Imm		Qua	
L2	Obj	Part	Bou	Site	[NR]	
L3			2-D		[Col]	[Sha]
# questions	16	38	54	18	36	12
GPT-4o	88 (75)	**77** (74)	94 (72)	100 (100)	86 (64)	92 (83)
Claude	100 (81)	90 (90)	93 (93)	94 (100)	**83** (89)	100 (83)
LLaVA -34B	69 (62)	**46** (44)	56 (50)	72 (61)	56 (50)	75 (67)
LLaVA -72B	62 (50)	69 (41)	57 (48)	100 (100)	**56** (31)	67 (75)
LLaVA -110B	62 (63)	**51** (49)	57 (52)	89 (83)	53 (50)	67 (67)
Human	95	94	94	92	92	95

Q1) How do models achieve accuracy above chance on dummy questions?

Q2) Why do some models perform better on dummy images than on original images for certain categories?

Q3) Why does LLaVA-110B underperform compared to the smaller LLaVA-72B?

We also conducted the same experiments using English prompts. In contrast to the Japanese results, the English experiments showed a more expected trend: model performance improved consistently with larger base LLM sizes, and models generally performed better when original images were presented compared to dummy images (except the "site" category). These observations suggest that the abovementioned anomalies are specific to the Japanese experiment. Below, we propose possible explanations for each question.

Hypothesis 1 Texts affirming the existence of amodal completion appear more frequently in the training dataset than those denying it.

If the models are biased toward selecting gold or miscompletion answers over no-completion alternatives because of the distribution of training data, then this could explain their tendency to perform above chance (50%) even when presented with dummy images.

Hypothesis 2 The models have visual limitations in amodal completion.

How can Q2 and Q3 be addressed? One possible explanation centers on the models' visual capability. While LVLMs may select answers correctly based on textual cues alone, they

Table 4: Error analysis. The number of incorrect answers by a pattern of erroneous choices. The percentage of no-completions among all incorrectly answered questions increases when the original image is presented. Numbers in parenthesis indicate the accuracy in dummy questions.

	Miscompletion	No-completion	% no-completion among errors
GPT-4o	3 (7)	10 (21)	77% (75%)
Claude	4 (8)	5 (3)	56% (27%)
LLaVA-34B	25 (30)	29 (31)	54% (51%)
LLaVA-72B	17 (27)	23 (30)	%57 (%53)
LLaVA-110B	24 (30)	25 (24)	51% (44%)

may choose incorrectly when the original image is presented because of limited visual competence in AC. The textual input may assign higher probability to the correct answer, but the visual input could introduce conflicting cues, thereby lowering accuracy. Although this hypothesis is plausible, it does not rule out other contributing factors, which we explore below.

Hypothesis 3 The models do not recognize the evidential function of Japanese perceptual verbs.

Hypothesis 3 answers Q2 and Q3 from the perspective of the models' language understanding capabilities. Because the task involves multimodal inputs, linguistic comprehension plays a crucial role in performance. In particular, the models may struggle to interpret the perceptual verb *mieru*, which appears in every correct description of AC in our dataset. While the literal meaning of *mieru* is "can be seen," according to Shiba (2023), the verb *mieru* undergoes what is called "grammaticalization," a phenomenon in which words that have substantive meaning and can function as independent elements, known as "content words," change into words that have the character of "function words," which are elements that solely serve grammatical functions (Miyake, 2005). Shiba (2023) points out, when combined with the evidential morpheme *yooda* to form *yoo-ni mieru*, it functions as an evidential marker, indicating that the speaker's statement is based on visual evidence. In this evidential usage, native Japanese speakers select the correct answer easily because the marker permits indirect evidence, making it compatible with the fact that the object in question is not directly visible.

Hypothesis 3 may answer Q2—why some models perform better on dummy images than on original images for certain categories—by suggesting that some LVLMs fail to recognize the evidential use of *mieru*. As a result, models may avoid descriptions that (to them) imply an invisible object "can be seen," even if they are capable of amodally completing the occluded parts of the image. This tendency to select the no-completion option is manifested in the fact that all the models choose no-completion options more when they are presented with the original image than with the dummies (Table 4).

It may also answer Q3, namely, why LLaVA-72B outperformed the larger LLaVA-110B. According to (Li et al., 2024), LLaVA-110B performs better than 72B on multimodal benchmarks. This suggests that LLaVA-110B may be more sensitive to visual input. However, if it still struggles to interpret the evidential meaning of *mieru*, then the model may be more hesitant to select the correct answer, particularly when textual cues hinge on this subtle linguistic distinction.

Evidentiality manifests differently across languages. For example, Japanese has both grammaticalized and non-grammaticalized evidential forms, whereas English lacks grammaticalized evidentials, relying instead on verbs and adverbs such as "seem" or "appear" to express similar functions (Yang, 2014). Given the findings of the present study, the diversity of evidential expressions across languages may pose a considerable challenge for LVLMs. Because evidentiality is a fundamental linguistic mechanism for expressing subjective perceptual experiences including AC, its cross-linguistic variability represents an important direction for future research in improving the multilingual reasoning and generalization of LVLMs.

Conclusion

In this study, we evaluated the multimodal capabilities of LVLMs on tasks involving both text and images related to AC. Rather than focusing solely on image processing, our objective was to assess systematically how models comprehended texts describing AC by using a classification framework derived from BFO.

The results showed that while LVLMs demonstrate a generally strong understanding of AC-related texts (comparable to human performance), their accuracy varies considerably depending on the type of object being completed, in contrast to the relatively stable performance observed in human participants. Moreover, experiments using dummy images revealed an intriguing language-specific effect: in Japanese, models such as LLaVA-NeXT and Claude sometimes performed worse with original images than with dummy images. This pattern suggests a limited understanding of Japanese perceptual verbs, which are essential for the correct interpretation of AC-related descriptions.

In future work, we will investigate further the hypothesis that these models fail to grasp the evidential use of perceptual verbs in Japanese. Advancing this line of research may enhance the ability of LVLMs to interpret human-like descriptions of amodal perception and support applications in the automated analysis of subjective narratives related to conscious experience.

Acknowledgments

This work was supported by JSPS KAKENHI grant number JP24H00809, 24K22328, 24KK0189, and 23K00001.

References

Ao, J., Ke, Q., & Ehinger, K. A. (2023). Image amodal completion: A survey. *Computer Vision and Image Understanding*, *229*, 103661. doi: 10.1016/j.cviu.2023.103661

Arp, R., Smith, B., & Spear, A. D. (2015). *Building ontologies with basic formal ontology*. London, England: MIT Press.

Gallagher, S., & Zahavi, D. (2008). *Phenomenological mind: An introduction to philosophy of mind and cognitive science*. New York, NY: Routledge.

Gerbino, W. (2020). Amodal completion revisited. *i-Perception*, *11*(4), 2041669520937323. doi: 10.1177/2041669520937323

Li, B., Zhang, K., Zhang, H., Guo, D., Zhang, R., Li, F., ... Li, C. (2024). *LLaVA-NeXT: Stronger LLMs Supercharge Multimodal Capabilities in the Wild.* Retrieved 2025-4-25, from https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/

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

Michotte, A. (1991). Amodal completion of perceptual structures. In G. Thinés, A. Costall, & G. Butterworth (Eds.), *Michotte's experimental phenomenology of perception* (pp. 140–167). Hillsdale, N.J.: Lawrence Erlbaum, Associates.

Miyake, T. (2005). Gendai nihongo ni okeru bunpōka: Naiyōgo to kinōgo no renzokusei o megutte [grammaticalization in modern japanese: On the continuity of content words and functional words]. *Nihongo no Kenkyū*, *1*(3), 61–76. doi: 10.20666/nihongonokenkyu.1.3_61

Nanay, B. (2018). The importance of amodal completion in everyday perception. *i-Perception*, *9*(4), 1–16. doi: 10.1177/2041669518788887

Nishida, S., Hamada, H. T., Niikawa, T., & Miyahara, K. (2024). Neural correlates of phenomenological attitude toward perceptual experience. *bioRxiv*. Retrieved from https://www.biorxiv.org/content/early/2024/07/10/2024.07.07.602347 doi: 10.1101/2024.07.07.602347

Pessoa, L., Thompson, E., & Noë, A. (2001). Filling-in: One or many? *Behavioral and Brain Sciences*, *24*(6), 1137–1139. doi: 10.1017/S0140525X01230143

Shiba, A. (2023). Chikaku dōshi "mieru" no suitei kōbun e no hirogari: Kyōjitai ni okeru bunpōka [extension of the japanese perception verb mieru to the evidential construction]. *The Journal of Humanities, Nagoya University*, *6*, 57–78. doi: 10.18999/jouhunu.6.57

Tse, P. U. (1999). Volume completion. *Cognitive Psychology*, *39*(1), 37–68. doi: 10.1006/cogp.1999.0715

Van Lier, R., & Gerbino, W. (2015). Perceptual completions. In *The oxford handbook of perceptual organization.* Oxford University Press. doi: 10.1093/oxfordhb/9780199686858.013.040

Yang, L. (2014). Evidentiality in english research articles of applied linguistics: From the perspective of metadiscourse. *Journal of Language Teaching and Research*, *5*(3), 581–591. doi: 10.4304/jltr.5.3.581-591