

# Exploring Affordance and Situated Meaning in Image Captions: A Multimodal Analysis

Pin-Er Chen, Po-Ya Angela Wang, Hsin-Yu Chou, Yu-Hsiang Tseng, Shu-Kai Hsieh

Graduate Institute of Linguistics, National Taiwan University

cckk2913@gmail.com, differe94nt@gmail.com

r10142008@ntu.edu.tw, seantyh@gmail.com, shukaihsieh@ntu.edu.tw

## Abstract

This paper explores the grounding issue regarding multimodal semantic representation from a computational cognitive-linguistic view. We annotate images from the Flickr30k dataset with five perceptual properties: *Affordance*, *Perceptual Salience*, *Object Number*, *Gaze Cueing*, and *Ecological Niche Association (ENA)*, and examine their association with textual elements in the image captions. Our findings reveal that images with Gibsonian affordance show a higher frequency of captions containing ‘holding-verbs’ and ‘container-nouns’ compared to images displaying telic affordance. *Perceptual Salience*, *Object Number*, and *ENA* are also associated with the choice of linguistic expressions. Our study demonstrates that comprehensive understanding of objects or events requires cognitive attention, semantic nuances in language, and integration across multiple modalities. We highlight the vital importance of situated meaning and affordance grounding in natural language understanding, with the potential to advance human-like interpretation in various scenarios.

## 1 Introduction

With the rapid advancement of (multimodal) language models, there has been an urgent demand for advanced natural human-machine interactions, as users expect more native-like interactions with AI systems. To attain this sophistication in multimodal communication, the challenge of *multimodal grounding*, i.e., the pairing of language and other modalities (vision, audio, haptics, etc.), as well as active interaction with the world, has emerged both in the natural language processing (NLP) and computer vision communities.

Basically, *grounding* refers to associating a word or concept with a perceptual experience in the environment, such as an object or event. Recent tasks such as Visual Grounding (VG) or Natural Lan-

guage Visual Grounding<sup>1</sup> have attracted increasing attention, aiming to localize objects/regions in images via natural language expressions (Yang et al., 2022). Transformer-based approaches and pretrained vision-and-language (VL) models have greatly succeeded in image and video captioning (Sun et al., 2019; Radford et al., 2021; Li et al., 2023). However, it is worth noting that the term *grounding* carries different meanings in the NLP and cognitive science community. As Chandu et al. (2021) pointed out, the grounding studies in NLP focus more on the *linking* of text to other modalities. In contrast, the later ones emphasize the *cognitive process* by which the speakers build the common ground to share their mutual information. During this cognitive process, a set of abstract symbols acquire meaning through speakers’ perceptions and situated actions based on sensorimotor experiences. The process is similarly proposed and elaborated in cognitive linguistics with the concept of *construal*, representing how individuals mentally interpret a situation or scene (Langacker, 2008) and account for the choice of alternative linguistic expressions; i.e., two grammatical possibilities for expressing the same situation are two ways of ‘construing’ that situation (Divjak et al., 2020).

We hypothesize the construal of scenes involves common sense knowledge of the presented objects and the visuospatial properties in the images. Therefore, this study systematically examines the grounding issue concerning multimodal semantic representation from a computational cognitive-linguistic view. We operationalize the visuospatial information in the images with five perceptual properties and how they relate to the construal, which is reflected in the presence of two types of textual elements in the captions. Our research addresses the following questions: (1) How do the five perceptual properties in the images correlate? (2) Does the

<sup>1</sup><https://paperswithcode.com/task/natural-language-visual-grounding>

object *Affordance* in an image relate to the distribution of the two types of textual elements ('holding-verbs' and 'container-nouns')<sup>2</sup> in its captions? and (3) How do the other perceptual properties associate with the usage of these textual elements in the captions?

The rest of the paper is organized as follows. We first review related works on multimodal cognitive linguistics and the five perceptual properties regarding objects and scenes in Section 2. In Section 3, we illustrate the dataset and the annotation framework regarding the perceptual properties. Additionally, we conduct exploratory analysis (Section 4.1 and 4.2) and adopt statistical modeling (Section 4.3) on the perceptual properties and the textual elements. Finally, Section 5 concludes the paper.<sup>3</sup>

## 2 Related work

### 2.1 Theoretical framework on multimodal cognitive linguistics

The fundamental assumptions in Cognitive Linguistics are (i) language is an autonomous, self-contained system; (ii) the linguistic structure is usage-based; and (iii) grammar is inherently symbolic conceptualization (Croft and Cruse, 2004; Langacker, 2008; Hart and Marmol Queralto, 2021). These recognize that meaning construction can occur through various semiotic forms of expression within language usage. In other words, Cognitive Linguistics is "particularly well-equipped to unite the natural interest of linguistics in the units that define the language systems with the multimodality of language use" (Zima and Bergs, 2017).

Recently, Cognitive Linguistics has experienced a multimodal turn, focusing on the interplay between visual perception, linguistic expressions, and the collaborative impact on event conception. Hart and Marmol Queralto (2021) examine the phenomenon of *intersemiotic convergence*, which occurs when language and images share similar forms and create a cohesive relation. They also investigate how linguistic expressions and images converge to shape shared construal in conceptualization by exploring various dimensions<sup>4</sup>. Similarly, Divjak et al. (2020) employ a Visual World

Paradigm to study how alternative linguistic constructions (i.e., *location/preposition*, *voice*, and *dative*) modulate the distribution of attention and evoke different conceptualizations. These studies highlight the intricate connection between cognitive mechanisms and linguistic behaviors.

In terms of linguistic expressions, studies have demonstrated that word meanings are rooted in perception. This connection between language and perception has been extensively explored in Frame semantics (Fillmore et al., 1976) and Generative Lexicon Theory (GLT) (Pustejovsky, 1998), particularly in the context of human-object interaction (HOI) tasks, which serve as a solid foundation for addressing the research questions in our study.

In Frame semantics (Fillmore et al., 1976), the understanding of objects is based on accumulated experiences, represented as frames. Words are comprehended through the conceptual scenes (frames) they evoke. Building on this framework, Belcavello et al. (2020) apply fine-grained cognitive semantics in multimodal analysis using FrameNet to investigate how visual objects grounded in the aural modality create frames. Additionally, objects are contextualized to establish their habitat (Pustejovsky, 2013). The object's habitat, along with the verb's internal event structure, forms the event simulation. Krishnaswamy and Pustejovsky (2016) apply these insights to the HOI task, modeling events in a computational virtual environment, and suggest that incorporating affordance learning helps address challenges faced in the robot community.

### 2.2 Perceptual property

As previously discussed, the contextualization of objects is proposed to establish their habitat (Pustejovsky, 2013), influenced by various affordances in Gibson (1977)'s theory. *Affordance* refers to the actions enabled by an object for an agent, often termed as "action possibilities" within the surroundings. Henlein et al. (2023) distinguish affordance into two categories: *Gibsonian* and *telic*, presenting a model that better detects affordances for novel objects and actions. Gibsonian affordance denotes the "mere interaction with an object," for instance, a *cup* provides Gibsonian affordance for *carrying* or *holding*. On the other hand, *telic* affordance is related to an object's typical use or purpose in a scene, activating a conventionalized function for the agent (Pustejovsky, 2013). For example, in a kitchen, a *cup* naturally affords *telic* actions such

<sup>2</sup>The two types will be defined in Section 3.1.

<sup>3</sup>The dataset, annotation, and analysis in this study will be publicly available at <https://github.com/XXX>

<sup>4</sup>E.g., schematization, viewpoint, window of attention, and metaphor (see also Talmy, 2000; Forceville, 2008; Langacker, 2008; Hart, 2015; Hart and Marmol Queralto, 2021).

as *drinking, sipping, or pouring*.

In addition to *Affordance*, our study incorporates four other perceptual properties: *Perceptual Salience*, *Object Number*, *Gaze Cueing*, and *Ecological Niche Association (ENA)*. These properties are crucial for understanding the context of the scene and the affordances offered by the objects. A brief review of each property is provided below:

**Object Number** provides contextual clues for image interpretation; plural objects may have more than a mere cumulative effect (Link, 1983). The distinction between focused attention and global attention modes suggests different processing for singular and plural objects (Treisman, 2006).

**Gaze Cueing** is included since attentional connections can be established using visual signals, such as the speaker’s gesture or gaze, to emphasize information. (Enfield, 2009). In situations with deficient speech, the speaker’s gestures gain conversational value for the audience (Özer et al., 2023). This can also apply to image observation, where the gaze of agent(s) depicted in the image guides the attention of the image-viewer(s)<sup>5</sup>.

**Perceptual Salience** refers to how much attention a perceived object or event attracts, meaning certain features make an object stand out. In language, we often emphasize a specific part of a scene as the main focus (Talmy, 1983). This prominence can be conveyed in two ways: firstly, by plainly specifying emphasized semantic elements; and secondly, by inferring the primacy of an event’s participants from its internal semantic structure, even when not directly addressed (Langacker, 2008). For instance, both *he drives a car* and *he is driving* emphasize the salience of ‘car’, with the latter omitting the noun phrase.

**Ecological Niche Association (ENA)** introduced in this study refers to the conventionality of an object co-occurring with its environment, denoted as the "ecological niche association." This term captures the mutual dependence and co-adaptation between objects and their surroundings in specific ways. ENA expands the concept of *habitat* as a prerequisite (Pustejovsky, 2013) for actions, emphasizing the importance of context in shaping the meaning and function of objects. It highlights the

<sup>5</sup>To avoid confusion, we use the term ‘agent(s)’ to denote the agent(s) portrayed in the image. These agents are the ones interacting with the container-like objects in our analyses. In contrast, the term ‘viewer(s)’ is employed to refer to both (1) people who have observed the images and provided captions in the Flickr30k dataset and (2) our annotators who observe the images and annotate them.

dynamic relationship between objects and their contexts, enriching our understanding of object utilization and interpretation in natural language processing and other applications.

### 3 Methodology

#### 3.1 Data

This study focuses on exploring the association between grounding attributes in images and the conceptualized semantic representation in the corresponding captions. To achieve this, we manually select images featuring objects resembling containers from the Flickr30k dataset (Young et al., 2014)<sup>6</sup>. This selection is guided by our interest in object affordance. Each image from the dataset is paired with five captions, resulting in a total of 733 images and 3665 captions.<sup>7</sup>

Regarding the captions, they are lemmatized and POS-tagged via spaCy<sup>8</sup>. We define two categories of textual elements found in captions: ‘holding-verbs’ and ‘container-nouns’. The ‘holding-verbs’ includes motion verbs such as *hold, carry, grasp, grip, lift, grab, and take*<sup>9</sup>. These verbs generally indicate physical contact with hands, while the potentially subsequent actions of the agent may not be explicit.<sup>10</sup> On the other hand, the ‘container-nouns’ consist of specific nouns chosen based on the hyponyms of *container* in the English Wordnet<sup>11</sup>, including *cup, mug, beaker, goblet, chalice, teacup, container, bin, tin, glass, pan, pot, and bowl*.

Regarding the images, we have identified and reviewed five visuospatial properties related to how viewers perceive images (see Section 2.2), as we aim to investigate the associations between properties of groundness in images and conceptualized expressions in captions. The details of the properties

<sup>6</sup>The Flickr30k dataset, centering around humans engaging in everyday activities and events, consists of 31,783 images commonly employed for image captioning tasks.

<sup>7</sup>Referring to the comments from the reviewers, the captions, as discussed in (Young et al., 2014), are provided by five annotators who lack familiarity with the specific entities and situations depicted in each image.

<sup>8</sup><https://spacy.io/>

<sup>9</sup>To account for the productive verb *take*, as it also frequently occurs in phrases like *take a bite, take a break, take a sip*, and so on, we have used regular expressions to filter out irrelevant constructions.

<sup>10</sup>The holding-verbs are opposed to verbs like *drink* and *sip*, which imply purposeful actions and presuppose the act of holding a container. For example, when the verbs *drink* or *sip* are used in a caption, they implicitly involve a sequence of actions: *holding* the cup, *lifting* it to the mouth, and *drinking* the liquid.

<sup>11</sup><https://wordnet.princeton.edu/>

Table 1: Perceptual properties for annotation of images involving container-like objects.

Property	Variable	Description
Affordance	G / T	Whether the object shows Gibsonian affordance (G) or telic affordance (T).
Object Number	S / P	Whether the number of container-like objects is singular or plural.
Gaze Cueing	Yes / No	Whether the image-viewer accordingly follows the gaze attention of the agent(s) depicted in the image toward the ‘container-like object(s)’.
Perceptual Salience	1 (low) - 5 (high)	The degree to which an object or event captures attention, specifically referring to the features that cause an object to be visually distinctive.
ENA	1 (low) - 5 (high)	The degree of conventional interconnectedness and interdependence between an object and its surrounding environment (scene), describing how they relate and co-adapt to each other in specific ways.

*Note.* ENA: Ecological Niche Association.

will be illustrated with the annotation framework in Section 3.2, and the processing flow of the images and captions is displayed in Appendix A.

### 3.2 Annotation Framework

The annotation of the images with the five properties, as shown in Table 1, includes two types of variables: binary labels (as in *Affordance*, *Object Number*, and *Gaze Cueing*) and a rating scale from 1 to 5 (as in *Perceptual Salience* and *ENA*). For example, images are labeled as "T" in *Affordance* when the depicted actions between the agent and container are purposeful, intentional, and active. On the other hand, images are labeled as "G" when the agent and container lack clear, intentional actions. The labeling of *Gaze Cueing* depends on whether the annotator follows the gaze of the agent towards the container upon first viewing. The *ENA* property is based on the conventional relationship between the container and its surrounding context. For instance, an image of ‘wine glasses in a restaurant’ would receive a higher *ENA* rating than that of ‘wine glasses in a park’.

With a clear understanding of these perceptual properties and trial annotations, four linguists are asked to annotate the 733 images. Two linguists annotated the first half of the dataset (G1), while

Table 2: Inter-annotator agreement rate on the five perceptual properties within each group.

Property	Aff	ON	GC	PS	ENA
G1	.89	.92	.97	.89	.68
G2	.72	.83	.85	.40	.38
Avg.	.80	.88	.91	.65	.53

*Note.* The properties are shown in abbreviated form: Aff: *Affordance*, ON: *Object Number*, GC: *Gaze Cueing*, and PS: *Perceptual Salience*.

the other two annotated the second half (G2). After annotation, we have normalized the labels<sup>12</sup> and calculated the inter-annotator agreement rate for each group. Table 2 presents the statistics of inter-annotator agreement, showing high agreement for *Gaze Cueing*, *Object Number*, and *Affordance*. This indicates that the annotations are consistent and suitable for subsequent analysis. In cases where there were disagreements within a group (e.g., G1), we involve annotators from the other group (G2), who have not seen the images, to discuss and provide a third annotation to determine the final labeling for those images.

## 4 Discussion

### 4.1 Correlation between Perceptual Properties

Firstly, we investigate the correlation structures between the perceptual properties. In Figure 1, we compute a matrix of Pearson’s bivariate correlations for each pair of independent properties.<sup>13</sup> We follow Mason and Perreault Jr (1991)’s suggestion in using bivariate correlation 0.8-0.9 as the cutoff threshold, above which indicates strong linear associations and a linearity problem. With all the values being less than 0.8, we are more confident the multi-collinearity will not be a significant issue when interpreting each model predictor’s contribution.

Some observations in the correlation matrix are noteworthy. Figure 1 shows positive correlations between *ENA* and *Affordance* (corr = 0.4) as well as *ENA* and *Object Number* (corr = 0.38). The positive correlation between *ENA* and *Affordance* indicates that as the container(s) appear more conventional

<sup>12</sup>*Perceptual Salience* and *ENA* are rated on a 1-5 scale, so we group the original labels into three categories (i.e., < 3, 3, and > 3).

<sup>13</sup>In computing the correlation, the values of *Affordance* (i.e., "T" and "G") and *Object Number* (i.e., "P" (plural) and "S" (singular)) are both transformed to 1 and 0.

Figure 1: Correlation matrix between the independent properties.



in their natural environment (higher *ENA*), they are more likely to be linked with a telic affordance. This discovery aligns with the concept of *telic* affordance by Henlein et al. (2023), denoted as "a conventionalized configuration to activate a conventionalized function." Namely, a container with purposeful functions in a conventionalized scene is likely to be intentionally used.

A positive correlation between *ENA* and *Object Number* is also observed. In settings such as kitchens or grocery stores, multiple containers are considered a collective noun, which serves as a situational and conventional cue. They become part of the "conventionalized configuration" that activates a customary function (Henlein et al., 2023). For instance, in Figure 2a (see Appendix B), the numerous containers create the arranged setup, as "the products of the vendor," which prompts the agent to engage in "selling." This configuration of container(s) and functional arrangements leads to a higher *ENA* score.

## 4.2 Affordance & Distribution of Textual Elements

We delve into the distribution of 'holding-verbs' and 'container-nouns' concerning the research question: "Is the object *Affordance* in an image related to the presence of the two types of textual elements in its captions?" Our hypothesis proposes that object *Affordance* is associated with the textual elements in linguistic expressions, and that viewers are more likely to use holding-verbs/container-nouns when describing images involving objects with 'Gibsonian affordance'. As objects with 'telic affordance' imply more purposeful uses, they lead viewers to use more specific and informative language in conceptualization of such images.

For each image, we quantify the number of captions with holding-verbs and container-nouns separately. As we aim to explore the different *Affordance* of the container-like objects, we categorize the 733 images into two groups: T-group (346) and G-group (387) based on our annotations. "T" denotes the telic affordance, and "G" denotes the Gibsonian affordance.

Table 3: Distribution of textual elements for T-group and G-group images.

Textual Element	Cap <sub>N</sub> <sup>*</sup>	Image <sub>N</sub> (%) <sup>**</sup>	
		T-group	G-group
Holding-verbs	0/5	67.6	47.0
	1/5	21.1	26.4
	2/5	6.1	13.4
	3/5	2.9	10.6
	4/5	2.0	2.1
Container-nouns	0/5	86.1	78.3
	1/5	9.8	15.8
	2/5	2.6	4.1
	3/5	0.6	1.6
	4/5	0.3	0.3
5/5	0.6	0.0	

Note. <sup>\*</sup> Cap<sub>N</sub>: The number of captions containing the target textual elements, out of the five captions per image. <sup>\*\*</sup> Image<sub>N</sub>: The percentage of number of images; for example, 0.3% of T-group images and 0.5% of G-group images contain holding-verbs in all of their captions (5 out of 5 captions per image).

Table 3 shows interesting patterns in holding-verbs and container-nouns within T-group and G-group image captions. For T-group images, 67.6% (234) of them do not contain any holding-verbs in captions (i.e., 0 out of 5 captions per image), whereas a significant proportion (86.1%, 298) of them include no container-nouns in captions. Contrarily, approximately 47% (182) of G-group image lack holding-verbs in captions, and 78% (303) of them do not contain any container-nouns in captions. In general, G-group exhibits a higher proportion of images incorporating either holding-verbs or container-nouns in their captions.<sup>14</sup> This finding supports our hypothesis of a stronger association between objects with Gibsonian affordance in images and the occurrence of the two textual elements, in comparison to objects with telic affordance.

Besides *Affordance*, we believe that the other aforementioned perceptual properties, regarding human attention and dynamic relationship between

<sup>14</sup>Among the G-group images, 53% include holding-verbs in at least one caption, while the T-group exhibit a percentage of 32%. Also, 22% of G-group images contain container-nouns in at least one caption, while the T-group exhibit a lower percentage of 14%.

container(s) and scene in images, also play essential roles in conceptualized linguistic expressions (see Section 2.2). We attempt to adopt multiple linear regression models to evaluate the association between different perceptual properties and the usage of textual elements in image captions.

### 4.3 Statistical Modeling

We have employed two multiple linear regression models to investigate the question on "how the perceptual properties in an image contribute to the occurrence of textual elements within the captions." Both models include five independent variables (i.e., perceptual properties): *Affordance* (T/G), *Object Number* (Singular/Plural), *Gaze Cueing* (Yes/No), *Perceptual Saliency* (1-5), and *ENA* (1-5). The dependent variables are the numbers of captions with holding-verbs or container-nouns for each image, which are normalized to a range of 0 to 1. For example, if an image owns 3 out of 5 captions containing holding-verbs, the value for its holding-verb usage would be 0.6; if the image owns 2 out of 5 captions containing container-nouns, the value for its container-noun usage would be 0.4. The results of the models for holding-verbs and container-nouns will be discussed separately in Section 4.3.1 and 4.3.2, with an overview at the end of Section 4.3.2.

Table 4: Results of multiple linear regression model for holding-verbs.

Variable	Coeff	SE	t	P
(Intercept)	-0.01	0.035	-0.296	.768
<b>Perceptual Saliency</b>	0.063	0.007	8.886	<.001 ***
<b>Object Number_S</b>	0.057	0.015	3.733	<.001 ***
<b>ENA</b>	-0.021	0.007	-2.881	<.005 **
<b>Affordance_T</b>	-0.073	0.016	-4.608	<.001 ***
<b>Gaze Cueing</b>	-0.096	0.018	-5.414	<.001 ***

Note. *Affordance\_T*: *Affordance* labeled as T (telic). *Object Number\_S*: *Object Number* labeled as S (singular).

#### 4.3.1 Regression model for holding-verbs

The results of the model are presented in Table 4. Firstly, *Perceptual Saliency* shows a positive relationship with the holding-verbs usage (estimate: 0.0625,  $p < .001$ ). This suggests that when a container in an image is attention-grabbing to viewers, there is a higher likelihood for viewers to use holding-verbs in captions. For instance, a cup is at the center of image (A) in Table 5. While the cup is apparently not the female's attention nor the main theme of the image, it tends to be mentioned



in the captions. Among the five captions for (A), as shown in the upper row of Table 5, four of them contain the verb *hold*; specifically, the majority of the captions involve participle constructions (i.e., *holding a drink*; *holding a cup*; *holding a beverage*) to modify the agent (i.e., *the woman*; *the girl*).

Regarding *Object Number*, a significant positive correlation with holding-verbs usage is observed (estimate: 0.0574,  $p = .0002$ ). This suggests a strong tendency for viewers to use holding-verbs in captions when a singular container appears in an image. In contrast, *ENA* shows a negative relationship with holding-verbs usage (estimate: -0.0208,  $p = .0041$ ). This implies that holding-verbs are more likely to be employed by viewers when an image portrays a scene that is less conventional. For instance, Figure 2b (see Appendix B) presents a scene where the conventional function of the cup is not evident (i.e., low *ENA*). In this case, three out of five captions contain the act of *carrying a drink*, *holding a cup*, and *holding a coffee cup*, while the other two captions do not refer to the cup. On the other hand, Figure 2c (see Appendix B) displays a scenario where multiple containers are situated within a kitchen or party setting. It reflects higher conventionality of the containers co-occurring with their environment (i.e., high *ENA*), which aligns with Pustejovsky (2013)'s definition of a habitat as the precondition for an action involving the object. This activation prompts viewers to use verbs that directly describe the container's function, such as *drink*, *sip*, *pour*, *stir*, rather than holding-verbs.

In terms of *Affordance*, the coefficient for the *Affordance\_T* variable is negatively significant ( $p < .001$ ). This is because annotators assign the "T" label for *Affordance* when the depicted relations between the agent and container in images seem telic and purposeful. As these actions are explicit, it is reasonable for viewers to choose more specific verbs rather than less specific holding-verbs in their captions. Conversely, images are labeled as "G" when annotators perceive no clear intentional actions. As agent(s) of these images typically has mere contact with the container, i.e., "behaviors afforded due to the physical object structure" (Henlein et al., 2023), viewers tend to use more general verbs, the holding-verbs, to describe the relationship between the agent and the container.

As for *Gaze Cueing*, it shows a significant negative relationship with the use of holding-verbs (estimate: -0.0956,  $p < .001$ ). This suggests that

Table 5: Example images with captions: (A) above and (B) below

Example Image (A) & (B)	Captions
	<ol style="list-style-type: none"> <li>1. A blond woman in a short denim skirt, black top, and beige jacket, is reaching toward a part of a painting that is propped up on a windowsill.</li> <li>2. A woman is <b>holding a drink</b> in one hand and pointing at a painting with the other.</li> <li>3. Woman with a jean skirt <b>holding a drink</b> points to an object in a painting.</li> <li>4. A girl <b>holding an empty plastic cup</b> is pointing to a painting.</li> <li>5. A girl <b>holding a beverage</b> points at a painting.</li> </ol>
	<ol style="list-style-type: none"> <li>1. A man and a smiling woman sit at a dining table with many plastic <b>cups</b> on it as a person next to them eats out of a <b>bowl</b> with chopsticks.</li> <li>2. A group of people eat a meal in a crowded outdoor location.</li> <li>3. A group of people enjoy food and drinks at an outdoor party.</li> <li>4. A group of people eating and talking around a table.</li> <li>5. People are gathered at a table to enjoy drinks.</li> </ol>

when the agent of an image employs explicit *Gaze Cueing*, directing viewer’s attention, viewers are less likely to use holding-verbs in captions. In Figure 2d (see Appendix B), the agents look directly at the containers (i.e., *Gaze Cueing*: yes), showing intentional engagement with the containers; the captions for this image include telic verbs like *mix*, *pour*, and *perform*, rather than holding-verbs. Conversely, in Figure 2e (see Appendix B), the agents’ gaze is not at the container but at the screen (i.e., *Gaze Cueing*: no)<sup>15</sup>. In this case, viewers tend to use verbs related to ‘looking’ as the main action and use holding-verbs only to modify the agent (in relation to the container).

#### 4.3.2 Regression model for container-nouns

We also conducted multiple linear regression analysis to examine the usage of container-nouns. The same variables as in the model for holding-verbs were utilized, as displayed in Table 6.

For each unit increase in *Perceptual Salience*, there is a positive estimate of 0.1073 ( $p < .001$ ) in the number of captions containing container-nouns. This suggests that when a container in an image is more visually noticeable, viewers tend to use container-nouns more frequently in their captions. This aligns with our earlier discussion on holding-verbs. Even though the container is not the primary

<sup>15</sup>It is noted that *Gaze Cueing* in this study only represents the agent’s gaze attention toward “container-like object(s).” The agent’s gaze at other objects may be taken into account as another type of *Gaze Cueing* in future studies.

Table 6: Results of multiple linear regression model for container-nouns.

Variable	Coeff	SE	t	P
(Intercept)	-0.055	0.114	-0.482	.63
<b>Perceptual Salience</b>	0.107	0.023	4.701	<.001 ***
<b>Object Number_S</b>	0.182	0.05	3.649	<.001 ***
<b>ENA</b>	-0.061	0.023	-2.632	.008 **
<b>Affordance_T</b>	-0.047	0.052	-0.921	.358
<b>Gaze Cueing</b>	0.021	0.057	-0.374	.708

Note. *Affordance\_T*: *Affordance* labeled as T (telic). *Object Number\_S*: *Object Number* labeled as S (singular).

focus of the scene, its salience prompts viewers to include its description when conceptualizing the image. Consequently, container-nouns (e.g., *cup*) are used in participial phrases to modify the main agent/focus of the image, as in caption 4 (*A girl ‘holding an empty plastic cup’ is pointing to a painting.*) in the upper row of Table 5.

The *Object Number\_S* (singular) also demonstrates a significant positive relationship with the use of container-nouns (estimate: 0.1821,  $p < .001$ ), suggesting that when a solitary container is presented, viewers tend to use container-nouns more frequently in captions. This preference arises from the ability to concentrate attention on a singular object, leading to the expectation of more precise distinctions (Treisman, 2006). In contrast, if the number of container in an image is plural, the captions are less likely to include container-nouns. This can

be observed in image (B) in Table 5. In scenarios with an abundance of container, such as in a café or gathering, the individual significance and distinctiveness of containers decrease. Viewers tend to either concentrate on describing specific elements of the scene (e.g., agent(s) engaged in a purposeful action) or depict the scene as a whole. This can be seen in captions 2-5 in the second row of Table 5.

On the contrary, *ENA* shows a slightly significant negative relationship, with an estimate of  $-0.0616$  ( $p = .008$ ). This indicates that when an image depicts a less conventional scene, viewers tend to use container-nouns more frequently in captions. This observation is consistent with findings concerning *Object Number*. In Figure 2b (see Appendix B), where there is only one cup and a scene difficult for viewers to identify the conventional function (i.e., singular object & low *ENA*), the captions contain more phrases with container-nouns (e.g., *holding a cup*). In contrast, captions for Figure 2c contain fewer container-nouns as this image presents an accumulation of containers and a scene with higher conventionality that can be easily identified as a party (i.e., plural objects & high *ENA*). As for the other variables, *Affordance (T)* and *Gaze Cueing* did not exhibit statistical significance.

Table 7: Statistically significant factors for the presence of holding-verbs and container-nouns in captions.

<b>Holding-verbs</b>	
<i>Perceptual Salience</i>	The container is perceptually noticeable to viewer (high).
<i>Object Number</i>	The number of the container is singular (S).
<i>Gaze Cueing</i>	The agent does not employ explicit gaze cueing to the container (low).
<i>ENA</i>	The scene depicted in the image is less conventional (low).
<i>Affordance</i>	The object shows Gibsonian affordance (G).
<b>Container-nouns</b>	
<i>Perceptual Salience</i>	The container is perceptually noticeable (high).
<i>Object Number</i>	The number of the container is singular (S).
<i>ENA</i>	The scene depicted in the image is less conventional (low).

Table 7 presents a summary of significant factors in the two models, highlighting specific properties in images that prompt viewers to use these textual elements more frequently in captions. The results

strongly support our hypothesis, indicating a preference for holding-verbs in conceptualizing objects with Gibsonian affordance. When viewers observe an image depicting agent(s) and container(s), they determine if the container serves a purposeful function for the agent in such scene. If it does not, i.e., indicating Gibsonian affordance, viewers tend to use holding-verbs like *hold* or *take* to describe the container while modifying the agent (e.g., *girl holding a glass*). In terms of the other perceptual properties, *Perceptual Salience* and *Object Number\_S* exhibit significantly positive relationships with the usage of the two textual elements, while *ENA* shows less significant negative correlation with them; *Gaze Cueing* shows significant negative relationship only with the usage of holding-verbs. They facilitate the dynamic convergence between the container and its habitat (Pustejovsky, 2013) within the image, improve context comprehension, and contribute to the selection of linguistic expression. Overall, the analyses highlight the crucial role played by human cognitive mechanisms, object affordance, and contextual information in shaping shared construal by integrating visually-perceived events and text (Hart and Marmol Queraltó, 2021).

## 5 Conclusion

This study investigates the grounding issue in multimodal semantic representation, focusing on five perceptual properties in images and their associations with two types of textual elements in captions. Regarding *Affordance*, images featuring Gibsonian affordance show higher frequency of captions containing ‘holding-verbs’ and ‘container-nouns’ compared to images featuring telic affordance. The other properties, namely *Perceptual Salience*, *ENA*, *Gaze Cueing*, and *Object Number*, also play vital roles in shaping linguistic expressions of scenes. Our findings highlight the significance of situated meaning and object affordance in human conceptualization of visual input, transcending mere combination of text and other modalities. They offer insights for computational cognitive science, multimodal communication, and the contextually grounded AI models.

Despite limitations such as subjective selection of target images and the need for evaluations of provided captions and annotations, our study opens up possibilities for bidirectional tasks involving visual and textual elements for machines. Regarding future work, we plan to extend our research



to multimodal datasets that contain scenes where the visual cues and affordance are not as obvious or entirely absent, ensuring that the insights we've gained can be applied beyond images with clear affordance. To effectively handle situations where images lack evident affordance, we will explore the incorporation of additional contextual cues and the advanced deep learning techniques, which will help us bridge the gap between the visual characteristics of scenes and the language used to describe them in more intricate visual contexts. Overall, by integrating situatedness into multimodal semantics, we can improve our understanding of human interpretation in diverse real-world situations and facilitate further research on groundedness in natural language understanding systems.

## References

- Frederico Belcavello, Marcelo Viridiano, Alexandre Diniz da Costa, Ely Edison da Silva Matos, and Tiago Timponi Torrent. 2020. Frame-based annotation of multimodal corpora: Tracking (a) synchronies in meaning construction. In *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*, pages 23–30.
- Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W Black. 2021. Grounding 'grounding' in nlp. *arXiv preprint arXiv:2106.02192*.
- William Croft and D Alan Cruse. 2004. *Cognitive linguistics*. Cambridge University Press.
- Dagmar Divjak, Petar Milin, and Srđan Medimorec. 2020. Construal in language: A visual-world approach to the effects of linguistic alternations on event perception and conception. *Cognitive Linguistics*, 31(1):37–72.
- Nicholas J Enfield. 2009. *The anatomy of meaning: Speech, gesture, and composite utterances*. 8. Cambridge University Press.
- Charles J Fillmore et al. 1976. Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech*, volume 280, pages 20–32. New York.
- Charles Forceville. 2008. Metaphor in pictures and multimodal representations. *The Cambridge handbook of metaphor and thought*, pages 462–482.
- James J Gibson. 1977. The theory of affordances. *Hilldale, USA*, 1(2):67–82.
- Christopher Hart. 2015. [Viewpoint in linguistic discourse: Space and evaluation in news reports of political protests](#). *Critical Discourse Studies*, 12(3):238–260.
- Christopher Hart and Javier Marmol Queralto. 2021. What can cognitive linguistics tell us about language-image relations? a multidimensional approach to intersemiotic convergence in multimodal texts. *Cognitive Linguistics*, 32(4):529–562.
- Alexander Henlein, Anju Gopinath, Nikhil Krishnaswamy, Alexander Mehler, and James Pustejovsky. 2023. Grounding human-object interaction to affordance behavior in multimodal datasets. *Frontiers in Artificial Intelligence*, 6.
- Nikhil Krishnaswamy and James Pustejovsky. 2016. Voxsim: A visual platform for modeling motion language. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 54–58.
- Ronald W Langacker. 2008. Cognitive grammar. *Basic Readings*, 34:29.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.
- Godehard Link. 1983. The logical analysis of plurals and mass terms: A lattice-theoretical approach in r. bjuerle et al (eds.), meaning, use, and interpretation of language.
- Charlotte H Mason and William D Perreault Jr. 1991. Collinearity, power, and interpretation of multiple regression analysis. *Journal of marketing research*, 28(3):268–280.
- Demet Özer, Dilay Z Karadöller, Aslı Özyürek, and Tilbe Gökşun. 2023. Gestures cued by demonstratives in speech guide listeners' visual attention during spatial language comprehension. *Journal of Experimental Psychology: General*.
- James Pustejovsky. 1998. *The generative lexicon*. MIT press.
- James Pustejovsky. 2013. Dynamic event structure and habitat theory. In *Proceedings of the 6th International Conference on Generative Approaches to the Lexicon (GL2013)*, pages 1–10.
- 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.
- Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473.
- Leonard Talmy. 1983. *How language structures space*. CiteSeer.

Leonard Talmy. 2000. *Toward a cognitive semantics*. Cambridge, MA: MIT Press.

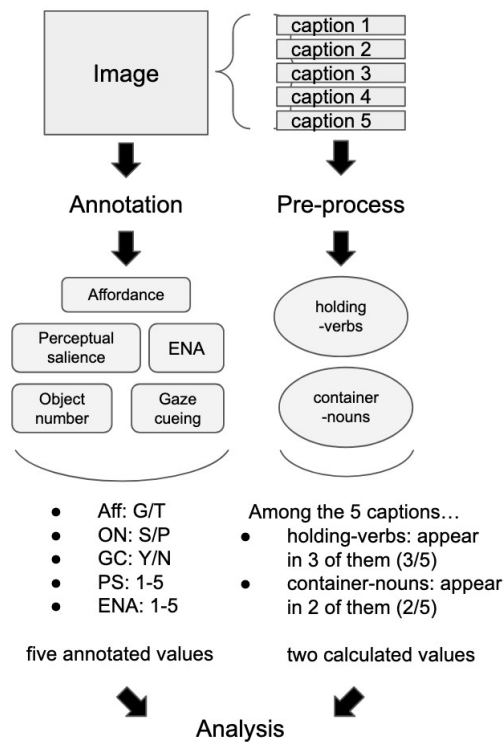
Anne Treisman. 2006. How the deployment of attention determines what we see. *Visual cognition*, 14(4-8):411–443.

Li Yang, Yan Xu, Chunfeng Yuan, Wei Liu, Bing Li, and Weiming Hu. 2022. Improving visual grounding with visual-linguistic verification and iterative reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9499–9508.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Elisabeth Zima and Alexander Bergs. 2017. Multimodality and construction grammar. *Linguistics Vanguard*, 3(s1).

## A Data Processing Flow



## B Example Pictures



(a) Correlation between Object Number and ENA.

(b) ENA: 1



(c) ENA: 5

(d) Gaze Cueing: Y



(e) Gaze Cueing: N

Figure 2: Example images. The value of *ENA* scales from 1 to 5; The value of *Gaze Cueing* is either Y (Yes) or N (No).