Towards an Exhaustive Evaluation of Vision-Language Foundation Models

Emmanuelle Salin, Stéphane Ayache, Benoit Favre

Towards an Exhaustive Evaluation of Vision-Language Foundation Models

Emmanuelle Salin Stéphane Ayache
Benoit Favre
Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
{emmanuelle.salin,sephane.ayache,benoit.favre}@lis-lab.fr

Abstract

Vision-language foundation models have had considerable increase in performances in the last few years. However, there is still a lack of comprehensive evaluation methods able to clearly explain their performances. We argue that a more systematic approach to foundation model evaluation would be beneficial to their use in real-world applications. In particular, we think that those models should be evaluated on a broad range of specific capabilities, in order to bring awareness to the width of their scope and their potential weaknesses. To that end, we propose a methodology to build a taxonomy of multimodal capabilities for vision-language foundation models. The proposed taxonomy is intended as a first step towards an exhaustive evaluation of vision-language foundation models.

1. Introduction

The development of foundation models in the last few years has enabled new state-of-the-art performances across a wide range of in the fields of computer vision and natural language processing tasks [73, 107]. Yet, monomodal models have shown to be limited in their ability to perform real-world tasks [4], as they are not sufficiently grounded in real-world experiences to be able to grasp multimodal concepts. Multimodality can be considered as an effective approach to ground models and reach a better understanding of human semantics. This has resulted in a growing focus on multimodal foundation models. In this paper, we specifically consider vision-language foundation models, which are use visual and textual inputs [92, 17, 49, 39, 97, 2, 53]. These models have been tested on a wide range of tasks, from image-to-text generation to cross-modal retrieval or classification. Yet, recent work has brought to light weaknesses in their understanding of multimodal concepts, i.e. concepts that cannot be captured by a single modality. For instance, vision-language models have a limited multimodal understanding of position [79, 80], vision-language compositionality [65] and word order [93], even though they are able to understand the basis of those concepts at a monomodal level [79, 80]. This has prompted the creation of dedicated evaluation tasks to assess those specific capabilities [108, 59]. Although benchmarks have also attempted to consider a wider spectrum of vision-language capabilities [67, 59], no attempt has been made to provide an exhaustive evaluation of those models.

Drawing inspiration from the work that has been accomplished in the evaluation of monomodal models, we aim at starting a discussion on the comprehensive evaluation of vision-language foundation models. Our goal is to reach a better explainability of the *capabilities* of foundation models. Other important aspects that should be taken into account when evaluating a foundation model, such as environmental and societal impact, are not the focus of this work. Foundation models are notoriously more difficult to evaluate than task-specific models. Indeed, the latter can be reliably evaluated on one specific task. Foundation models, on the other hand, are applicable to a wide range of tasks and domains. Thus, they must be evaluated on their whole scope of application. While researchers have developed benchmarks committed to a comprehensive evaluation of monomodal foundation models [96, 55, 110], to our knowledge, there has been no such proposal in the case of vision-language models. We argue that it is essential to assess the performance of multimodal vision-language foundation models on a wide range of specific capabilities. This would be the first step towards an exhaustive evaluation of such models. In this work, we propose an attempt at a taxonomy of vision-language capabilities. Figure 1 shows a summary of this taxonomy, presented in Section 4.

2. Evaluating Foundation Models

In this work, we consider vision-language foundation models. In [8], the authors propose a definition of those models: "A foundation model is any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks" That is to say, the goal of vision-language foundation models is to serve as the basis of multiple tasks

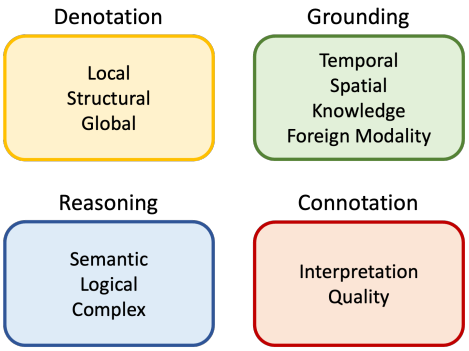

Figure 1. Summary of the proposed taxonomy

by learning general representations of texts or images on a large amount of data. The question of the evaluation of foundation models has still no answer. In particular, researchers can have different goals when evaluating a foundation model. One of those goals can be the comparison to human intelligence. In that respect, it is important to focus on its generalization ability and its capacity to solve previously unseen tasks [21]. Yet, the evaluation of a foundation model also aims to reach a better understanding of the precise capabilities of the model and its possible scope of operation. Indeed, foundation models are being used in real-world environments, where failures can have considerable consequences. Those are more likely to happen if users are unaware of the potential weaknesses of those models, or the extent of their reliability.

2.1. Monomodal Foundation Models

There have been standardization efforts in the evaluation of general-purpose models in Natural Language Processing and Computer Vision, following the development of multitask models. The fast development of language models has led to benchmarks designed to test the multitask abilities of those models. For instance, GLUE [96] and SuperGLUE [95] have gathered complex tasks to compare models to human performance. Similar benchmarks have been developed in Computer Vision. For instance, VTAB [110] aims to evaluate representation learning algorithms on a diverse range of 19 tasks (e.g. object counting, location recognition, fine-grained classification, disease classification) in diverse domains. However, these benchmarks offer limited insight on the explanation of a model's performance. To reach a better understanding of those black box models, new methods have been developed [78]. Among those methods, there has been an emergence of studies evaluating specific skills of models using probing tasks or evaluation methods [23, 74]. These have been established as a way to better understand what information is encoded in learned representations. Yet, probing tasks have also shown that they can lack in robustness, being highly dependent on syntactic variations [76]. This has led to the development of methods to stress test NLP models such as Checklist [77] or HELM [55] with regard to robustness, but also bias and fairness. Similar studies have also tested the robustness and bias of models learning visual representations [38, 100].

With the emergence of foundation models, the question of evaluation methods shifted from fine-tuning to zero-shot or few-shot evaluations on a wide range of tasks, which is less resource consuming. For instance, [99], the authors develop 1600 few-shot evaluation tasks for generative language models. While some studies focus on gathering numerous evaluation tasks [30], others have chosen to evaluate those models on human examinations rather than machine learning benchmarks [112]. For the visual modality, Florence [107] and CLIP [72] authors also use a wide range of visual and vision-language tasks and datasets to assess their models. Some methods tackle the evaluation problem from a capability-centric perspective [89], or attempt to build a taxonomy for the evaluation of language foundation models[55]. This enables a more precise explanation of the performances of foundation models. However, building a comprehensive evaluation benchmark is complicated, due to the variety of possible applications of foundation models. As a solution, authors rely on existing work in the field [55]. Thus, it is not aimed to be frozen but to evolve with the inclusion of new applications[89].

Other difficulties impact the evaluation of foundation models. First, the metrics used to evaluate those models are not always appropriate, especially in the case of generative models, either for texts [36] or for images [9]. The use of human evaluation enables researchers to avoid the flaws of existing metrics, but lack in standardization ability. In addition, the evaluation of foundation models relies on data dependent on bias and subjectivity [52]. The use of appropriate datasets and metrics to evaluate on a task as well as the development of exhaustive evaluation methods are both decisive to better diagnose and analyze foundation models.

2.2. Vision-Language Foundation Models

In the case of multimodal models, it can be difficult to clearly assess a model's understanding. Indeed, models rely on spurious correlations, and may exploit information from only one modality, without using crucial information from the other. This has been shown in vision-language models, where visual information can be ignored in favor of textual bias [33]. Therefore, to be able to trust a vision-language model's performance in a real-world application, it is important to be aware of what concept this model is able to understand at a multimodal level.

In recent years, several benchmarks have been developed [113, 11] to evaluate vision-language models. Some have also built tasks based on a multimodal phenomenon they want to assess, such as counting objects [68, 111]. On

the contrary, some works focus on the evaluation of models on tasks requiring complex reasoning abilities, such as generalization or abstraction [19]. Those methods give us an overview to compare the capabilities of vision-language models, and can point out their weaknesses.

However, as the field of vision-language multimodality is less mature than those of language only or vision-only machine learning, there is also a lack of hindsight on what issues vision-language foundation models will be facing. There are several aspects to consider in order to create a thorough overview of such a model: the understanding of each of the modalities, as well as the combination of monomodal information to understand multimodal concepts. To our knowledge, there has been no attempt at evaluating a broad coverage of vision-language capabilities.

3. Methodology

Through this work, we aim to discuss an exhaustive evaluation of vision-language foundation models, to help point out precise failures in the multimodal understanding of foundation models. With access to such information, users would be able to make an informed decision on the use of a model. In order to get a precise overview of the general multimodal understanding of a vision-language foundation model, we want to study its performances on a diverse set of multimodal capabilities. Such methods have indeed proven beneficial in natural language processing and computer vision to better understand the inner workings of large black-box models. Indeed, a more granular evaluation will help to point out limiting factors of vision-language models. Contrary to current works in the evaluation of foundation models in natural language processing, we do not focus on specific tasks (e.g. retrieval, inference, generation) but the capabilities required for multimodal understanding. To that end, we propose a taxonomy of vision-language capabilities. The goal of this taxonomy is to cover a broad range of vision-language capabilities. Indeed, the capabilities used to evaluate foundation models should be as complete as possible to avoid blind spots. In this section, we explain the categorization of vision-language capabilities into the taxonomy, as well as how to determine granular vision-language capabilities relevant in real-world applications.

3.1. Categorization

Indeed, multiple types of broad abilities are required when a foundation model performs a vision-language task. The categorization of granular vision-language capabilities into those broad abilities can help identify potential blind spots. To organize those abilities, we draw a parallel with the human understanding. Indeed, we refer to the field of visual literacy, which studies the human understanding of images, to help us establish different stages of visual literacy for machine learning systems. There is no clear definition of what it means to be visually literate, due to the complex nature of the concept [48]. Visual literacy is defined by aggregating sets of skills in two main categories: 'denotation' and 'connotation' [3]. *Denotation* refers to the perception of visual elements in an image, while *Connotation* associates the image with an ideological or affective meaning. However, those specific abilities are not sufficient to evaluate the capabilities of a model, which may fail at skills considered basic for a human. As a result, we propose four broad categories of vision-language capabilities, with the following definitions. The first letters of those categories will be used to refer to them in the next section.

Definition 1 (Grounding G)**.** *Capabilities requiring the use of information that is not directly accessible using the given inputs (2D image and text) or the understanding of concepts that cannot be described using those modalities (e.g. time, space, knowledge, sound, mathematical documents).*

Definition 2 (Reasoning R)**.** *Capabilities requiring the application of abstract thinking or logic to the analysis of an image-text instance.*

Definition 3 (Connotation C)**.** *Capabilities related to the subjective analysis of a text-image instance, from symbolic interpretation to qualitative evaluation.*

Definition 4 (Denotation D)**.** *Text explicitly depicts or refers to image elements and does not require grounding, reasoning or evokes connotation.*

3.2. Determining vision-language capabilities

In order to build this taxonomy, we must consider the context in which it operates, meaning the current state of the vision-language field. Indeed, the evaluation of vision-language foundation models should be to be appropriate, considering the use-cases and challenges of vision-language models. By precisely analyzing the context, we can identify relevant vision-language capabilities at a granular level. We are inspired by HELM [55], which uses conference tracks to assess the coverage of their evaluation of language models. However, the field of vision-language machine learning is less mature than Natural Language Processing, and not all challenges have been clearly identified.

Since foundation models are aimed at real-world applications, we select some that could be a use-case for vision-language models from current research. There is a growing number of complex applications, with common challenges that have not yet been resolved, as detailed in in Appendix A. A foundation model would have to be evaluated on challenges linked to those various applications. We argue that those challenges should be tackled as a common goal, and that it should reflect in the evaluation of those models. However, the complex nature of those applications may make it difficult to interpret the performance of a model. To that

end, we encourage the evaluation of foundation models to go from a task-centric perspective to a capability-centric perspective, by creating a list of vision-language capabilities needed for real-world applications.

In this section, we study more precisely several of the identified real-world applications to get as complete a picture as possible of the capabilities involved in those tasks: news captioning, medical visual question answering (VQA) [1] and vision-language navigation [85] to determine associated vision-language capabilities. As observed previously, those applications do not cover the whole range of vision-language multimodality, but they offer insight into different capabilities relevant to multimodality. For each of those applications, we proceed with a specific method to identify related vision-language capabilities. These methods could then be applied to other vision-language applications to identify capabilities.

Manually studying relevant data Vision-language foundation models can be used with news-related data for fake-news detection algorithms. We study the capabilities necessary for such applications from a data-centric perspective: we collect examples and manually identify relevant capabilities. News-related data varies across cultures, periods, as well as topic of interest. We choose to study examples from selected newspapers to extract different types of multimodal interaction, as well as capabilities needed for a vision-language system to correctly understand those examples. More details are available in the Appendix B. We notice that news images and their captions follow two main different types: either the image is described by the caption, with possibly a bit of context added by the text, or the image is used as an illustration of the text, and the link between text and image is less direct. Following the vocabulary introduced by [66], we call the first text-image relationship *anchorage* and the second situation *illustration*. The studied examples are evenly split along those two categories. From the studied examples, we extract several capabilities necessary for a good understanding of the instances:

- Object Recognition D: Understand the content of an instance. For instance, in the case of war reporting, it is important to differentiate between systems belonging to two different armies.
- Text Understanding G-R: Understand written text in an image, and its role with respect to the object it is written on. For instance, text written on a protest board or a shop window have widely different intents.
- Named Entity Recognition G: Link famous people or monuments in an image to the corresponding entity.
- Semantic Role Understanding G: Understand the role of both objects and people. For instance, understanding the job of someone with respect to the context.
- Sentiment Understanding D-G: Understand the stance, gaze, expressions and interaction of a person (or animal) with their environment.
- Structural Understanding D: This can relate to the understanding of the structure of an image (e.g. counting, understanding position). For instance, it can help understand how each part of the instance relates to each other (e.g. interaction between people).
- Context Grounding G: Identify when the picture was taken, where it was taken, or the event it depicts.
- Image Interpretation C: Some instances show a discrepancy between text and image, which can help understand the intent of the journalists. For instance, the use of the words 'is investigated' in a caption gives a new meaning to a picture.
- Style understanding C: This can relate to the understanding of art or style, and the understanding of iconography.

Relying on existing datasets Vision-language foundation models can be used as part of multiple real-world applications, as detailed in Appendix A. Those applications often require specific technical knowledge to understand the underlying challenges. To compensate for our lack of technical knowledge, we can rely on already existing tasks and datasets to identify relevant capabilities. In this section, we specifically study Computer-Aided Diagnosis systems as an example. These systems can provide doctors with another tool to reach a medical diagnosis or help communication. Some datasets have already identified relevant problems of vision-language multimodality applied to medical data. To that end, we refer to the question types identified in medical VQA tasks [1].

- Data Collection Context D: In medical imaging, data can vary following what is being observed, with what machine, options, at what angle.
- Object Recognition D: Recognize different organs or body parts, as well as to be able to segment them.
- Semantic Object Understanding G-R: Differentiate between 'normal' or 'abnormal' organs.
- Focus Understanding D: Understand the main 'abnormality' in an image, which requires the system to understand the focus of a medical instance.
- Knowledge Grounding G: Medical technical knowledge is necessary to correctly describe and differentiate technical terms.

- Logical Reasoning R: The system may need to perform logical reasoning to aggregate multiple factors.
- Multi-source understanding D-R: Summarize and compare several sources of data.

Relying on extensive research in a field Vision-language foundation models can be used to build agents that can interact with their environment using human language and visual information. This field is referred to as vision-language navigation (VLN). To identify relevant vision-language capabilities, we rely on extensive research consisting of various datasets and models [35] that have studied the challenges and problems related to this field. To be able to perform VLN, a system must have a good understanding of:

- Spatial Understanding D-G: Understand the position of an agent relative to other objects in the scene, as well as the depth and size of other objects. This skill depends on the point of view of the system.
- Space-based Reasoning R: The ability to design a path based on available information.
- Object Recognition D: Recognize objects in the scene.
- Object Role Understanding G: A model should be able to recognize the role objects, as well as their associated physics. In particular, some objects can be obstacles, and others can be interacted with.
- Object State Understanding G: Recognize the state objects, and the semantic change in those states. For instance, a cup can be empty or full and will not have the same role depending on its state.
- Action Understanding G-R: Understand the sequence of actions necessary for a task, and their effect on the environment. For instance, washing something implies changing the state of an object from 'dirty' to 'clean'.
- Structure Understanding D: Recognize the structure of a scene, as well as the dependency between objects.
- Intent Understanding C: Understand the intent, even in the case of a misalignment between modalities. The model must be able to correctly understand the intent despite this discrepancy.

Discussion In this section, we study a few diverse applications of vision language systems to determine a set of skills necessary for vision-language systems. In addition to downstream applications, we also rely on previous works in the fields of computer vision and natural language processing [60, 56, 10, 105, 13] to identify relevant capabilities to complete the taxonomy. Due to the breadth of the vision-language field, it is difficult to enumerate all possible vision-language capabilities. To further this study, several other applications (Appendix A) could help provide a more complete understanding of vision-language skills. Before using a vision-language foundation model on a specific application, we encourage studying the task to uncover relevant vision-language capabilities.

4. Taxonomy

In this section, we propose a preliminary attempt at a taxonomy of vision-language capabilities. We supplement the previously determined capabilities (Section 3.2) using previous work in natural language processing, computer vision and cognitive sciences to build a taxonomy of vision-language capabilities. The taxonomy is presented in more details in Appendix C.

Denotation The capabilities of a vision-language model to explicitly associate a text and an image are conditioned on its ability to take into account information at different levels. At a **local** level, denotation capabilities evaluate the understanding of a single element of a text-image instance, independently of the rest of the instance. Among the previously determined capabilities, the ability to recognize objects is such an ability. A parallel can be made with the Communicative Development Inventories (CDIs) [28], where recognizing objects such as animals or vehicles is among the first skills evaluated for children. Several datasets have focused on the evaluation of the presence of objects [82, 67]. A related category that appears in CDIs but not in previously determined capabilities is the understanding of descriptive words (e.g. 'dark', 'blue'). We infer from it the capability to detect basic descriptive attributes, which is often included as part of more complex tasks [43, 44].

At a **structural** level, denotation capabilities evaluate the understanding of the dependency between an element and the rest of the instance, or between several elements of an instance, i.e. the compositionality of an instance. As a whole, those skills also require local understanding, because the model needs to also understand each element individually. Although we have identified in the previous section the need for structural understanding of an instance, we specify here more granular capabilities using as basis previous work in the field of vision-language multimodality. As the structure of text and that of an image are radically different, we first consider the understanding of the two structures individually: scene understanding and syntactic understanding. Scene understanding, which also groups positional understanding and counting, is an active field of research in vision-language multimodality [43, 67, 80]. Similarly, the multimodal understanding of syntax remains part of ongoing research, as works have shown the difficulty of

vision-language models to understand word order at a multimodal level [94]. In addition, understanding the multimodal alignment between elements of the instance is also important, through the understanding of multimodal dependencies [65] and coreferences [18].

At a **global** level, denotation capabilities evaluate the understanding of the whole instance. Two main capabilities determined in the previous section correspond to this category: the ability to understand document type (e.g. the context behind the data collection) or the focus. However, to our knowledge, beside domain-specific datasets, no multimodal dataset evaluates these precise capabilities. Denotation skills characterize factual understanding of a vision-language instance and its components. We listed in this section several skills that, to our knowledge, are necessary to establish this understanding of a vision-language instance. This list does not include the ability to ground the instance in the world or use knowledge specific to a domain.

Grounding First, **temporal** grounding capabilities evaluate a model's ability to understand the situation of an instance in time. The ability of action understanding, context understanding and object state understanding described in the previous section are related capabilities. Several datasets already evaluate the grounding in time of a model, through tasks such as event captioning or procedural understanding [51, 104], but not all capabilities are covered.

Then, **spatial** grounding capabilities evaluate a model's ability to understand a scene as part of a wider spatial context. Among the applications studied in the previous section, it is especially useful in Vision-Language Navigation, but also in context understanding. Several datasets and tasks focus on spatial grounding capabilities, mainly relating to 3D understanding [32, 22, 15, 50].

In addition, technical or cultural **knowledge** can be necessary to understand a vision-language instance. This can be relevant to context understanding in news data or to the understanding of specific knowledge in medical data. In the case of technical grounding, evaluations specific to the domain are necessary [37, 98, 5, 75].

Finally, vision-language models can also be evaluated on their understanding of other **foreign modalities** not present in the instance. For instance, they can be used in applications which refer to time series, such as financial data understanding. In this case, evaluation tasks for those capabilities are very specific and depend on the domain. The understanding of temporality, as well as other forms of grounding, is complex, and requires precise data to be appropriately evaluated. If a vision-language model is destined at being used in this context, evaluating it on more granular skills (described in C.2) can be necessary to understand weaknesses.

Reasoning We identify a few reasoning tasks necessary for vision-language models, using as inspiration existing monomodal tasks [60, 56, 10, 105, 13]. First, some reasoning capabilities can require a good understanding of **semantic** knowledge, which can be useful in applications requiring some kind of technical knowledge such as medical assisted diagnosis. We can for instance list the detection of abnormality. However, there is to our knowledge no specific dataset evaluating multimodal knowledge-based reasoning.

Then, reasoning skills can be based on **logic**, or the understanding of mathematical concepts. Several evaluation tasks have focused on logical and mathematical reasoning [19], as such tests are generally used as a metric to measure human intelligence. Another kind of skills linked to logical reasoning are those based on comparison between instances. Those are well known in the field of natural language processing, being evaluated through tasks such as natural language inference[27].

Finally, some reasoning capabilities are more **complex**, due to the use of abstraction or several steps of reasoning. For instance, this is the case of multi-hop reasoning that can be encountered in vision-language navigation. As such tasks are complex and specific, they are mostly evaluated with respect to the relevant application domain. We also group in this subcategory the ability to perform introspection, i.e. to explain the reasoning of a prediction, which is an active field of research [45, 109, 24]. These reasoning capabilities can be complemented by other monomodal capabilities transferred to multimodality.

Connotation The skills listed in this section may not be useful to all applications of vision-language models, as they rely on individual interpretation of multimodal instances. In addition, their evaluation is subjective and can widely vary depending on the annotations. The connotation capabilities can evaluate a model's ability to **interpret** the meaning or intent of an instance. In particular, this relates to the previously identified capability of intent understanding. Some related evaluation tasks interpret the emotion [64] or the style techniques [69].

In addition to interpretation, connotation capabilities can also relate to the evaluation of the **quality** of an instance. These are mostly evaluated using user judgment, and evaluate stylistic appreciation [71, 91]. In the connotation category, we also list several capabilities for which we have found no related evaluation tasks in Appendix C.4. Those are inspired from human evaluation methods of visual literacy, which often rely on interpretation and assessment of instances. These skills can be used in real-world applications where the interpretation of an instance is important, such as applications related to art.

Category	Subtype	Datasets	Task description
Denotation D	Local	GQA[40], Foil it![82], TDIUC[44], VQA[34], VALSE[67], Toolbox[111]	Object and attribute recognition
	Structural	GQA[40], Daquar[61], CLEVR[43], TDIUC[44], Probing[81], VALSE[67], Toolbox[111]	Position understanding and counting
		Winoground [94]	Understanding word order
		Noun-Predicate Dep [65], Abstract Semantics [114], CREPE [59], ARO [108]	Understanding compositionality
		Cops-ref [18], RefCOCO [46], CLEVRRef [57], VALSE [67]	Multimodal referring expressions
Grounding G	Temporal	Dense Event Captioning[51], RecipeQA [104]	Event and procedure understanding
	Spatial	IQUAD [32], VQA360 [22], Matterport3D [15], AI2-THOR [50], RemoteSensing [58]	Spatial understanding (3D & aerial)
	Knowledge	OK-VQA [63], TDIUC	Object role understanding
		TextVQA [88], SceneText VQA [6], TextCaps [87]	Optical character recognition
		OK-VQA [63]	VQA with cultural knowledge
		GoodNews [5], BreakingNews [75]	News-related tasks with NER
		PathVQA [37], Chest Xrays [98]	Medical tasks
Reasoning R	Logical	E-SNLI-VE [27], NLVR2 [90]	Multimodal inference and comparison
		SMART [19]	Logical and mathematical reasoning
	Complex	E-vil [45], VCR [109], VQA-HAT [24]	Explanations for VQA
		Visual Dialog [25], FashionIQ [103], GuessWhat?! [26]	Dialog with multimodal context
Connotation C	Interpretation	AVA [69]	Image style understanding
		SentiCaps [64]	Caption generation with sentiments
	Quality	New Yorker Caption Contest [71], ICQD [91]	Rating Caption quality
		DPC [41], VizWizQuality [20], AVA [69], Aesthetic Cap[31], VILA [47]	Image Quality Evaluation

Table 1. Projection of a range of existing vision-language evaluation tasks in the proposed taxonomy

5. Evaluating Foundation Models

The taxonomy presented in the previous section aims at providing a guideline for an extensive evaluation of vision-language foundation models, taking into account their real-world applications. To that end, we argue that foundation models should be evaluated on granular capabilities, more easily interpretable than complex tasks. These capabilities should have the broadest possible coverage, and be useful in real-world applications. Indeed, it is important to be aware of the main weaknesses of a foundation model, as well as the scope of tasks and datasets it can be applied to. In Table 1, we give a projection of vision-language evaluation tasks in our proposed taxonomy. Depending on the application and domain of a vision-language foundation model, it can be unnecessary to evaluate it on every possible capability, and all capabilities may not have the same usefulness. For instance, a foundation model geared towards medical assisted diagnosis would have no use for connotation capabilities. Moreover, the goal of this taxonomy is not to help compute a ranking score from an aggregation of tasks, but to bring back the focus on multimodal understanding capabilities relevant to real-world vision-language applications. The use of several pre-defined tasks for the evaluation of vision-language foundation models may encourage a specific focus on raising the performance on those tasks, but they should be used as an introspective evaluation to establish a diagnosis of the performance of a foundation model. In addition, the datasets presented in Table 1 may always be appropriate for the multimodal evaluation of models. Indeed, among the existing evaluation tasks for vision-language models, some of them evaluate an aggregate of complex skills more or less directly linked to a specific capability. They may not be granular enough to identify potential blind spots. Another aspect is that they may not truly

evaluate multimodal understanding. Indeed, some of those tasks present considerable textual bias, which hampers the multimodal evaluation of those models. For instance, a language model can reach good performances on 'Foil it!' [83] [67]. In other cases, the task itself may not be built with multimodality in mind. This is the case for datasets of the connotation category, where the evaluation of instance quality can often be associated to a vision-only task. The difference between monomodal and multimodal capabilities can be blurry, as shown by the use of vision-language models to perform vision-only tasks [72]. This is why some of the capabilities we present in this taxonomy may belong to both multimodal and monomodal understanding.

6. Limits of the current taxonomy

This taxonomy is aimed at guiding the evaluation of foundation models for real-world applications. However, the use of such a taxonomy also presents its own limitations. First, it may not reflect the possible applications of vision-language foundation models, and may be more specifically biased towards already existing tasks. Indeed, capabilities were selected from a range of English language vision-language applications, which may hide challenges or needs more present in other languages or cultures. In Table 1, we give an overview of vision-language evaluation tasks related to the categories listed in the taxonomy. These evaluation tasks are not evenly distributed through the categories, and this taxonomy can help us identify potential gaps in the evaluation of vision-language models. These gaps can be due to the lack of interest, available data or known research challenges, but still hide potential blind spots of those models. This taxonomy is not final, but the gaps can also be used to guide the way towards other evaluation tasks relevant for vision-language applications. The taxonomy we presented in this section establishes a set of skills relating to vision-language multimodal understanding. However, evaluation tasks for foundation models may not necessarily clearly fit into this taxonomy. Indeed, there can be overlap in the skills that different tasks evaluated. In addition, more complex skills are built on simpler skills. For instance, most reasoning skills require first an understanding of denotation skills. As a result, this taxonomy is not intended to be complete, but a first step towards building a more comprehensive evaluation of multimodal foundation models.

Although we focus in this paper on *capabilities* of vision-language models, other factors are to take into account to provide a comprehensive evaluation of a foundation model. In particular, a foundation model should have a good ability to generalize to unseen examples from different domains. This diversity could be ensured by selecting instances from a broad range of semantic categories. For instance, vocabulary from Communicative Development Inventories for various cultures [29] can be used to ensure diversity, as well as images from diversified sources. In addition, we do not mention limiting bias and ensuring fairness and robustness, which are major aspects of foundation models evaluation, and should be taken into account when building evaluation tasks and datasets. In this taxonomy, we also do not take into account the type of task (e.g. generation, classification), but the evaluation of a multimodal capability can vary depending on the type of task used. As this taxonomy is based on a sample of tasks that is not necessarily representative of the whole range of possible vision-language applications, it is also not yet complete. It is intended to evolve, and in particular to be more specified, for instance regarding the various uses of a foundation model.

Evaluating the taxonomy An important question is how to evaluate such a taxonomy, in particular in terms of the coverage of all capabilities. Indeed, it is difficult to be both granular and exhaustive. One way could be to study a range of tasks from Appendix A in the same way as Section 3.2 to ensure a coverage of necessary capabilities. This is particularly difficult to assess, as it depends on how models are used in downstream applications. This taxonomy is incomplete, and is aimed at evolving with the improvement of vision-language foundation models and the creation of new applications, which would lead to new challenges. In addition, evaluating a model on the whole taxonomy is time- and resource-consuming, this is why our goal in presenting this taxonomy is above all to serve as a guideline.

7. Conclusion

Foundation models are notoriously difficult to evaluate. To our knowledge, no exhaustive evaluation method of vision-language foundation models has been developed yet. In this work, we argue that such a method should aim at evaluating a wide range of precise multimodal capabilities, in order to better apprehend the possible weaknesses of such models, rather than evaluating a few complex tasks that may be more difficult to analyze and diagnose. To that end, we propose a methodology to build a taxonomy of vision-language capabilities. We rely on existing vision-language tasks to establish vision-language capabilities useful for vision-language applications. We also relate this taxonomy to existing evaluation tasks. The goal of such a taxonomy is to establish a comprehensive evaluation method of vision-language foundation models. Thus, it would help highlight potential weaknesses of those models that may impact their performances in real-world applications. However, the use of such a taxonomy also presents its own limitations, in particular due to potential bias in determining useful capabilities. In the future, it would be interesting to further strengthen this taxonomy with additional perspectives, as well as to further complete its coverage of vision-language real-world applications.

8. Acknowledgments

The project leading to this publication has received funding from Excellence Initiative of Aix-Marseille - A*MIDEX (Archimedes Institute AMX-19-IET-009), a French "Investissements d'Avenir" Programme.

References

[1] Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. *CLEF (working notes)*, 2(6), 2019. 4

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 1

[3] Laurence Bardin. Le texte et l'image. *Communication & Langages*, 26(1):98–112, 1975. 3

[4] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA, Mar. 2021. Association for Computing Machinery. 1

[5] Ali Furkan Biten, Lluis Gomez, Marçal Rusinol, and Dimosthenis Karatzas. Good news, everyone! context driven entity-aware captioning for news images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12466–12475, 2019. 6, 7

[6] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301, 2019. 7

[7] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer, 2022. 14

[8] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 1

[9] Ali Borji. Pros and cons of gan evaluation measures. *Computer Vision and Image Understanding*, 179:41–65, 2019. 2

[10] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015. 5, 6, 16

[11] Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. Iglue: A benchmark for transfer learning across modalities, tasks, and languages. In *International Conference on Machine Learning*, pages 2370–2392. PMLR, 2022. 2

[12] Andrea Burns, Deniz Arsan, Sanjna Agrawal, Ranjitha Kumar, Kate Saenko, and Bryan A Plummer. A dataset for interactive vision-language navigation with unknown command feasibility. In *European Conference on Computer Vision*, pages 312–328. Springer, 2022. 14

[13] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31, 2018. 5, 6, 16

[14] Yiyi Cao, Lei Chen, Yuan Yuan, and Guangling Sun. Cucumber disease recognition with small samples using image-text-label-based multi-modal language model. *Computers and Electronics in Agriculture*, 211:107993, 2023. 14

[15] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 6, 7

[16] Yuyang Chen and Feng Pan. Multimodal detection of hateful memes by applying a vision-language pre-training model. *Plos one*, 17(9):e0274300, 2022. 14

[17] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. 2019. 1

[18] Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee K. Wong, and Qi Wu. Cops-Ref: A New Dataset and Task on Compositional Referring Expression Comprehension. pages 10086–10095, 2020. 6, 7

[19] Anoop Cherian, Kuan-Chuan Peng, Suhas Lohit, Kevin A Smith, and Joshua B Tenenbaum. Are deep neural networks smarter than second graders? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10834–10844, 2023. 3, 6, 7

[20] Tai-Yin Chiu, Yinan Zhao, and Danna Gurari. Assessing image quality issues for real-world problems. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3646–3656, 2020. 7

[21] François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019. 2

[22] Shih-Han Chou, Wei-Lun Chao, Wei-Sheng Lai, Min Sun, and Ming-Hsuan Yang. Visual question answering on 360

images. In *2020 IEEE winter conference on applications of computer vision (WACV)*, pages 1596–1605. IEEE, 2020. 6, 7

[23] Alexis Conneau and Douwe Kiela. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*, 2018. 2

[24] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, 2017. 6, 7

[25] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335, 2017. 7, 14

[26] Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512, 2017. 7

[27] Virginie Do, Oana-Maria Camburu, Zeynep Akata, and Thomas Lukasiewicz. e-snli-ve: Corrected visual-textual entailment with natural language explanations. *arXiv preprint arXiv:2004.03744*, 2020. 6, 7

[28] Larry Fenson et al. Macarthur-bates communicative development inventories. 2007. 5

[29] Michael C Frank, Mika Braginsky, Daniel Yurovsky, and Virginia A Marchman. Wordbank: An open repository for developmental vocabulary data. *Journal of child language*, 44(3):677–694, 2017. 8

[30] Leo Gao, Jonathan Tow, Stella Biderman, Charles Lovering, Jason Phang, Anish Thite, Fazz, Niklas Muennighoff, Thomas Wang, sdtblck, tttyuntian, researcher2, Zdeněk Kasner, Khalid Almubarak, Jeffrey Hsu, Pawan Sasanka Ammanamanchi, Dirk Groeneveld, Eric Tang, Charles Foster, kkawamu1, xagi dev, uyhcire, Andy Zou, Ben Wang, Jordan Clive, igor0, Kevin Wang, Nicholas Kross, Fabrizio Milo, and silentv0x. Eleutherai/lm-evaluation-harness: v0.3.0, Dec. 2022. 2

[31] Koustav Ghosal, Aakanksha Rana, and Aljosa Smolic. Aesthetic image captioning from weakly-labelled photographs. *arXiv preprint arXiv:1908.11310*, 2019. 7

[32] Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. Iqa: Visual question answering in interactive environments. In *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*. IEEE, 2018. 6, 7

[33] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 2

[34] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. pages 6904–6913, 2017. 7

[35] Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Eric Wang. Vision-and-language navigation: A survey of tasks, methods, and future directions. *arXiv preprint arXiv:2203.12667*, 2022. 5

[36] Tianxing He, Jingyu Zhang, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James Glass, and Yulia Tsvetkov. On the blind spots of model-based evaluation metrics for text generation. *arXiv preprint arXiv:2212.10020*, 2022. 2

[37] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020. 6, 7

[38] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 2

[39] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023. 1

[40] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 7

[41] Xin Jin, Le Wu, Geng Zhao, Xiaodong Li, Xiaokun Zhang, Shiming Ge, Dongqing Zou, Bin Zhou, and Xinghui Zhou. Aesthetic attributes assessment of images. In *Proceedings of the 27th ACM international conference on multimedia*, pages 311–319, 2019. 7

[42] Quanliang Jing, Di Yao, Xinxin Fan, Baoli Wang, Haining Tan, Xiangpeng Bu, and Jingping Bi. Transfake: multi-task transformer for multimodal enhanced fake news detection. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021. 14

[43] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 5, 7

[44] Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *ICCV*, 2017. 5, 7

[45] Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1244–1254, 2021. 6, 7

[46] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 7

[47] Junjie Ke, Keren Ye, Jiahui Yu, Yonghui Wu, Peyman Milanfar, and Feng Yang. Vila: Learning image aesthetics from user comments with vision-language pretraining. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10041–10051, 2023. 7

[48] Joanna Kedra. What does it mean to be visually literate? examination of visual literacy definitions in a context of higher education. *Journal of Visual Literacy*, 37(2):67–84, 2018. 3

[49] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 1

[50] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, Aniruddha Kembhavi, Abhinav Kumar Gupta, and Ali Farhadi. Ai2-thor: An interactive 3d environment for visual ai. *ArXiv*, abs/1712.05474, 2017. 6, 7

[51] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *International Conference on Computer Vision (ICCV)*, 2017. 6, 7

[52] Bart Lamiroy. Interpretation, evaluation and the semantic gap... what if we were on a side-track? In *Graphics Recognition. Current Trends and Challenges: 10th International Workshop, GREC 2013, Bethlehem, PA, USA, August 20-21, 2013, Revised Selected Papers 10*, pages 221–233. Springer, 2014. 2

[53] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1

[54] Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. Clip-event: Connecting text and images with event structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16420–16429, June 2022. 14

[55] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022. 1, 2, 3

[56] Xi Victoria Lin, Richard Socher, and Caiming Xiong. Multi-hop knowledge graph reasoning with reward shaping. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3243–3253, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. 5, 6, 16

[57] Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L Yuille. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4185–4194, 2019. 7

[58] Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195. 7

[59] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921, 2023. 1, 7

[60] Bill MacCartney. *Natural language inference*. Stanford University, 2009. 5, 6, 16

[61] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1682–1690. Curran Associates, Inc., 2014. 7

[62] Aboli Marathe, Deva Ramanan, Rahee Walambe, and Ketan Kotecha. Wedge: A multi-weather autonomous driving dataset built from generative vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3317–3326, 2023. 14

[63] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 7

[64] Alexander Mathews, Lexing Xie, and Xuming He. Senticap: Generating image descriptions with sentiments. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016. 6, 7

[65] Mitja Nikolaus, Emmanuelle Salin, Stephane Ayache, Abdellah Fourtassi, and Benoit Favre. Do vision-and-language transformers learn grounded predicate-noun dependencies? *arXiv preprint arXiv:2210.12079*, 2022. 1, 6, 7

[66] Christian Otto, Matthias Springstein, Avishek Anand, and Ralph Ewerth. Understanding, categorizing and predicting semantic image-text relations. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 168–176, 2019. 4

[67] Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. *arXiv preprint arXiv:2112.07566*, 2021. 1, 5, 7, 8

[68] Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Annual Meeting of the Association for Computational Linguistics*, 2021. 2

[69] Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *IEEE Conference on Computer Vision & Pattern Recognition*, 2012. 6, 7

[70] Bryan A Plummer, Matthew Brown, and Svetlana Lazebnik. Enhancing video summarization via vision-language embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5781–5789, 2017. 14

[71] Dragomir Radev, Amanda Stent, Joel Tetreault, Aasish Pappu, Aikaterini Iliakopoulou, Agustin Chanfreau, Paloma de Juan, Jordi Vallmitjana, Alejandro Jaimes, Rahul

Jha, et al. Humor in collective discourse: Unsupervised funniness detection in the new yorker cartoon caption contest. *arXiv preprint arXiv:1506.08126*, 2015. 6, 7

[72] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 8

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

[74] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34:12116–12128, 2021. 2

[75] Arnau Ramisa, Fei Yan, Francesc Moreno-Noguer, and Krystian Mikolajczyk. Breakingnews: Article annotation by image and text processing. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1072–1085, 2017. 6, 7

[76] Abhilasha Ravichander, Eduard Hovy, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. On the systematicity of probing contextualized word representations: The case of hypernymy in bert. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 88–102, 2020. 2

[77] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online, July 2020. Association for Computational Linguistics. 2

[78] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2021. 2

[79] Philipp J. Rösch and Jindřich Libovický. Probing the role of positional information in vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1031–1041, Seattle, United States, July 2022. Association for Computational Linguistics. 1

[80] Emmanuelle Salin, Badreddine Farah, Stéphane Ayache, and Benoit Favre. Are Vision-Language Transformers Learning Multimodal Representations? A Probing Perspective. *AAAI*, 2022. 1, 5

[81] Emmanuelle Salin, Badreddine Farah, Stéphane Ayache, and Benoit Favre. Are vision-language transformers learning multimodal representations? a probing perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11248–11257, 2022. 7

[82] Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. FOIL it! find one mismatch between image and language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265, Vancouver, Canada, July 2017. Association for Computational Linguistics. 5, 7

[83] Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. FOIL it! Find One mismatch between Image and Language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265, Vancouver, Canada, July 2017. Association for Computational Linguistics. 8

[84] Wonyoung Shin, Jonghun Park, Taekang Woo, Yongwoo Cho, Kwangjin Oh, and Hwanjun Song. e-clip: Large-scale vision-language representation learning in e-commerce. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 3484–3494, 2022. 14

[85] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749, 2020. 4

[86] Mustafa Shukor, Nicolas Thome, and Matthieu Cord. Structured vision-language pretraining for computational cooking. *arXiv preprint arXiv:2212.04267*, 2022. 14

[87] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer, 2020. 7

[88] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019. 7

[89] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022. 2

[90] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A Corpus for Reasoning about Natural Language Grounded in Photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy, July 2019. Association for Computational Linguistics. 7

[91] P. Sharma T. Levinboim, A. Thapliyal and R. Soricut. Quality estimation for image captions based on large-scale human evaluations. *arXiv preprint arXiv:1909.03396*, 2019. 6, 7

[92] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*

(EMNLP-IJCNLP), pages 5100–5111, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. 1

[93] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 1

[94] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality. *arXiv:2204.03162 [cs]*, Apr. 2022. arXiv: 2204.03162. 6, 7

[95] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019. 2

[96] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019. 1, 2

[97] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19175–19186, 2023. 1

[98] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. 6, 7

[99] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, M. Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddharth Deepak Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hanna Hajishirzi, and Daniel Khashabi. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Conference on Empirical Methods in Natural Language Processing*, 2022. 2

[100] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8919–8928, 2020. 2

[101] Congcong Wen, Yuan Hu, Xiang Li, Zhenghang Yuan, and Xiao Xiang Zhu. Vision-language models in remote sensing: Current progress and future trends. *arXiv preprint arXiv:2305.05726*, 2023. 14

[102] Christopher Wimmer and Navid Rekabsaz. Leveraging vision-language models for granular market change prediction. *arXiv preprint arXiv:2301.10166*, 2023. 14

[103] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. The fashion iq dataset: Retrieving images by combining side information and relative natural language feedback. *CVPR*, 2021. 7

[104] Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. Recipeqa: A challenge dataset for multimodal comprehension of cooking recipes. *arXiv preprint arXiv:1809.00812*, 2018. 6, 7

[105] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*, 2018. 5, 6, 16

[106] Zongming Yang, Liang Yang, Liren Kong, Ailin Wei, Jesse Leaman, Johnell Brooks, and Bing Li. Seeway: Vision-language assistive navigation for the visually impaired. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 52–58. IEEE, 2022. 14

[107] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 1, 2

[108] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv e-prints*, pages arXiv–2210, 2022. 1, 7

[109] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019. 6, 7

[110] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, André Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv: Computer Vision and Pattern Recognition*, 2019. 1, 2

[111] Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. An explainable toolbox for evaluating pre-trained vision-language models. In *Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 30–37, 2022. 2, 7

[112] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*, 2023. 2

[113] Wangchunshu Zhou, Yan Zeng, Shizhe Diao, and Xinsong Zhang. Vlue: A multi-task benchmark for evaluating vision-language models. *arXiv preprint arXiv:2205.15237*, 2022. 2

[114] C Lawrence Zitnick and Devi Parikh. Bringing semantics into focus using visual abstraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3009–3016, 2013. 7

A. Vision-Language Applications

In this section, we list several real-world applications based on vision-language data, that are or could be a use-case for vision-language foundation models. For each application, we identify one capability necessary for this application that could pose a challenge for vision-language models.

- **Multimodal Dialog** [25]: Use textual and visual context for dialog with a user.
  Example capability: Understand the subjective meaning of some instances, such as jokes, memes (C).

- **Fake News Detection** [42]: Identify fake news in social media.
  Example capability: Understand the intent behind a specific text-image combination (C).

- **Vision-Language Navigation** [12]: Understand natural language instructions in a visual environment.
  Example capability: Understand if there is a mismatch between a text command and the available visual information (R).

- **Tools for Visually Impaired People** [106]: Help a visually impaired person navigate or answer questions on an image.
  Example capability: Precisely describe the structure of a scene (D).

- **Crisis/Event Analysis** [54]: Understand a crisis, the relevant actors and its context based on text-image data.
  Example capability: Understand spatial and temporal context of a text-image instance (G).

- **Video Summarization** [70]:
  Vision-language models can be used in some cases with to complement applications based on video. *Example capability*: Describe visual elements relevant to temporal data in still images (G).

- **Computer-assisted Food Analysis** [86]: For instance, it can consist in image-text retrieval applied to food, and can have applications in health and nutrition.
  Example capability: Understand the temporal and spatial structure of text-image food or recipe data (G).

- **Biomedical Vision-Language Processing** [7]: Interpreting visual and textual biomedical data for clinical care.
  Example capability: Understand and reason on complex biomedical semantics (G& R).

- **Agriculture** [14]: Identify plant disease for agricultural purposes and differentiating between healthy and diseased plants.
  Example capability: fine-grained classification from limited examples (D).

- **Autonomous Driving** [62]: For instance, vision-language models can help design datasets geared towards autonomous driving that are not present in sufficient quantity in real datasets.
  Example capability: Semantic understanding of events such as weather, accidents or other incidents (D& G).

- **E-commerce Recommendation** [84]: Product recommendation based on textual and visual information. There are several possible subtasks such as product matching, classification, clustering.
  Example capability: Associate text to the corresponding semantic information using visual data despite limited grammatical structure (D).

- **Multimodal Hate Speech Detection** [16]: Detecting hate speech that is present in multimodal data.
  Example capability: Understanding subjective and ambiguous meaning of text-image data (C).

- **Remote Sensing Understanding** [101]: Study of satellite mages in correlation with text data.
  Example capability: Differentiate semantically between atmospheric visual data and relevant ground visual data (G).

- **Market Prediction** [102]:
  Predict the evolution of the stock market using text and image data. *Example capability*: Identify patterns in time series data represented using text or images (G).

B. Details on Methodology for News-related data

In order to get a comprehensive perspective of news data, we select 5 online news sources from several countries and varying demographics. We restrict ourselves to English language newspapers.

- The New York Times, a daily American newspaper [1]
- Daily Mail, a daily British tabloid [2]
- Wall Street Journal, a daily American business newspaper [3]
- France 24, a French international news network [4]
- Al Jazeera, a Qatari international news network [5]
- Global Times, a daily Chinese English-language newspaper. [6]

We select three dates and study a captioned image from those newspapers for each of those dates, selecting a topic at random for each example. These examples vary across topics: ranging from business to culture.

C. Detailed Taxonomy

The taxonomy presented in this section is a preliminary attempt at classifying vision-language capabilities. It is not exhaustive. In this section, an instance is composed of at least a text and an image.

C.1. Denotation

The capabilities of a vision-language model to associate a text and an image are conditioned on its ability to take into account information at different structural levels, from local information to information relating to the whole instance.

Denotation skills, local: These capabilities evaluate the understanding of a single element of a text-image instance, independently of the rest of the instance.

- Basic Property Detection: *Def.* The ability to detect the presence of a basic property (e.g. color, texture) and associate it to a corresponding word.
  Ex. Associate the color red with the word 'red'.
- Object Perception: *Def.* The ability to differentiate between objects, both at coarse and fine-grained level. Includes the understanding of the continuity of an object (e.g. segmentation).
  Ex. Identify a flower from its picture.

[1] https://www.nytimes.com/
[2] https://www.dailymail.co.uk/
[3] https://www.wsj.com/
[4] https://www.france24.com/en/
[5] https://www.aljazeera.com/en
[6] https://www.globaltimes.cn

Denotation skills, structural: These capabilities evaluate the understanding of the dependency between an element and the rest of the instance, or between several elements of an instance, i.e. the compositionality of an instance. As a whole, those skills also require local understanding, because the model needs to also understand each element individually. A compositional instance depends, in addition to the individual elements, on the structure of those elements.

- Syntactic Understanding: *Def.* The ability to grasp the syntactic structure of a sentence and deduce the relation between different words using visual information. Includes the resolution of polysemy.
  Ex. Differentiate 'bear' as a verb or a noun.
- Scene Understanding: *Def.* The ability to grasp the structure of an image using textual information. Includes counting and positional understanding (i.e. the ability to understand depth, distance and position between objects in the referential of the image).
  Ex. Count people in a crowd.
- Multimodal Alignment Understanding: *Def.* The ability to correctly associate textual elements using visual information. The textual elements can be non-explicit (i.e. co-reference resolution). Includes understanding the static interaction between people and objects in an instance.
  Ex. Associate a predicate to the correct noun.

Denotation skills, global: These capabilities evaluate the understanding of the whole instance.

- Document Type Understanding: *Def.* The ability to detect the topic of an instance, its source (e.g. author, machine used to capture it), its date or its style.
  Ex. Specify how a medical image was captured.
- Focus Identification: Understanding what elements are or are not the focus of an instance using its textual and visual information.
  Ex: Identify which person is the focus of a newspaper image/caption pair.

Denotation skills characterize factual understanding of a vision-language instance and its components. We listed in this section several skills that, to our knowledge, are necessary to establish this understanding of a vision-language instance. This list does not include the ability to ground the instance in the world or use knowledge specific to a domain.

C.2. Grounding

In this section, we identify several types of grounding.

Grounding skills, temporal: These capabilities evaluate a model's ability to understand the situation of an instance in time.

- Temporality Perception: *Def.* The ability to detect if time affects the instance. For the image modality, it includes whether an object/structure changes state and position in the immediate past or future. For the textual modality, it means using text information (e.g. verb tense) to detect temporality.
  Ex. Detect which element of an instance are moving.

- Object State Understanding: *Def.* The ability to associate the state of an object with corresponding words and differentiate the role of an object depending on its state.
  Ex. Differentiate between an empty or full glass.

- Temporal Extrapolation: *Def.* The ability to extrapolate the past or future structure of a scene using multimodal information.
  Ex. Understand that a glass will break if pushed.

- Time Period Identification: *Def.* The ability to identify a specific period in a multimodal instance.
  Ex. Recognize that an instance depicts medieval times.

Grounding skills, spatial: These capabilities evaluate a model's ability to understand a scene as part of a wider spatial context.

- Spatial Understanding: *Def.* The ability to ground an instance in the world using textual and visual information. Includes the understanding of perspective, depth, size and spatial referential.
  Ex. Recognize that a plane in the sky is the same size as at the airport.

- Physical Spatial Understanding: *Def.* The ability to understand how physics affect the position of objects in an image. Includes occlusion, obstacles, contact.
  Ex. A partially hidden object is still the same.

- Spatial Extrapolation: *Def.* The ability to extrapolate the spatial context not seen in the instance using multimodal information.
  Ex. Extrapolate what is behind the photograph taking a picture.

- Location Identification: *Def.* The ability to recognize known places using multimodal information.
  Ex. Recognize a specific country using street furniture.

Grounding skills, knowledge: These capabilities evaluate a model's ability to use specific technical or cultural knowledge.

- Semantic Grounding: *Def.* The ability to exploit knowledge from semantic relations (e.g. roles, synonyms, antonyms and hypernyms).
  Ex. Understand that 'robin' and 'bird' can refer to the same element.

- Technical Grounding: *Def.* The ability to exploit knowledge from a specific domain (e.g. medical). Includes the understanding of specialized objects, technical terms, events, or specific named entities. *Ex.* Associate visual information to the term 'pneumothorax'.

- Cultural Grounding: *Def.* The ability of a model to understand the cultural context of an instance, with respect to textual or visual elements, and differentiate across cultures.
  Ex. A mask can mean a medical mask or a mold that represents someone else. The latter, following cultures, can be traditional, religious, used for theater or for carnivals.

- Symbolic System Grounding: *Def.* The ability to recognize symbols and characters in an image. Ranges from Optical Character Recognition to the ability to recognize the meaning of a symbol.
  Ex. Describe signs held at a demonstration.

Grounding skills, multimodal: These capabilities evaluate the understanding of concepts related to a foreign modality not present in the instance.

- Human Senses Grounding: *Def.* Detecting and associating words or objects that can refer to human senses not linked to vision, such as hearing, touch or taste.
  Ex. Associate a waterfall with the word 'loud'.

The use of grounding can be necessary for specific applications. For instance, the spatial and temporal grounding skills can be used for vision-language navigation. However, those applications can also require other types of skills, such as reasoning.

C.3. Reasoning

We identify a few reasoning tasks necessary for vision-language models, using as inspiration existing tasks such as NLP tasks [60, 56, 10, 105, 13]. As a whole, monomodal reasoning tasks can be adapted to multimodality. Reasoning skills can require prior understanding of several other skills, for instance related denotation or grounding.

Reasoning skills, semantic: These capabilities evaluate a model's ability to reason semantic knowledge.

- Abnormality Detection: *Def.* The ability to detect an abnormal instance. Includes making the distinction between something rare and something unrealistic. Can be local, structural or global.
  Ex. Detect that an object is at an unrealistic position.

- Mismatch Detection: *Def.* The ability to spot if information is missing from one of the two modalities.
  Ex. Detect that a sentence asks a question about an object which isn't present in the image.

Reasoning skills, logic: These capabilities evaluate a model's ability to reason using logic or mathematical concepts.

- Logical Operations: *Def.* The ability to understand logic operations (e.g. negation, *or*, *and*).
  Ex. Understand 'no' in 'There is no cat'.

- Comparison: *Def.* The ability to compare two parts of an instance. Can also be applied between multiple instances.
  Ex. Compare the size of two objects in an image.

- Multimodal Inference: *Def.* The ability to detect whether one instance can be entailed from another.
  Ex. Use context and a medical image to assist in a diagnosis.

- Mathematical Reasoning: *Def.* The ability to use topological, geometrical, arithmetical or algebraic skills.
  Ex. Answer a math-related IQ question.

Reasoning skills, complex: These capabilities evaluate a model's ability to reason using abstract reasoning or in multiple stages.

- Extrapolation: *Def.* The ability to complete an instance from incomplete visual or textual information. Includes the ability to distinguish between extrapolation and hallucinations.
  Ex. Deduce part of an obstructed text in an image without hallucinating.

- Multi-hop Reasoning: *Def.* The ability to perform reasoning using multiple steps.
  Ex. Path computing in vision-language navigation.

- Introspection: *Def.* The ability to explain the prediction of a task.
  Ex. Explain the reasoning when answering a question.

These can be complemented by other monomodal reasoning tasks transferred to the multimodal domain. Some of those tasks can require task-specific data or fine-tuning, and be difficult to achieve using only a foundation model.

C.4. Connotation

The skills listed in this section may not be useful to all applications of vision-language models, as they rely on individual interpretation of multimodal instances. In addition, their evaluation is subjective and can widely vary depending on the annotations.

Connotation skills, interpretation: These capabilities evaluate a model's ability to interpret the meaning or intent of an instance:

- Symbolism Understanding: *Def.* The ability to understand the intent behind the symbolism in multimodal elements (e.g. metaphors).
  Ex. Associate a person holding a scale with 'justice'.

- Ambiguity Understanding: *Def.* The ability to understand voluntary ambiguity (e.g. optical illusions, word plays).
  Ex. Understand that an image shows a duck or a rabbit.

- Sentiment Understanding: *Def.* The ability to understand the emotions evoked by an instance. Includes the detection of humor and irony.
  Ex. Understand that the gap between an image and its associated text conveys humor.

Connotation skills, criticism: These capabilities evaluate the understanding of the quality of an instance.

- Stylistic Appreciation: *Def.* The ability to evaluate whether stylistic elements are appropriately and consistently used.
  Ex. Criticize the symmetry in an image.

- Effectiveness Evaluation: *Def.* The ability to evaluate whether an instance is effective at expressing its intended meaning.
  Ex. Evaluate whether a cartoon transmits the intended message.

These skills can be used in real-world applications where the interpretation of an instance is important, such as applications related to art.