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 recommendations 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]

[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

- 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.

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 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 is 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.

[6] https://www.globaltimes.cn

- 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 the 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.