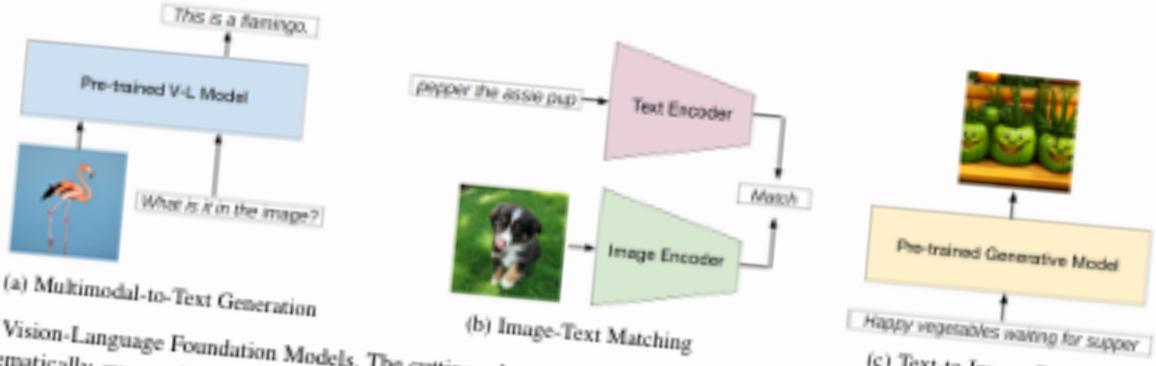

(a) Multimodal-to-Text Generation (b) Image-Text Matching (c) Text-to-Image Generation

Fig. 1: Vision-Language Foundation Models. The cutting-edge research in prompt engineering on Vision-Language Foundation Models is systematically summarized. Three main types of vision-language models are focused in this work, namely, multimodal-to-text generation models (*e.g.*, Flamingo [1]) in subfigure a, image-text matching models (*e.g.*, CLIP [2]) in subfigure b, and text-to-image generation models (*e.g.*, Stable Diffusion [3]) in the subfigure c. More details of each type are introduced in the later sections.

neering of pre-trained VLMs. Specifically, we classify prompting methods into two main categories based on the readability of the templates, *i.e.*, hard prompt and soft prompt. hard prompts can be further divided into four subcategories, namely task instruction, in-context learning, retrieval-based prompting, and chain-of-thought prompting. Soft prompts, on the other hand, are continuous vectors that can be fine-tuned using gradient-based methods. Note that this survey primarily focuses on prompting methods that maintain the model's architecture, and thus, the methods such as P-tuning [13] and LoRa [14] that introduce additional modules into the model, are not the primary scope of this survey.

We investigate the prompt engineering on three types of VL models, which are *multimodal-to-text generation models*, *image-text-matching models*, and *text-to-image generation models*. A clear definition of each model type is provided in Sec. 2.1. Moreover, we categorize existing prompt-engineering approaches from an encoder-decoder perspective as shown in Fig. 1, *i.e.*, encode-side prompting or decode-side prompting, where the prompts are added to the encoder and decoder, respectively.

The rest of this paper is organized as follows. In Sec. 2, we summarize and define the taxonomy and notations that we use across this survey. Sec. 3, 4, and 5 present the current progress of prompt engineering on multimodal-to-text generation models, image-text-matching models, and text-to-image generation models, where each section first presents the preliminaries of the corresponding models followed by a detailed discussion of the prompting methods, then investigates the applications and the responsible AI considerations of such prompting methods. Sec. 6 provides a comparison between prompting unimodal models and VLMs, and we make an in-depth discussion about their analogies and differences. Finally, in Sec. 7, we highlight the challenges and potential research directions.

In order to facilitate the literature search, we also build and release a project page [1] where the papers relevant to our topic are organized and listed.

2 Taxonomy

In this section, terms and notations related to Prompting Engineering on VLMs used throughout the paper are introduced.

1. https://github.com/JindongGu/Awesome-Prompting-on-Vision-Language-Model/

2.1 Terminology

This is a list of terms along with their descriptions. Note that instead of formally defining the following concepts, we provide a general description for readers.

- *Prompt:* Additional information or hints provided to a model to guide its behavior or help it perform a specific task;
- *Prompting Method:* An approach used to incorporate prompts into the input to guide model behavior or enhance model performance;
- *Multimodal-to-Text Generation:* Generating textual descriptions or narratives from multimodal input data, *e.g.*, a combination of vision and language data;
- *Image-Text Matching:* Establishing a semantic relationship or alignment between images and textual descriptions;
- *Text-to-Image Generation:* Generating visual images from textual descriptions.
- *In-context Learning:* A prompting method by providing models with instructions or demonstrations within relevant contexts to solve new tasks without requiring additional training.
- *Chain-of-thought:* A prompting method that enhances reasoning skills by instructing a model to generate a sequence of intermediary actions that guide towards solving a multi-step problem and reaching the ultimate solution.

2.2 Notations

These are the mathematical notations that are followed throughout the paper (Tab. 1). All the formulations of this work will stick to these notations unless otherwise specified.

3 Prompting Model in Multimodal-to-Text Generation

3.1 Preliminaries of Multimodal-to-Text Generation

Large language models (LLMs) have demonstrated impressive capabilities in the field of NLP, prompting researchers to explore ways of integrating visual modality into these models' training framework. This integration aims to enhance their linguistic prowess and expand their applicability to multimodal tasks.