A Systematic Survey of Prompt Engineering on Vision-Language Foundation Models

Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, Philip Torr

Abstract—Prompt engineering is a technique that involves augmenting a large pre-trained model with task-specific hints, known as prompts, to adapt the model to new tasks. Prompts can be created manually as natural language instructions or generated automatically as either natural language instructions or vector representations. Prompt engineering enables the ability to perform predictions based solely on prompts without updating model parameters, and the easier application of large pre-trained models in real-world tasks. In past years, Prompt engineering has been well-studied in natural language processing. Recently, it has also been intensively studied in vision-language modeling. However, there is currently a lack of a systematic overview of prompt engineering on pre-trained vision-language models. This paper aims to provide a comprehensive survey of cutting-edge research in prompt engineering on three types of vision-language models: multimodal-to-text generation models (*e.g.,* Flamingo), image-text matching models (*e.g.,* CLIP), and text-to-image generation models (*e.g.,* Stable Diffusion). For each type of model, a brief model summary, prompting methods, prompting-based applications, and the corresponding responsibility and integrity issues are summarized and discussed. Furthermore, the commonalities and differences between prompting on vision-language models, language models, and vision models are also discussed. The challenges, future directions, and research opportunities are summarized to foster future research on this topic.

Index Terms—Prompt Engineering, Vision Language Model, Multi-modal Model, Natural Language Processing, Computer Vision.

1 Introduction

Prompt engineering is an approach to adapting a large pre-trained model, also known as a foundation model, to new tasks by augmenting the model input with task-specific hints. Specifically, the model's input is augmented by an additional part, called prompt, which could be manually created natural language instructions [4], automated generated natural language instructions [5], or automated generated vector representations [6]. The natural language instructions have been also referred to as *discrete prompts* or *hard prompts*, while the vector representations are called *continuous prompts* or *soft prompts*.

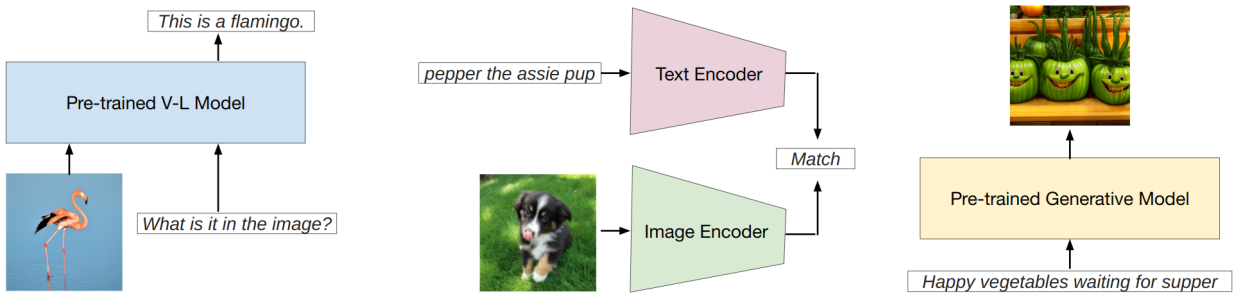
Prompt engineering has indeed co-appeared and gained prominence with the emergence of large pre-trained models and together led to a paradigm shift in machine learning (ML). The traditional paradigm requires labeling a considerable amount of data and then training a task-specific ML model from scratch or fine-tuning a pre-trained large model. The model's performance heavily relies on the quality and amount of labeled data, which can be resource-intensive to acquire. Besides, the traditional paradigm requires tuning the model's parameters to some extent, *i.e.*, entire parameters in the case of training an ML model from scratch or fully fine-tuning a pre-trained model and partial parameters in the case of parameter-efficient finetuning. This limits the extensibility of an ML model and requires a specific model copy for each task. Recently, prompting a pre-trained large model to adapt it for specific tasks has become a new trend. The key idea of prompt engineering is to provide hints along with input to guide a pre-trained model for solving a new task using its existing knowledge. If the hints are human-interpretable natural language (*hard prompts*), the related studies have been referred to as *In-Context Learning* [7], which enable the model to learn from task instructions, demonstrations with a few examples, or supporting information in the context. Also, the hints could be continuous vector representations (*soft prompts*). The related work has been referred to as *Prompt-Tuning* [6], which optimizes prompts directly in the embedding space of the model.

Compared to the traditional paradigm, prompt engineering has multiple advantages. Firstly, it requires a few labeled data to adapt a pre-trained model to new tasks, which greatly reduces the effort of human supervision and computation resource for fine-tuning. Secondly, prompt engineering enables a pre-trained model to perform predictions on new tasks solely based on the prompt without updating any of the model's parameters, allowing serving a large scale of downstream tasks using the same model. This makes it possible to apply large-scale pre-trained models for real-world applications.

Prompt engineering has been first studied and popularized in natural language processing (NLP) [8, 9], and then gained great attention in computer vision [10, 11], as well as in vision-language modeling [1, 12]. While there is an abundance of literature on prompt engineering in the NLP domain, there is currently no systematic overview available to provide insight into the current state of prompt engineering on pre-trained vision-language models (VLMs), which present their own unique challenges.

In this paper, we aim to bridge this gap by providing a comprehensive survey of cutting-edge research in prompt engi-

- *Jindong Gu and Philip Torr are with University of Oxford.*
- *Zhen Han, Shuo Chen, Bailan He are with Ludwig Maximilian University of Munich.*
- *Ahmad Beirami is with Google Research.*
- *Gengyuan Zhang, Ruotong Liao, and Volker Tresp are with Ludwig Maximilian University of Munich and Munich Center for Machine Learning.*
- *Yao Qin is with Google Research and University of California, Santa Barbara.*
- *Corresponding E-mail from Jindong Gu: {jindong.gu}@outlook.com.*



(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 *muiltimodal-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.

TABLE 1: The used mathematical notations are listed. They are followed throughout the paper.

x	A clean input image
t	A sentence paired with an image
y	A ground-truth class label of an image
χ	Input distribution
$f(\cdot)$	A vision-language model
$f_v(\cdot)$	A visual encoder
$f_e(\cdot)$	A textual encoder
$\{v_i\}_{i=1}^M$	visual tokens
$\{c_i\}_{i=1}^M$	textual tokens
$\{z_i\}_{i=1}^M$	visual prompt tokens
$\{t_i\}_{i=1}^M$	textual prompt tokens
H^l	l^{th} A layer of the target network
L	Label word token
H_k^i	the k^{th} activation in l^{th} layer of the target model
z^i	Model output logits

To maintain consistency with the training methodologies employed by LLMs, generation-based vision-language models (VLMs) typically comprise three essential components: *text feature*, *visual feature*, and *fusion module*. These components synergistically collaborate, enabling the models to effectively leverage textual and visual information to generate coherent and contextually relevant outputs.

Incorporating the visual modality into LLMs has opened up exciting opportunities for various applications, such as visual commonsense reasoning [15], visual question answering [1, 16, 17, 15], multimodal dialogue systems [18, 1, 12], *etc*. By combining textual and visual cues, VLMs have the potential to provide a more comprehensive understanding of multimodal data and produce outputs that align with human-like reasoning and perception [19]. Furthermore, the fusion of text and visual features within VLMs plays a crucial role in seamlessly integrating information from both modalities. This fusion process enables the model to capture interdependencies and interactions between textual and visual elements, resulting in more accurate and contextually grounded generations [19].

Text Feature. Early studies on VLMs commonly employed the preprocessing technique introduced by BERT [20]. The raw text undergoes tokenization and is concatenated with special tokens, `[CLS]` and `[SEP]`, represented as `<[CLS],`c_1`, . . . ,`c_m`,[SEP]>`, where token c_i is associated with a word embedding. However, with the progression of language model research, more advanced models have emerged, showcasing emergent abilities such as in-context learning [21] and chain-of-thought reasoning [4]. Building upon these advancements, the latest generation of VLMs has embraced powerful language models like T5 [22] and GPTs [23], which further enhances their linguistic capabilities.

To accommodate different modalities in the input, recent works have introduced new special tokens. For example, [24] incorporate an additional image classification token `[CLS_I]`, while [15] use `<image>` and `</image>` to indicate the beginning and end of the encoded image embedding and `<s>` and `</s>` to mark the beginning and end of a sequence. In another approach, [1] employs `<BOS>` to represent the "beginning of sequence" and `<EOC>` to denote "end of chunk". These special tokens serve to differentiate and identify the boundaries between different modalities, allowing the model to effectively process and leverage multimodal information.

Visual Feature. To obtain a consistent representation of input as a sequence of embeddings for both modalities, the image x is transformed into a sequence of embedding vectors: $x =< v_1, v_2, ..., v_M >$. Accurately representing the information conveyed by images is crucial for downstream tasks but can be challenging. CNN structures have been commonly used in prior research for extracting image features. For instance, models like ViLBERT [25] and VL-T5 [26] employ faster R-CNN [27] to detect object regions in images and encode them as a sequence of Region-Of-Interest (ROI) features. However, this approach may overlook important regions in an image. To address this limitation, approaches like OFA [28] and Flamingo [1] utilize ResNet to encode information from the entire image, considering a broader context. Additionally, leveraging the powerful feature extraction capabilities of the transformer architecture, models such as SimVLM [29], PaLI [30], MAGMA [31], and BLIP2 [32] adopt the Vision Transformer (ViT) [33] architecture for image representation. This allows them to effectively capture visual information and incorporate it into the multimodal framework.

Fusion Module. The fusion module plays a crucial role in integrating text and image embeddings to create a joint representation. A well-designed fusion module can capture interactions and relationships between modalities, prevent information loss, avoid semantic mismatch, mitigate biases, and enables comprehensive understanding. For example, in Visual Question Answering (VQA), the fusion module enables the model to leverage both textual and visual information to understand the question and the corresponding image, leading to accurate answers. To improve the ability of answer generation, prompts can be manually designed for different tasks and included as part of the input to the fusion module. These prompts serve as additional information or cues that guide the model's understanding of the question and the image. As for generation-based VLMs, there are two main types of fusion module approaches based on the integration of visual and textual modalities: *encoder-decoder as a multi-modal fusion module* and *decoder-only as a multi-modal fusion module*.

In the encoder-decoder as a multi-modal fusion module approach, models like VL-T5 [26], SimVLM [29], OFA [28], and PaLI [30] focus on creating a joint representation that combines both modalities at an early stage. The overall formulation can be represented as:

$$y = \mathcal{G}(\mathcal{E}(x_{input})) \tag{1}$$

where the x_{input} represents the given input and y denotes the corresponding ground-truth, respectively. The fusion encoder function $\mathcal{E}$ integrates the visual and textual information to create a joint representation that captures their interactions and dependencies. This fused representation is then fed into the generating module $\mathcal{G}$, which performs further processing and generates the desired outputs for the downstream tasks.

In the decoder-only as a multi-modal fusion module approach, models like Frozen [17], Flamingo [1], and MAGMA [31] directly combines the visual and textual information in the decoding stage, without explicitly creating a joint representation at an earlier stage.

This approach allows the model to effectively incorporate both modalities during the generation process and produce contextually relevant outputs. The formulation can be represented as:

$$y = \mathcal{G}(x_{input}) \tag{2}$$

A special case is PICa [16], which represents images as textual descriptions and utilizes GPT-3 as the fusion module. This approach treats images as text and leverages a pre-trained language model like GPT-3 to generate outputs based on the pure text input.

In addition, BLIP-2 [32] examines the fusion of two distinct modules for integration: the decoder-based OPT [34] and the encoder-decoder based FlanT5 [35]. The study further offers an analysis of the respective strengths and benefits offered by these fusion modules.

3.2 Multimodal-Text Prompting Methods

Fig. 2 illustrates the classification of prompting methods. Prompting methods fall into two categories: hard prompts, which are labor-intensive, manually crafted text prompts with discrete tokens, and soft prompts, which are optimizable, learnable tensors concatenated with input embeddings, but lack human readability due to their non-alignment with real word embeddings.

3.2.1 Hard prompt

Hard prompts involve manually crafted, interpretable text tokens, *e.g.,* adding "*A photo of* " before the input for captioning tasks. Hard prompts can be further divided into four subcategories: *task instruction*, *in-context learning*, *retrieval-based prompting*, and *chain-of-thought prompting*. It is important to note that retrieval-based prompting is often used to select samples for in-context learning.

Task Instruction Prompting. This method involves the use of carefully designed prompts that provide explicit task-related instructions to guide the model's behavior [36, 37]. The formulation for this method can be represented as $x_{\text{input}} = \mathcal{H}(x, t)$. Here, $\mathcal{H}$ serves as the task instruction function, taking the image x and text t as inputs and producing the modified input representation x_{input}.

In-context Learning. In-context Learning [7, 23] is a method where the model is exposed to a sequence of related examples or prompts, enabling it to learn and generalize from the provided context. The in-context learning method can be represented using the equation $x_{\text{input}} = \mathcal{H}(\mathcal{C}, x, t)$. Here, $\mathcal{H}$ denotes the task instruction function which integrates the given context $\mathcal{C}$ with the image x and text t inputs. The resulting modified input representation x_{input} captures the model's understanding of the context and is used to generate coherent and contextually relevant responses. By exposing the model to a sequence of related examples or prompts, the in-context learning method promotes improved performance in understanding and generating responses [4].

Retrieval-based Prompting. Retrieval-based Prompting [16, 38, 39, 40] is a method that involves selecting prompts or context using retrieval techniques. In this approach, the model retrieves relevant prompts or context from a prompt pool or external knowledge base to guide its generation or decision-making process. The retrieval-based prompting method can be denoted by the formulation: $\mathcal{C} = \mathcal{R}(x, t)$. In this equation, $\mathcal{R}$ signifies the retrieval method that garners pertinent prompts or context based on the image x and text t inputs. The retrieved context $\mathcal{C}$ is then used to guide the model's generation or decision-making process. It is worth noting that the retrieval method $\mathcal{R}$ can vary depending on the specific approach and the available prompt pool or knowledge base. This method allows the model to benefit from existing information and improve its performance by leveraging relevant prompts or context during the generation process [38, 39, 40].

Chain-of-Thought Prompting. Chain-of-Thought Prompting [4, 9, 5] is a method where the model is prompted with a series of instructions or questions that progressively build upon each other. Each prompt in the chain adds context or narrows down the focus, enabling the model to generate more coherent and contextually appropriate responses. This method helps the model maintain a logical "chain" throughout the conversation. The formulation for the chain-of-thought prompting method does not involve a specific equation but rather the iterative process of applying prompts [4]. At each step l in the chain, the model's response from the previous prompt is used as input for the next prompt. This can be represented as $\mathcal{T}^{l+1} = \mathcal{T}^{l}(x, t)$. Here $\mathcal{T}$ represents the prompt function that takes the image x and text t inputs and generates a response. The output of the l-th prompt, denoted as $\mathcal{T}^{l}(x, t)$, serves as the input for the $(l+1)$-th prompt $\mathcal{T}^{l+1}$. By progressively building upon the previous prompts, the iterative nature of the chain-of-thought prompting method helps the model maintain coherence and generate responses that align with the evolving context of the conversation [4].

3.2.2 Soft prompt

Unlike hard prompts, soft prompts are characterized as continuous vectors that can be fine-tuned using gradient-based methods [6, 41]. For example, this process might involve concatenating a learnable vector with the input embeddings and subsequently optimizing these to align with a particular dataset. Soft prompts can be classified according to whether new tokens are internally incorporated within the model's architecture or simply attached to the input. This distinction generally relates to two specific strategies: *prompt tuning* and *prefix token tuning*. However, this survey focuses exclusively on prompt methods that do not involve modifying the underlying model itself, and thus techniques like P-tuning [13] and LoRa [14], which alter the fundamental structure of the model, are not within the primary scope of this study.

Prompt Tuning. Prompt tuning [6] creates continuous vector representations as input hints. During the training process, the model learns to refine the prompts, aiming to improve its performance on specific tasks. This method enables the model to dynamically generate effective prompts based on its understanding of the task. The objective of prompt tuning, with the prompting parameter x_p, can be demonstrated as follows:

$$\underset{x_p}{\operatorname{argmin}} \mathcal{L}(\mathcal{F}(y_i, x_p)|y_{<i}, x_{input}) \tag{3}$$

where $\mathcal{F}(y_i, x_p)$ represents the model's output given the prompting parameter x_p. Here $y_{<i}$ denotes the previously generated outputs, and x_{input} refers to the modified input based on the prompt. The objective of prompt tuning is to minimize the loss $\mathcal{L}$ between the model's output and the desired output, given the previously generated outputs and the modified input. By continuously refining the prompts through prompt tuning, the model adapts its behavior and improves its performance on specific tasks. The dynamic generation of effective prompts based on the model's understanding enhances its capability to generate accurate and contextually relevant responses.

Prefix Token Tuning. Similar to prompt tuning, prefix token tuning [42] involves adding task-specific vectors to the input.

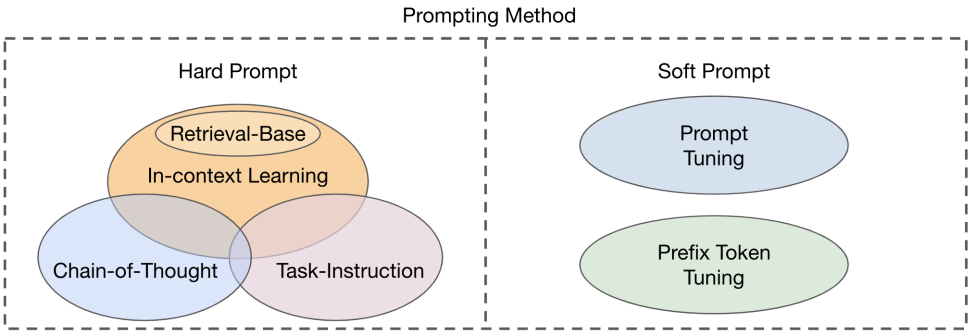


Fig. 2: Prompting methods in multimodal-to-text generation. Prompting methods can be divided into two main categories based on the readability of the templates: hard prompt and soft prompt. Hard prompt encompasses four subcategories: task instruction, in-context learning, retrieval-based prompting, and chain-of-thought prompting. Soft prompts are classified into two strategies: prompt tuning and prefix token tuning, based on whether they internally add new tokens to the model's architecture or simply append them to the input. However, this study primarily concentrates on prompt methods that avoid altering the base model, excluding techniques like P-tuning and LoRa that modify the model's core structure.

However, in this case, these vectors are inserted in all model layers and can be trained and updated independently while keeping the rest of the pre-trained model's parameters frozen.

It's worth noting that these prompting methods are not mutually exclusive. They can be combined and used together to achieve desired results in various settings and tasks. The choice of prompting method depends on the specific task, dataset availability, and the desired level of control and customization required for the model's behavior.

3.3 Advances in Prompting Techniques for VLM

This section will overview the use of prompting techniques in various VLMs to boost performance. For a clear and structured presentation, models will be divided into two main types based on their fusion modules: 1) models utilizing an encoder-decoder as the fusion module, and 2) models employing a decoder-only as the fusion module.

Prompting Models with Encoder-decoder as the Fusion Module. Early studies in VLMs often involved designing task-specific architectures on top of transformer encoders. However, recent advancements have introduced a unified vision-language framework that incorporates an encoder as the fusion module. Notable examples of such models include VL-T5 [26], SimVLM [29], and OFA [28]. They employ two main prompting methods: handcrafted instructions and prompt tuning.

Both VL-T5 [26] and OFA [28] utilize text prefixes as prompts. For example, "*vqa:*" is used for vision question answering, and "*caption:*" is employed for image captioning tasks. SimVLM [29] introduces the prefix "*a photo of:*" to enhance the quality of decoded captions. In addition, VL-T5 [26] introduces shared visual sentinel tokens (<vis_i>) to specify corresponding image regions of Region of Interest (RoI) features. Text sentinel tokens (<$text\_i$>) are used to replace contiguous text. Similarly, OFA [28] generates location tokens that specify the position of the region (<x_1, y_1, x_2, y_2>). These special tokens facilitate the structured incorporation of visual and textual information.

Building upon these special tokens, VL-T5 [26] utilizes the prompt "*caption region:* <vis_i>" for the grounded captioning task, indicating that the model should generate a caption based on the specified visual region. OFA [28] prompts the proposed grounded question answering task using the template "*Q: what color is the car in the region? region:* <x_1, y_1, x_2, y_2> *A:*", providing instructions for the model to answer the question by referring to the specified visual region. Prompt tuning on OFA is explored by [43], who introduce tunable prompt embeddings at each layer. Experimental results demonstrate that this lightweight prompt-tuning approach is not only efficient but also resilient against adversarial attacks.

Prompting Models with Decoder-based Fusion Module. Another line of research focuses on utilizing the decoder as a fusion module in VLMs. Frozen [44] and BLIP-2 [45] exemplify models that employ image conditional prefix tuning. Frozen [17] introduces the concept of preserving the language capabilities of a LLM while incorporating visual information as a prefix. It achieves this by freezing the model and training a separate vision encoder to represent images. In Frozen, visual information is represented as a sequence of two embeddings, serving as a visual prefix. The authors also propose task induction techniques, such as instructing the model to "*Answer with dax or blicket,*" and evaluate the model's performance with various forms and amounts of in-context learning for downstream tasks. To effectively facilitate cross-modal alignment, BLIP-2 [45] does not fine-tune the vision encoder. Instead, it introduces a Querying Transformer (Q-Former) to extract visual features from the frozen image encoder, using the extracted query embeddings as soft visual prompts. MAGMA [31] follows a similar approach to Frozen, incorporating a new image prefix encoder while keeping the language model frozen. Task instructions, such as "*A picture of* " are used for image captioning. Flamingo [1] explores the capabilities of few-shot learning and employs various prompt techniques. The authors introduce special tokens, `<BOS>` (beginning of sequence) and `<EOC>` (end of chunk), to differentiate sample pairs. In the zero-shot scenario, text prompts that do not contain corresponding vision information are used. In the few-shot setting, different formatting is employed for various tasks (*e.g.,* "*Question:* {*question*} *Answer:* {*answer*}" for visual question-answering tasks), and the retrieval-based in-context example selection (RICES) [16] approach is utilized to select suitable sample pairs as prompts. Prompt ensembling techniques are also employed to calculate the final scores. For specific

tasks such as HatefulMemes, prompts are designed to incorporate provided OCR information. Additionally, hand-crafted dialogue prompts are specifically designed for presented dialogues. [30] extends the multilingual capabilities of LLMs to VLMs without freezing any parameters. They achieved this by explicitly specifying the intended language in the prompt instruction. For example, a prompt may be formulated as "*Generate the alt_text in* `<lang>`", where `<lang>` represents the language code associated with the desired text string. Furthermore, Microsoft proposes a series of Multimodal Large Language Models (MLLM), namely Kosmos-1 [15] and Kosmos-2 [46]. These models possess the ability to perceive diverse modalities and evaluate a wide range of tasks, including zero-shot, few-shot, and multimodal chain-of-thought prompting scenarios. Textual instructions are used to enable the model to better understand downstream tasks. For example, in Kosmos-1 [15] phrases like "*Here are three/four/eight images:*" and "*The following image is:*" are employed for the Raven IQ test. In chain-of-thought prompting, Kosmos-1 first uses a prompt (*e.g.,* "*Introduce this picture in detail:*") to guide the model to generate a rationale. Then, a task-aware prompt incorporating the generated rationale is utilized to produce the final results. Based on Kosmos-1 [15], Kosmos-2 [46] incorporating grounding and referring capabilities by using text span with bounding box as prompt, *i.e.*, "`<p> text span </p><box><loc1><loc2></box>`, where `<loc1>` and `<loc2>` are location tokens `<p>`, `</p>`, `<box>` and `</box>` are special boundary and text span tokens, respectively.

PICa [16] takes a different approach by not learning visual embeddings. Instead, it converts images into textual descriptions and queries GPT-3 directly to predict the answer. Leveraging the few-shot learning ability of GPT-3, PICa adapts to the visual question-answering (VQA) task with only a few in-context examples during inference time. GPT-4 [18], the latest version of ChatGPT, has been introduced as an advanced VLM. In addition to employing task-specific hard prompts, GPT-4 also incorporates the in-context learning approach to tackle complex tasks such as AP Art History [47].

3.4 Understanding Prompting

To deeply understand the factors impacting prompting in multimodal-to-text generation models, the following aspects will be introduced:

Dataset-specific Prefixes. The choice of text prompts can have a significant impact on the performance of models. VL-T5 [26] experimented with a single prefix "*vqa*" for both Visual Question Answering (VQA) and GQA [48] tasks. The results demonstrated that a single model can effectively handle multiple VQA tasks without the need for dataset-specific prefixes.

Freezing the Language Model. Many explorations of prompting in multimodal-to-text generation models rely on the powerful generative capabilities of language models. To preserve the extensive capabilities of LLMs, approaches like Frozen [17], MAGMA [31], Flamingo [1], and BLiP2 [32] freeze the language model during training. This prevents knowledge loss and enables the retention of prompt capabilities. On the other hand, approaches like OFA [49] and KOSMOS-1 [15] directly adopt the encoder-decoder structure without additional model components to pursue unified models. However, fine-tuning the language model alone can lead to a decrease in language ability. To address this, both OFA [49] and KOSMOS-1 [15] add language-only tasks during training to prevent the loss of language ability.

In-context Learning. Recent studies have demonstrated that the in-context learning capabilities of language models can be successfully transferred to vision-language-generating models. Frozen [17] exhibits the capability of fast concept binding, enabling the model to associate a new word with a visual category using only a few examples and immediately utilize that word appropriately. While the model performs well in the two-way setting (two new words), this ability fails to transfer to the five-way setting (five new words). Experimental results also indicate that increasing the number of in-context learning samples enhances model performance, but there is a saturation point, and additional repeated content can even lead to a decline in performance. Similar conclusions have been drawn in the case of Flamingo [1]. Both Flamingo [1] and Kosmos-1 [15] demonstrate that employing individual text prompts instead of image-text pairs can improve model performance. However, it is important to note that using individual text prompts can introduce bias to the model [1].

Prompt Tuning. [43] conducted a study on Prompt tuning in generative multimodal models. Their findings indicate that prompt tuning consistently exhibits greater robustness than finetuning across various tasks. The study also highlights the impact of different setups on prompting performance, revealing that longer prompts with more parameters can facilitate improvements. However, there is a diminishing marginal utility, and excessively long prompts may even have a detrimental effect on performance. Furthermore, the results suggest that inserting prompts at the bottom layers might lead to better performance.

3.5 Application of Prompting

Prompting has been widely adopted in many vision-language tasks evolving text generation, demonstrated promising results, and inspired a new learning paradigm, *i.e.*, in-context learning.

Visual Question Answering. The goal of visual question answering (VQA) is to train models to understand the information in an image and answer questions about it in natural language. In-context prompts show surprising results in few-shot [1, 16, 17, 15] and zero-shot scenarios [29, 30, 1, 15]. Some work also applies prompts to web page question answering [15] and grounded question answering [28]. Web page question answering aims to find answers to questions from web pages which requires comprehension of both the semantics and structures of texts. Huang *et al.* [15] uses the template prompt *"Given the context below from the web page, extract the answer from the given text like this: Question: Who is the publisher of this book? Answer: Penguin Books Ltd. Context:* $\{$*WebText*$\}$ *Q:* $\{$*question*$\}$ *A:* $\{$*answer*$\}$*"* where $\{$*WebText*$\}$ stands for the text extracted from web pages. Grounded question answering is firstly designed to reflect the strong transferability of the One For All (OFA) model [28]. In this task, the model should answer a question about a certain region, and special region tokens for hard prompts are designed.

Visual Commonsense Reasoning. This task requires an understanding of the properties of everyday objects in the real world, such as object size reasoning and object color reasoning[15]. The model is required to predict the size or color relation between The Kosmos model [15] uses example prompts like *Is* $\{$*Item1*$\}$ *larger than* $\{$*Item2*$\}$*?* $\{$*Answer*$\}$ and *The color of* $\{$*Object*$\}$ *is?* $\{$*Answer*$\}$ in the zero-shot scenarios and achieves promising results.

Zero-shot Image Classification. Prompting combined with large-pre-trained multimodal models has shown great transferability on out-of-domain test data such as zero-shot image classification.

Kosmos [15] concatenates the input image with a prompt like *The photo of the* and lets the model complete the prompt sentence with the predicted class. Besides, to incorporate additional rules in the classification, Kosmos also sends class descriptions along with prompts to prompt the model for a specific category.

Image Captioning. Generating descriptions given an image is a typical multimodal-to-text generation task that requires the comprehension of both vision and language information. Prompts are used mostly in few-shot and zero-shot scenarios and demonstrate powerful capacity. Flamingo [1] and PaLI [30] adopt prompts to generate image captions in few-shot settings. For example, PaLI [30] uses the prompt *Generate the alt_text in EN* to generate image captions. Prompts are also studied in zero-shot settings, such as BLIP-2[32], MAGMA [31], SimVLM [29], and OFA [28].

Chatbot. The advent of chatbots such as ChatGPT [50] is one of the most remarkable breakthroughs in AI research. Following work such as Visual ChatGPT [12] and GPT4 [18] extend chatbots to multimodal applications which support both images and text prompts. Visual ChatGPT [12] is built based on ChatGPT and visual foundation models. It uses a Prompt Manager which specifies input-output formats, converts visual information to language format, and handles histories of different visual foundation models. GPT4 [18] is able to accept prompts consisting of both images and texts, which lets users specify any vision and language task by generating text outputs given arbitrarily interlaced text and image prompts. Besides, some work migrates GPT to a specific domain such as BiomedGPT on biomedical research [51].

3.6 Responsible AI Considerations of Prompting

Language-based VLMs inherit the risks of the underlying LLMs and vision models, such as gender and racial biases when prompted with images [52]. Several surveys on the ethics of LLMs are available [52, 53]. Some work studies the robustness of VLMs against both natural distribution shifts [54] and adversarial robustness [55]. A recent study [56] investigates the robustness of prompt tuning on VLMs against natural distribution shifts. Moreover, [57] proposes robust prompt tuning on VLMs by integrating multiple-scale image features into the prompt.

4 Prompting Model in Image-Text Matching

4.1 Preliminary of Image-Text Matching Models

Matching-based VLMs have introduced a novel training paradigm that facilitates the acquisition of joint multi-modal representations. Prominent models in this field, such as CLIP [2], ALIGN [58], ALBEF [58] and Multi-Event CLIP [59], leverage contrastive learning techniques to achieve joint representations for images and texts with a learning objective that aims to bring the representation of an image-text pair closer together while pushing non-pairs further apart.

By expanding training datasets [2] and scaling up the model parameter, matching-based models exhibit adaptability across a broad spectrum of downstream tasks, including zero-shot benchmarks and fine-tuning scenarios.

Depending on the target of prompting, existing methods can be classified into three categories: *prompting the text encoder*, *prompting the visual encoder*, or *jointly prompting both branches* as shown in Fig. 3. These approaches aim to enhance the flexibility and task-specific performance of VLMs in recent studies.

A classic matching loss is formulated as below to align the image and text embeddings with an Image-to-Text loss $\mathcal{L}_{i2t}$ and a Text-to-Image loss $\mathcal{L}_{t2i}$.

$$\mathcal{L}_{i2t} = -\frac{1}{N}\sum_{i}^{N}\log\left(\frac{\exp \mathrm{sim}(f_v^l(v_i), f_t^l(t_i))}{\sum_j^N \exp \mathrm{sim}(f_v^l(v_i), f_t^l(t_j)}\right) \quad (4)$$

$$\mathcal{L}_{t2i} = -\frac{1}{N}\sum_{i}^{N}\log\left(\frac{\exp \mathrm{sim}(f_t^l(t_i), f_v^l(v_i))}{\sum_j^N \exp \mathrm{sim}(f_t^l(t_i), f_v^l(v_j)}\right) \quad (5)$$

By prompting, we substitute model input by following learnable prompts: for textual prompts: $f_v^l([\{v_i\}_{i=1}^M, \{z_i\}_{i=1}^M])$ and for the general visual prompt: $f_t^l([\{c_i\}_{i=1}^M, \{t_i\}_{i=1}^M])$ where M is the number of prompts we use.

4.2 Prompting Text Encoder of VLM

Prompting language models has been long studied. As discussed in Sec. 3.2, we categorize prompts into hard prompts and soft prompts in this section. As shown in Fig. 3(a), learnable textual prompts are optimized on image-text pairs in a supervised manner. Recent works [60] also investigate a different scenario with unlabelled data. In this section, we will delve into the details of soft prompts, exploring different types, such as global prompts, task-specific prompts, and instance-specific prompts.

4.2.1 Hard prompt

Prompting language models within the context of VLMs has been extensively investigated. The introduction of prompts has played a pivotal role in discovering and utilizing large-scale pre-trained Language Models. Textual prompts mitigate handcrafted text templates (*e.g.,* ,"*a photo of a [CAT]*"), which enables the model to understand and respond to specific tasks without requiring explicit task-specific training, showcasing the flexibility and versatility of the model. [2] utilize hard prompts to test its zero-shot performance on several tasks. Hard prompts demand significant expertise in the domain and often involve high costs. This has given rise to a new learning paradigm as carefully refining prompts to optimize performance.

4.2.2 Soft prompt

The task of selecting an appropriate prompt is a complex endeavor that demands experience and domain expertise, and significantly impacts model performance. This raises an important question: can we dynamically 'search' for optimal prompts using gradient-descent-based learning methods? Soft prompts refer to prompts that incorporate learnable parameters within their design. We categorize soft prompts into three main types: global soft prompts, group-specific prompts, and instance-specific prompts.

Global Soft Prompt. A straightforward yet powerful approach to adapting language models for downstream tasks involves modifying template tokens specifically for those tasks. Studies such as [61, 62, 63] have employed learnable prompts as input token embeddings when dealing with new tasks. Compared to fine-tuning the entire model, learning a small set of prompt embedding parameters proves to be more parameter-efficient and data-efficient. These prompts, referred to as "global soft prompts," are utilized consistently across all instances within a given task. The term "global" signifies their universal usage throughout the task, enabling the model to generalize and perform well across inputs.

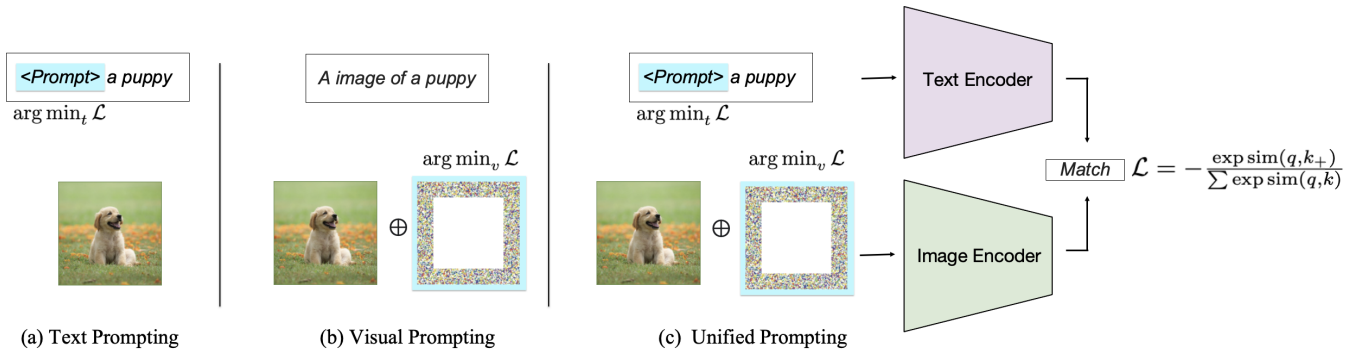


Fig. 3: Prompting tuning on Image-Text Matching VLMs can be applied in different branches: (a) Text prompting, (b) Visual prompting, and (c) Unified prompting, on the input data. Light blue boxes denote learnable parameters. A matching loss is employed to optimize over a small number of learnable parameters, in the loss formulation q denotes the querying modality and k denotes the target modality.

Group-specific Prompt. Several recent studies [63, 64, 65] have employed a group of soft prompts specifically tailored to adapt to different tasks or types of inputs. These models enable the models to query and select appropriate prompts dynamically. [63, 64, 65] use a group of soft prompts targeted to adapt different tasks/types. Different prompts are queried based on the input data. CoOp [63] finds that using different context prompts for classes (class-specific context) can enhance performance in fine-grained classification. [64] uses task-specific prompts to adapt CLIP on a wide range of video understanding tasks. [65] proposes to use MVLPT with different task prompts for source and target tasks to share knowledge across task-specific prompts.

Instance-specific Prompt. While effective in some cases, task-grouped prompts can suffer from overfitting issues and may struggle to adapt to unseen classes or novel samples. In contrast, instance-specific prompts aim to customize prompts for individual samples, allowing for a more personalized and adaptive approach. CoCoOp [66] is a model that adopts instance-adaptive prompts, specifically instance-specific prompts, instead of relying solely on global prompts. This approach has been shown to enhance the generability of the model.

4.3 Prompting Image Encoder of VLM

In line with the achievements of prompt tuning in Natural Language Processing, there have been endeavors to extend the concept to visual inputs. According to the way of designing visual prompts, we categorize them into two classes: patch-wise prompts, where prompts are added as visual patches that are prepended to the original images, and annotation prompts which involve annotating prompts directly on the raw images

Patch-wise Prompts. Adding learnable patches as visual prompts is an intuitive method to incorporate visual cues into VLMs. Just as textual soft prompts serve as input tokens, [67] introduces Visual Prompt Tuning (VPT), which learns a small set of visual prompts as visual patches. These patches are concatenated with input images to adapt pre-trained models to new tasks. VPT investigates visual prompts in the input and latent layers and outperforms most other adaptation methods like full fine-tuning. In a similar vein, [10] explores the use of visual perturbation as a visual prompting technique. Through adversarial reprogramming, the model learns to add visual prompts to input images. Additionally, [65] adopts patchified visual tokens as learnable input embeddings. [68] proposes applying normalized visual prompts to augmented images, unleashing the potential of visual prompting in diverse data settings. In terms of promoting diversity in prompts, [69] employs different visual prompts for distinct subsets of data.

Annotation Prompts. Visual prompting can also be performed explicitly by directly manipulating images, similar to the process of annotation, which we term annotation prompts. Colorful Prompt Tuning (CPT) introduced in [70] focuses on colorizing specific regions of images as visual prompts. By incorporating color cues, the model is guided to ground objects and better understand the visual context. [71] explores the use of annotations, such as red circles, as an innovative visual prompting design. These annotations serve as cues to guide the model's attention toward specific areas of interest, thereby enhancing its understanding of images. The study delves into CLIP's emergent ability to comprehend images through the clever use of visual prompts. Furthermore, [72] proposes the use of example input and output images as visual prompts. By providing a pair of images demonstrating a desired task, such as image inpainting, edge detection, or image colorization, the model is guided to complete similar tasks based on the provided examples. These studies demonstrate the creative and effective use of explicit visual prompting methods and offer a practical and interpretable approach to improving the model's performance.

4.4 Unified Prompting on VLM

As prompt engineering continues to advance in both the vision and language branches, there has been a recent development in joint prompting. This approach aims to enhance matching-based VLMs by leveraging prompts from both the visual and language domains. As in Fig. 3, learnable prompts in both branches are optimized. According to whether the visual prompt and textual prompt are independent of each other, they can be categorized into coupled and decoupled unified prompting, respectively, in out follow-up discussion.

Coupled Unified Prompting. UPT [49] issues that prompting single modality does not fit all cases: textual prompts may struggle to handle data with high intra-class visual variance, while visual prompts may struggle with data exhibiting high inter-class visual variance. Thus it employs a tiny neural network to optimize

learnable textual and visual prompts jointly and finds unified prompting outperforms any unimodal prompting.

Decoupled Unified Prompting. [65] employs soft textual prompts and VPT-like visual prompts on both the language and vision branches. This approach leverages the benefits of prompt engineering in both modalities [70] introduces an innovative approach that combines visual and textual sub-prompts for visual grounding tasks. By utilizing regions in the image as visual prompts and phrases in a sentence as textual sub-prompts, the model can establish co-reference across different modalities. [73] introduces MaPLe hierarchical prompts on both branches and synergizes prompt training in both modalities via a Vision-Language coupling function. By leveraging the strengths of both visual and textual prompts, joint prompting contributes to more effective and versatile VLMs that excel in multimodal understanding.

4.5 Application of Prompting

Prompting matching-based VLMs offers the promise of transferring representations learned by pre-trained models to downstream domains and niche tasks like pure-vision tasks including image/video classification, semantic segmentation, relation detection, and multimodal tasks.

Image Classification. Image classification is extensively researched in computer vision for many years. A new approach to object classification has been suggested by prompting the text encoders in VLMs. A Naïve solution to image classification is using a fixed prompt like "`An image of [CLASS]`" as in CLIP [2] and TPT [62], which uncovers pre-trained capacities of zero-shot classification performance. Accordingly, learnable prompts are adapted to image classification in works like [62] Prompt engineering also shows its efficacy in more challenging classification tasks like long-tailed classifcation [74], multi-label classification [75, 76].

Text Classification. Text classification appears to present a dual challenge akin to that of image classification. Due to the scope of this survey, we only cover works that focus on text classification of VLMs. [77] uses visual prompts concerning different classes to better leverage visual information for text classification.

Object Detection. Object detection is aimed at predicting class labels of object bounding boxes in an image. With abundant information on classes in texts, prompt engineering is also used for multi-label recognition. [76] proposes Dual Context Optimization (DualCoOp) using labels as a part of prompts and learning positive and negative prompt pairs to align images and prompts to solve multi-label recognition tasks. [75] proposed Texts-as-Images (TaI) prompting for multi-label detection. Open-vocabulary object classification is a promising application of prompt engineering in object classification, where the detectors can predict new classes that are not in training. ViLD [78] generate a fixed prompt template, *e.g.,* "*a photo of [CATEGORY] in the scene*". [79] introduces detection prompt (DetPro) to learn continuous prompt representations. PromptDet[80] uses regional prompt learning to align region features and text features.

Visual Relation Detection. Visual relation detection is a computer vision task that targets extracting relations between objects in an image. Prompt tuning boosts visual relation detection with its powerful commonsense knowledge contained in LLMs. [81] optimizes a small continuous task-specific vector for visual relation detection. [82] pre-trains VLMs with a matching-based strategy to align image regions and dense captions and fine-tune a decoder with soft prompts to generate relation predictions. [83] presents a Relation Prompt for video open-vocabulary relation detection by generating subject-object sensitive prompts based on object motion cues.

Semantic Segmentation. Semantic segmentation is a classic computer vision task with the goal of assigning each pixel to a class label. DenseCLIP [84] converts image-text matching to pixel-text matching to enable a pixel-wise dense prediction including semantic segmentation task; class-conditioned text prompts are used to contextualize visual cues in texts. Segment anything [85] presents a large-scale foundation model for segmentation which takes images and promptable segmentation queries as inputs.

Not only on a single task but prompting has also been proven to be beneficial for domain adaptation and generalization of pre-trained models. Further studies investigate prompt tuning on pre-trained model transferability under distribution shift.

Domain Adaptation. Prompt learning also enables continual learning of pre-trained models in tasks like test-time domain adaptation, which aims to adapt models to unlabeled test data under a distribution shift. [86] attempts to embed domain information discrepancy in domain-specif textual prompts and can preserve semantic features of pre-trained VLMs. [87] adds prompts to different stages of ViT and fine-tunes prompts in the unlabelled target domain.

Continual Learning. Continual learning is aimed at tackling catastrophic forgetting in non-stationary data distribution. Prompt tuning becomes a new methodology for continual learning. Learning to Prompt (L2P) [88] shows a brand new prompt-based approach for continual learning by querying trainable task-specific prompts from a prompt pool for each input instance and prepends it to input before pre-trained models to instruct the model. [89] presents DualPrompt, to learn task-invariant and task-specific instructions across tasks, unlike that L2P uses only one prompt tool and calls them General and Expert prompt space.

Domain Generalization Domain generalization targets adapting models to unseen domains in the training stage. [90] encapsulates domain-specific knowledge in domain prompts generated by a prompt adapter and prepends it with input data; at test-time, prompts are generated based on the similarities between domains.

4.6 Responsible AI Considerations of Prompting

The importance of AI Integrity and ethics has been attached to prompting matching-based VLMs to construct trustworthy multimodal models. The discussion covers model robustness, safety, fairness, bias, privacy, and the like.

Adversarial Robustness of Prompt. Robustness analysis evaluates the performance of the model under different conditions and perturbations. [91] studies how VPT and fine-tuning improve zero-shot robustness under adversarial attack on CLIP and finds that VPT is more effective in the absence of texts. [92] also attempt to leverage universal visual prompting to improve the adversarial robustness at test time. Visual prompting is more flexible compared to conventional adversarial defenses, as it allows universal (*i.e.*, data-agnostic) input prompting templates, which are capable of plug-and-play during testing. [93, 94] investigate the reasons behind VLMs' robustness to natural distribution shifts systematically and reveals that diverse training data is the primary reason for robustness gain. [95] explores the model vulnerability that injecting triggers brings to pre-trained models in prompt tuning.

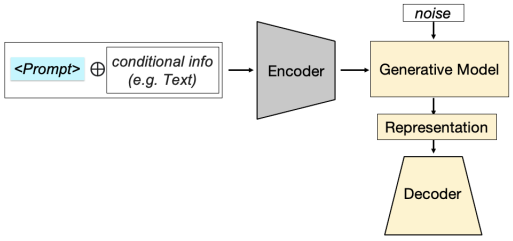


Fig. 4: Prompting Generative Models. A depiction of a typical text-to-image generation framework, detailing elements such as conditional information, an image encoder, a generative model, noise injection, latent space representation, and a decoder. Conditional information can take various forms such as hard prompts, learnable soft prompts, or a combination of the two. Furthermore, prompts can be presented in textual, visual, or both formats.

Backdoor Attack of Prompt Learning. [96] studies on the backdoor and poisoning attacks on CLIP and find CLIP trained on manually labeled data suffer badly from such attacks. It shows that the training on noise and uncurated datasets makes backdoor and poisoning attacks a significant threat. [97] proposes a new backdoor attack method named BadEncoder on CLIP and exposes this threat to VLMs. Once a pre-trained image encoder has been injected backdoors, the downstream classifiers built on it for different downstream tasks simultaneously inherit the backdoor behavior. Given such vulnerability to backdoor attacks, CleanCLIP [98] is proposed as a finetuning framework that weakens the learned spurious associations introduced by backdoor attacks.

Fairness and Bias. Social bias is an important topic in a fair AI system. A wide range of works have studied different aspects of biases. [99] showcases an analysis of bias regarding race and gender misclassification in the CLIP model. In the meantime, many existing works focus on de-biasing the model. In particular, [100] attempts to alleviate bias by calibrating the biased prompted texts to debiased content while [101] proposes to mitigate biased results in image retrieval tasks by post-processing of the VLMs output. In addition, [102] introduces a new dataset debiasing pipeline to augment the dataset with healthy data.

5 Prompting Model in Text-Image Generation

5.1 Preliminary of Text-Image Generation Models

This section provides an overview of the preliminaries required to understand the prompting model in text-image generation, with a specific focus on diffusion models.

Text-image generation automatically synthesis vivid and realistic images from natural language descriptions and has attracted much more attention. From the pioneering work DRAW [103], text-image generation models have seen numerous breakthroughs. Generative adversarial network (GAN) [104] then used to design end-to-end differentiable image generation structure [105] which is followed by many works [106, 107, 108]. Besides, variational auto-encoder (VAE) [109] is also adapted to generate images [110, 111]. However, these models are trained on small-scale data and lack generalization [112]. Autoregressive methods driven by large-scale datasets, such as DALL-E [112], and Parti [113], are proposed and demonstrate surprising zero-shot generation ability. Recently, the diffusion model (DM) has spurred another line of state-of-the-art models for text-image generation [114]. Diffusion models, also known as diffusion probabilistic models [115], originate from non-equilibrium statistical physics [116] and sequential Monte Carlo [117] and are designed to fit any data distribution while keeping tractable. The denoising diffusion probabilistic models (DDPMs) [118] first adopt DMs in the image generation domain and inspire the whole community of generative models. In inference, DDPMs build a Markov chain that generates images from noisy data within finite transitions which is called *reverse process*. In training, DDPMs learn from the *forward process* where noise is added to the natural images and estimated by the model. Given a clean image x_0 from a distribution q, diffusion step T and hyperparameters β_t, the forward process generates x_T following

$$q(x_{1:T}|x_0) \coloneqq \prod_{t=1}^{T} q\left(x_t \mid x_{t-1}\right), \tag{6}$$

$$q\left(x_t \mid x_{t-1}\right) \coloneqq \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I\right). \tag{7}$$

The noised image from any arbitrary step t then can be reformulated as

$$q\left(x_t \mid x_0\right) \coloneqq \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}x_0, \left(1-\bar{\alpha}_t\right) I\right), \tag{8}$$

where $\alpha_t \coloneqq 1-\beta_t, \bar{\alpha_t} \coloneqq \prod_{s=0}^{t} \alpha_s$.

Given the defined forward process, the DDPM is trained in the reverse process which starts from $p_\theta(x_T)$ by the loss defined as

$$L(\theta) \coloneqq \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t\right)\right\|^2\right], \tag{9}$$

where t is uniform between 1 and T, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is random noise, and $\boldsymbol{\epsilon}_\theta$ is known as noise predictor parametrized by θ.

By incorporating additional control information, typically in the form of textual prompts, the efficacy of the reverse process in diffusion models has been significantly enhanced to control the synthesis results rather than random sampling. This textual-based generation has solidified its position as the pioneering foundation in the field of text-to-image generation. Consequently, let Γ be an encoder that maps a conditioning input prompt P into a conditioning vector $\boldsymbol{c} \coloneqq \Gamma(P)$, the conditioned learning objective has been expanded with $\boldsymbol{c}$ that represents textual prompt.

$$L(\theta) \coloneqq \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon},\boldsymbol{c}}\left[\left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t, \boldsymbol{c}\right)\right\|^2\right], \tag{10}$$

Figure 2 illustrates a typical text-image generation framework, highlighting its key components and functionalities, including (1) fixed or learnable conditional information, such as hard textual prompts or learnable soft prompts. The conditional information can be in textual form or in other modalities; (2) An encoder $\mathcal{E}$ of the input image; (3) A generative model, such as diffusion model, autoregressive model, or GAN; (4) Noise injection or interference; (5) A representation of features in the latent space or low-resolution images; (6) A decoder $\mathcal{D}$ for image decoding or super-resolution for a high-fidelity generation. The training process involves dataset utilization, loss functions, and optimization techniques to train the model for generating coherent and visually appealing images based on text prompts. During the inference stage, the trained model is utilized to generate images based on user-specified prompts. The formulation of prompts plays a crucial role as it governs communication with the model and influences the desired outcomes of image generation. This section focuses on prompt engineering in text-image generation and its applications.

5.2 Understanding Prompting

To gain a deeper understanding of the factors that influence the generated images, we will introduce prompt design in text-to-image from the view of semantics, prompt diversity, and controllable prompts.

Semantic Prompt Design. The art of prompt semantics has significant impacts on image generation in diffusion models [119]. The linguistic components such as adjectives, nouns, and proper nouns in prompt influence image generation in different ways but consistently. While descriptors (simple adjectives) subtly affect the output, nouns introduce new content more effectively. Interestingly, using an artist's name tends to generate images deviating significantly from the original, and incorporating lighting phrases can dramatically modify image content and mood. Therefore, the quality of image generation can be enhanced through clear, noun-based statements, effective seeds, and the emulation of artist styles.

Diversify Generation with Prompt. Apart from direct handcrafting individual prompts in a semantic way, recent works experiment with various prompt modifiers $\mathcal{M}$ focusing on enhancing the diversity of initial prompts P by $\tilde{P} = \mathcal{M}(P)$ with $\tilde{P}$ be the diversified prompts. DiffuMask [120] explores two strategies in prompt modifiers $\mathcal{M}$, *i.e.*, retrieval-based prompt and prompt with Sub-class, with P set to *"Photo of a [sub-class] car in the street"*. Specifically, they retrieve real images and captions sets [121, 2], with captions as the prompt sets for generating synthetic images. Besides, they select sub-classes from Wiki based on the main class. ImaginaryNet [122] uses GPT2 [36] as $\mathcal{M}$ with a given class name y of the target object to generate a complete description of an imaginary scene $\tilde{P}_y$ under the guidance of prefix phrases of *"A photo of a"*. The prompt serves as generating diversified photo-realistic imaginary images for the imaginary supervised object detection task. Similarly, [123] uses a word-to-sentence T5 model [22] as $\mathcal{M}$ to generate detailed prompts $\tilde{P}_y$ targeted for a specific label space y, thereby maximizing the potential of synthesized data in data-scarce settings by enriching the diversity of prompts. These approaches further obtain diversified images I by $I = \mathcal{G}(\boldsymbol{\epsilon}|\tilde{P})$ where $\mathcal{G}$ represents the generative model.

Complex Control of Synthesis Results. As the synthesized image generation is usually inconsistent due to noise injection and randomness lying in the stochastic nature of diffusion models, recent work has been emerging in the area of complexly controllable generation. To avoid controllability limitations with user-provided masks that restrict the modified area [124, 125], prompt-based control is gaining attention. OneWord [126] aims to solve the problem of generating personalized images with specific subjects that are hard to describe with pure texts. Therefore they proposed a prompt method that designates a placeholder string S_* to represent the new concept such as *"a photograph of S_* on the beach"* with its associated learned embedding v_*. A similar design is done by DreamBooth [127]. Instead of creating new words, they design prompts with *(unique identifier, subject)* pairs that bind rare tokens from T5-XXL tokenizer [22] as unique identifiers for the specific subjects and the coarse class name of the subjects, such as *"A [V] dog"*, with *[V]* as the rare-token identifiers. They further retain the representation of class names in prompts by introducing extended class-prior preservation loss to the training objective. Custom Diffusion [128] extended the customization into a multi-concept scenario where multiple personalized concepts are composed in the same generated image, such as family members in the same family photo. They design prompts at this aim by including the use of a unique modifier token S_i^* for each concept i, initialized with different rare tokens and positioned ahead of category namex.

Over the days, textual prompts only cannot meet the specific needs of image-processing tasks, and controllable text-to-image generation is gaining attention [129, 130, 131]. A wide range of task-specific input conditions, such as canning edge encoded by image encoder [132], are added with trainable network architecture to the diffusion model in the work of ControlNet [133]. The additional task-specific conditions $\boldsymbol{c}_\mathrm{f}$ is added to the overall training objective as $L(\theta) \coloneqq \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{\epsilon},\boldsymbol{c},\boldsymbol{c}_\mathrm{f}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\boldsymbol{x}_t, t, \boldsymbol{c}, \boldsymbol{c}_\mathrm{f}\right)\right\|_2^2]$. Notably, in order to improve the semantic recognition ability of the encoder from control maps and optimize ControlNet's performance even when explicit prompts are absent, ControlNet's training utilizes a method where half of the text prompts are randomly replaced with empty strings.

Controlling the synthesis results can also be done after the generation process with prompt editing methods. To bypass the common demand of user-defined spatially fixed masks [124, 125], Prompt-to-Prompt [134] can edit images by only editing prompts by replacing a word, specifying a style, changing adjectives, *etc*. The manipulations are infiltrated by injecting the cross-attention maps controlling which pixels attend to which tokens of the prompt text during which diffusion steps. Prompt-based image editing methods that merely modify the text prompt provide more intuitive editing experiences.

5.3 Application of Prompting

Text-to-image diffusion models, aided by prompting techniques, have excelled in data generation applications. This section investigates their efficacy in generating training data that boost the scope and flexibility of learning procedures. Additionally, we explore the versatility of these models in crafting diverse data in target domains, spanning diverse output formats like images, videos, 3D models, and motion. Also, we unveil its potential in complex task-solving and adversarial attacks.

5.3.1 Generating Synthetic Training Data

Recent advancements have sparked a growing interest in prompting text-to-image models as innovative synthesized training data generators for various tasks downstream tasks such as segmentation, object detection, and image recognition. Challenges such as data scarcity and the need for high-resolution synthetic images can be mitigated through intricate prompt engineering. DiffuMask [120] automatically generates high-resolution synthetic training images with the aforementioned prompt engineering strategies in Sec. 5.2. Its created pixel-level semantic masks between prompts and generated images can be seamlessly applied for segmentation tasks, including semantic segmentation, open-vocabulary segmentation, and domain generalization on real images. ImaginaryNet [122] generates synthesis data to tackle the challenge of insufficient real images and annotations for training object detection. It generates scene descriptions with LLM from class labels and prompts the text-to-image model for creating imaginary training data. Under different training settings of pure or mixed imaginary and real data, object detectors are enhanced for the Imaginary Supervised Object Detection task (ISOD), especially under settings where real images and annotations are unavailable. Synthetic data is also proven feasible for image recognition tasks, specifically in zero-shot and few-shot settings. [123] creates the synthetic data for image recognition in a two-phase manner. Firstly, novel samples are synthesized using target

category names. Secondly, a fine-tuned language model is used to convert category names into richly contextual, diversified language prompts for diversifying the training data.

5.3.2 Generating Data in Target Domain

In addition to the role of training data generators, diffusion models also play a pivotal role as target data generators. Importantly, their capabilities extend beyond the generation of images. They can efficiently generate video data, three-dimensional data, and motion data, further broadening their application range and utility.

Text-to-Video Generation. Make-A-Video from [135] is the first approach for directly translating the tremendous recent progress in text-to-image (T2I) generation to text-to-video (T2V) without paired text-video data. It infers actions and events in the prompt and generates video by leveraging joint text-image priors to bypass the need for paired text-video data. Imagen Video [136] propels T2V generation towards a more efficient stage, delivering higher video resolution outputs by combining a frozen T5 text encoder [22], a base video diffusion model, and interleaved spatial and temporal super-resolution diffusion models, *i.e.*, cascaded diffusion models [137]. However, works on T2V commonly face challenges in editing capabilities and effective training on specific domains. FateZero [138] overcomes these limitations with a zero-shot text-based editing method capable of editing attributes, style, and shape on real-world videos without per-prompt training or use-specific mask. Specifically, FateZero utilizes a pair of user-provided source prompt P_{src} and the editing prompt P_{edit}. The source prompt is for obtaining a noisy latent representation $\boldsymbol{x}_t$ of the source video frame, then $\boldsymbol{x}_t$ is denoised conditioned on the editing prompt P_{edit}. Tune-A-Video [139] tackles the challenge of computational expensiveness with the one-shot tuning strategy on one text-video pair and only on the first and former video frames. This study is in the inspiration that T2I models attend well to verbs in the prompt in generating still images and exhibit surprisingly good motion consistency alignment with prompts when extended to T2V. Tune-A-Video is also equipped with editing capability by capturing essential motion information from the input video and synthesizing novel videos with edited prompts preserving the motion words. Moreover, textual prompt-based generation has advanced to multi-modal generation, *e.g.,* generating simultaneously aligned audio-video pairs [140, 141].

Text-to-3D Generation. Previous works face challenges of insufficient large-scale labeled 3D datasets and inefficient architectures for denoising 3D data. As a consequence, prompt-based generation has advanced from T2I to T2V models and also in text-to-3D scenarios where high-quality 3D objects and scenes are generated from text prompts [142]. DreamFusion [143] firstly randomly initializes the 3D object with NeRF [144] for each text prompt and produces 2D image renderings $x = g(\eta)$ with differentiable image generator $g(\eta)$. These renderings are generated from various angles and paired with view-dependent prompts prefixes such as *"overhead view"* and *"front view"* and then diffused and reconstructed by Imagen [136] with $q\left(x_t \mid x_0\right) \coloneqq q\left(g(\eta)_t \mid g(\eta)_0\right)$. The sampled noise $\boldsymbol{\epsilon}$ guides a gradient direction to be backpropagated to the NeRF parameters η. To tackle the issue in the growing popular DreamFusion regarding the optimization efficiency of NeRF which leads to low-quality 3D models with a long processing time, Magic3D [145] proposed a two-phase coarse-to-fine optimization framework, *i.e.*, firstly obtaining coarse diffusion prior from text prompts with Imagen [136] and then rendering efficiently with high-resolution latent diffusion models (LDM) [3]. Borrowing the idea from [127], Magic3D is capable of personalized prompt-based editing of 3D models by binding the *[V]* identifier in the prompt with the 3D object. Besides, prompt-based editing can be done through finetuning with LDM in the coarse-to-fine stage with the modified prompt. Inaccurate and unfaithful structures in text-to-3D generation due to random shape initialization without prior knowledge lead Dream3D [146] to explicit 3D shape priors into the CLIP-guided 3D optimization process [143, 147, 148]. Specifically, it connects the T2I model and a shape generator as the text-to-shape stage to produce a 3D shape prior with shape components in the prompts. Then it harnesses the 3D shape prior to the initialization of NeRF and optimizes it with the full prompt. To close the gap between the synthesis image and shape, and also inspired by [126, 127], Dream3D links renderings with stylized text prompt suffixes in the format of *"a CLS in the style of* $*$*"* where *CLS* represents the shape category and $*$ is a placeholder token that requires optimization of its text embedding jointly with the weights of Stable Diffusion for capturing the style of the rendered images.

Text-to-Motion Generation. Another area where the power of prompt-based generation is exemplified is in the realm of text-to-motion (T2M). MotionDiffuse [149] is a diffusion model-based text-driven motion generation framework with motion sequence as the input $\mathbf{x}_0$. It has a body part-independent controlling scheme that generates separate sequences for each body part under m fine-grained prompts P_i with $i \in [1, m]$ for each body part i and predicts each $\epsilon_i^{part} = \epsilon_\theta\left(\mathbf{x}_t, t, \Gamma(P_i)\right)$. Besides, it generates arbitrary-length continuous motion synthesis using time-varied text prompts with m intervals, denoted as array $\{P_{i,j}, [l_{i,j}, r_{i,j}]\}$ and predicts the ϵ_i^{time}. All noises are interpolated mutually with other parts for the continuous motion sequence generation. Similarly, as in T2I, T2V, and text-to-3D, it is also required for T2M synthesis with flexible editing capability. Thus, FLAME [150] enables editing with free-form language description with novel transformer-based diffusion architecture. It takes diffusion time-step tokens, motion length tokens, language tokens, and motion tokens as input tokens to the transformer and can therefore handle motion sequences of variable length. MDM [151] also introduces editability and controllability with a similar idea borrowed from image inpainting by adding suffixes and prefixes to the motion in the temporal domain. And the textual condition guides MDM to fill the missing body part with a specific motion while keeping the rest intact in the spatial domain.

Complex Conditional Scene Generation. The use of diffusion models has expanded beyond single target data generation, finding applications in various scenarios that involve generating more complex scenes tailored to specific use cases with more complex conditional inputs. In robotics, text guidance is used to perform aggressive data augmentation on top of our existing robotic manipulation datasets to generate robotic scenes via inpainting various unseen objects for manipulation, backgrounds, and distractors [152]. In autonomous driving, diffusion models are leveraged to generate controllable pedestrian trajectories that align with the surrounding environment's context that enables the simulation of realistic pedestrian behavior [153]. Additionally, diffusion models can incorporate conditional information in the form of graphs that represent individual rooms to generate house floorplans, facilitating the design and planning of residential spaces [154].

5.3.3 Prompt-centered Complex Task

Beyond the former direct applications of text-to-other generation, prompt-centered complex application in various scenarios reveals the field's true versatility and potential. In the context of storytelling, StoryBook [155] retains a visual narrative storybook with consistent character faces through a series of prompt-centered steps. It first generates prompts of scene descriptions with LLM, which are prompted to the latent diffusion model with designated special token placeholder S_* like [126], to ground consistent character faces during generation. Similarly, [156] proposed multimodal procedure planning (MPP) task, where the initial stepwise textual plan is generated with LLM and then serves as prompts to diffusion model for synthesizing text-grounded image plan. What's different is that the image plans are verbalized through image captioning backward to the LLM for revising the initial plan showing the potential for multimodal prompting.

5.4 Responsible AI Considerations of Prompting

Artificial Intelligence is revolutionizing our world through its formidable learning ability, transformative force, and profound influence across diverse areas of society. It also spurred intense debate about ethical issues, principles, and integrity in AI development and applications. There is a global convergence around five ethical principles [157]: transparency, justice and fairness, non-maleficence, responsibility, and privacy. In this subsection, we discuss ethical issues when prompting text-to-image generative models.

Adversarial Robustness of Prompt. The adversarial attacks have been introduced to text-to-image diffusion models for mainly 2 aims. Some work takes diffusion models as a tool to facilitate or defend against adversarial attacks [158, 159]. Some work directly attacks diffusion models [160] and aims to erase image content given character perturbations. As the pioneer to introduce diffusion models in the adversarial attack field, DiffAttack [158] unveils the potential of diffusion models for crafting adversarial examples with satisfactory imperceptibility and transferability by manipulating the latent space rather than pixel space. This approach maintains visual quality with embedding perturbations undetectable to humans and transferable across diverse model architectures. Diffusion models can be utilized for adversarial purification - a defense strategy that removes adversarial perturbations. DiffPure [159] implements this approach, adding a minimal amount of noise to an adversarial example before reversing the generative process to restore the original image, thus exhibiting robust defense capabilities against powerful adaptive attacks. Zhuang *et al.* [160] study the query-free attack generation on Stable Diffusions where an adversarial text prompt is obtained in the absence of end-to-end model queries. They show the vulnerability of Stable Diffusions rooted in the text encoders. A five-character text perturbation is able to shift the output content.

Backdoor Attack of Prompt Learning. Backdoor attacks on text-to-image generative models aim to control the content of generated images during inference by embedding inputs with predefined backdoor triggers. The attacker secretly injects backdoors, such as specific text characters, into the model during training to trigger the model to either generate images with pre-defined attributes or images following a hidden or even malicious description. The backdoor attack may lead to inappropriate outputs such as offensive content. On the other hand, it can also be used in copyright protection by watermarking the models. Struppek *et al.* [161] demonstrate that the text encoders pose a major tampering risk. The attack is a teacher-student approach and only involves fine-tuning a text encoder by generating backdoor targets and triggers on the fly. Zhai *et al.* [162] design three types of backdoor attacks, namely pixel-backdoor, object-backdoor, and style-backdoor, and demonstrate the text-to-image diffusion models' vulnerability to backdoor attacks. Huang *et al.* [163] explore the vulnerability to backdoor attacks via personalization for a more efficient attack. Text-to-image personalization guides the diffusion-based text-to-image model to generate user-provided novel concepts through natural language. Huang *et al.* [163] devised backdoor attacks on two families of personalization methods, Textual Inversion [126] and DreamBooth [127].

Fairness and Bias. Generative AI models are typically trained on web-scale datasets scraped from the internet and are inevitable to biased human behavior as shown in [164, 165, 166, 167]. For example, Stable Diffusion only generates images with white male-appearing persons as firefighters [164]. Some studies start to pay more attention to the fairness issues related to text-to-image generations and can be grouped into three paradigms: 1) training data pre-processing to remove bias before learning [102, 168], 2) enforcing fairness during training by introducing constraints on the learning objective [168], 3) post-processing approaches to modify the model outcome at the deployment stage [164, 100, 169].

Privacy. There might be privacy-sensitive information, *e.g.,* face identity, in the huge amount of training data for training text-to-image models. Such information may arise privacy risks in real-world applications such as information leaks. Membership inference attacks are an approach to investigating privacy leakage by inferring whether a specific data sample was used in the training phase (called member or non-member respectively) [170]. Some work [170, 171, 172] studies the privacy risks of text-to-image generation models from the perspective of membership attacks. From the perspective of prompting, Shen *et al.* [173] propose *prompt stealing attack*, which steals prompts from images generated by text-to-image generation models. The creation of high-quality prompts can be challenging, time-consuming, and costly. Hence successful prompt stealing attacks direct violate intellectual property and even jeopardize the business model of prompt trading markets.

6 Prompting VLM vs. Uni-modal Models

6.1 Prompting in Natural Language Processing

This section summarizes existing studies on prompt engineering on textual language models. Prompt engineering has been widely adopted in various natural language processing applications including question answering [174, 175], text classification [61, 176], text generation [36, 23, 177], and information extraction [178, 179], *etc*. Recent LLMs such as InstructGPT [180] and PALM2 [181] have shown incredible generalized inference ability through prompting. Early works [182] designed natural language templates to let pre-trained language models fill in to explain their predictions. Wei *et al.* [4] demonstrate that the performance of LLMs can be significantly improved by adding intermediate reasoning steps into the prompt. In particular, the prompt of each task contains a few manual demonstrations consisting of a question and a reasoning chain leading to the answer. The LLM learns to follow the prompt and thinks step-by-step to solve the given task. Liu *et al.* [183] find that the quality of the prompt, *i.e.*, the selection of examples in prompts and given explanations,

largely impacts LLM's performance. Fu *et al.* [184] demonstrate that prompting LLMs with complex example questions, which requires more intermediate reasoning steps, could achieve better performance and benefit the model's robustness regarding format perturbation and distribution shift.

Manually crafting prompts for each task strongly depends on human experience, and manual testing would be required to evaluate and improve the template, which would be time-consuming. Zhang *et al.* [5] work on eliminating manual efforts by leveraging LLMs to construct reasoning chains with demonstrations. Besides, a line of works [185, 186, 187] automates the prompt engineering by utilizing a dense retriever to augment the language models with external resources, which has also been referred to as retrieval-augmented language models. For a given question, the dense retriever retrieves relevant text from a knowledge source and appends it to the language model input. Such language models have recently demonstrated strong performance on knowledge-intensive tasks. [6] propose *prompt tuning* that appends the input embedding layer with extra trainable tokens and learns these tokens through backpropagation on downstream tasks. This opens a direction of learning soft prompts to enhance LLMs.

Many studies demonstrated that LLMs' performance considerably drops as the task complexity increases. A natural way for humans to solve complex tasks is to decompose them into a series of simple subtasks and solve the complex task by completing each simple subtask. A line of works investigated enhancing LLMs' performance on complex tasks by prompting LLMs multi-times, where the LLMs are expected to solve a subtask by each prompt. Press *et al.* [188] examine the capacity of language models to execute compositional reasoning tasks and found that LLMs are good at memorizing facts but do not compose them to answer questions. To narrow the compositionality gap, the authors let LLMs ask themselves follow-up questions, answer the questions, and decide whether they have sufficient information to give the final solution. Kazemi *et al.* [189] propose a backward chaining algorithm to decompose a complex task by starting from the objective and recursively breaking down the complex task into sub-tasks based on rules. Khot *et al.* [190] decompose a complex task into sub-tasks and use sub-task-specific LLMs to solve them, leading to improved performance on a line of textual multi-step reasoning tasks.

Researchers have also noticed ethical and integrity issues related to prompt engineering in NLP. Yang *et al.* [191] propose a prompt-based adversarial attack to compromise NLP models and robustness enhancement techniques. This work indicates that the prompting paradigm has the potential in probing fundamental vulnerabilities of large language models and fine-tuning them for downstream tasks. Dong *et al.* [192] adopt a prompt-based learning approach to automatically generate effective adversarial examples to probe Dialogue State Tracker models. The prompt may inherit the bias in the pre-trained models and [193] review the literature on fairness metrics for pre-trained language models and experimentally evaluate compatibility. Moreover, one can refer to several existing surveys [194, 195, 9] for a more comprehensive review.

6.2 Prompting on Pure Vision Models

Although prompt is first widely adopted in natural language models, many works also utilize prompts in pure vision models [85, 10, 196, 11, 197, 198, 199] and applications including image classification [10, 197, 198, 199], image segmentation [85, 196], depth estimation [196], keypoint detection [196], denoising [196], detaining [196], and image enhancement [196] *etc*.

Several studies have identified two main mechanisms for incorporating prompts into vision models. The first mechanism treats prompting as an adaptation method that facilitates the fine-tuning of pre-trained vision models [10, 200, 197]. The second mechanism utilizes prompts as a module that plays a role in both model pre-training and inference [85, 196, 199].

Pre-trained vision models have significantly improved performance, but their size has also increased drastically, making training and fine-tuning infeasible for most users. To address this issue, adapting pre-trained models to specific tasks in a parameter-efficient way is critical. Many studies have treated prompting as an adaptation method.

Bahng *et al.* [10] use a single visual prompt to adapt a frozen large-scale vision model to a new task. Adaptation approaches such as fine-tuning and linear probes require some level of access to the pre-trained model during both training and testing. However, visual prompting only requires model access during training, making it feasible for some applications [200]. Additionally, Tu *et al.* [198] propose Visual Query Tuning (VQT) to adapt pre-trained Transformers to downstream tasks while keeping the backbone frozen, allowing for more accurate predictions utilizing the intermediate features of a pre-trained model.

As classical fine-tuning methods become more limiting when models are hosted as inference APIs, visual prompt learning is emerging as a potential solution for adapting frozen and cloud-hosted models. Loedeman *et al.* [197] introduce the Prompt Generation Network (PGN), which generates input-dependent visual prompts to facilitate domain adaptation. PGN generates new prompts for every image by combining items from a commonly learned library of tokens. It consists of a lightweight neural network that learns the probability distribution for selecting prompt vectors from a token library.

In addition to using prompts as an adaptation for downstream tasks, some researchers have integrated prompting modules into the entire model to improve pre-training performance and enable more flexible inference. In [85], Kirillov *et al.* introduce the Segment Anything Model (SAM), which aims to build a foundation model for segmentation. Inspired by prompting techniques in NLP, they proposed the *promptable segmentation task* to generate a valid segmentation mask based on any segmentation prompt. The prompt can include spatial or text information that identifies an object in the image, and the output of the corresponding model is a reasonable mask for at least one target object. This promptable segmentation task is used in both pre-training and downstream segmentation tasks.

Painter, presented in [196], is a generalist model that can perform various vision tasks based on given task prompts. It can perform tasks such as semantic segmentation, instance segmentation, depth estimation, keypoint detection, denoising, detailing, and image enhancement, as well as out-of-domain tasks like open-category object segmentation. To address the issue of general-purpose prompt definition, Painter formulates the dense-prediction vision problem as *image inpainting*. This way, input/output paired images from the same task can be used as input to indicate the task the model should perform.

Following the work on Painter, Wang *et al.*propose SegGPT [11], which focuses on the segmentation task and enables

segmentation of everything with a generalist Painter. Zhang *et al.*utilize auxiliary prompts to approach the Generalized Novel Category Discovery (GNCD) setting by proposing a prompt-based Contrastive Affinity Learning (PromptCAL) method [199]. Existing semi-supervised learning methods fail to learn unlabeled data from novel semantic classes, but PromptCAL is discriminative to novel semantic information.

The combination of prompt engineering with visual models has also triggered a line of work focusing on integrity and ethics issues. Chen *et al.* [92] leveraged visual prompting to improve the adversarial robustness of a fixed, pre-trained model at testing time. Li *et al.* [201] explored the benefits of visual prompting in constructing compelling neural network classifiers with differential privacy. However, such studies are still relatively rare and more attention is required.

7 Challenges and Opportunities

Prompting Model in Multimodal-to-Text Generation. In addition to visual and textual modalities, the incorporation of other modalities such as audio and thermal is possible. It is crucial to address the inherent heterogeneity among these modalities, which includes variations in data formats, scales, and structures.

Two notable projects in this domain are Kosmos [15, 46], developed by Microsoft, and IMAGEBIND [202], developed by Meta AI. These projects aim to create unified models capable of handling diverse modalities, promoting the utilization of such unified models as a significant direction in the field.

However, it is important to note that most of the research on prompts for multimodal-to-text pre-trained models has primarily focused on hard prompts. Conversely, soft prompts based on image-text matching models, such as CLIP [2], have received extensive investigation. Models like CoOp [63] and CoCoOp [66] leverage soft prompts on it to enhance model performance. Nevertheless, the exploration of prompt tuning for popular generative multimodal-to-text pre-trained models remains largely unexplored.

Additionally, multimodal-to-text pre-trained models employ a range of challenging prompt techniques, including in-context learning [1] and instruction tuning [203]. Despite their effectiveness, the underlying mechanisms by which these models learn and the specific contributions of different aspects of the demonstrations remain largely unexplored. A deeper understanding of these factors is crucial for refining and optimizing the performance of multimodal-to-text pre-trained models.

Prompting Model in Image-Text Matching. Although pre-trained encoders prompted by a matching loss have been widely used for adaptation in downstream tasks, the exploration of visual prompting on pre-trained encoders remains relatively unexplored. Similar to the seamless adaptation of textual encoders through learning textual prompts, the investigation of visual prompts is an intriguing area that can unlock emergent abilities, especially in difficult scenarios such as dense objects, object hallucination, and the adaptation to modern VLMs. In the future, it is imperative to address questions regarding the specific type of visual prompts that are essential and the semantic information that these prompts introduce. By delving into these inquiries, we can gain a deeper understanding of the role and impact of visual prompts, thereby further advancing the field.

Meanwhile, the investigation of how unified prompting can enhance the performance of both branches remains understudied. In an intuitive sense, a unified prompt can provide us with referential information that spans across modalities, as discussed in [70]. This has the potential to facilitate the development of multimodal models that are capable of visual grounding and enable referential dialogues encompassing visual and textual co-reference.

Prompting Model in Text-Image Generation. One of the significant challenges in the field of prompting text-to-other generation models, particularly in the case of Text-to-Video (T2V) and Text-to-3D (T2-3D) models, is their dependency on Text-to-Image (T2I) models. These models often share the same concern due to the nature that they are extensions of T2I diffusion models. For example, the inconsistency of the input control maps in T2I models can lead to errors in the consequently generated videos and 3D objects, thereby affecting the overall performance and reliability of these extension scenarios.

Looking ahead, there are several promising directions for future research. One such direction is the incorporation of visual prompting into T2I, T2V, and T2-3D diffusion models. In the context of text-to-image generation, visual prompting and visual annotations can offer more visual cues, leading to the creation of more personalized images. This approach allows for more attention to be paid to specific areas of the image, enhancing the detail and accuracy of the generated output. Visual prompts can also be beneficial for video generation, either on a frame-by-frame basis or for T2-3D generation aimed at improving 2D renderings or shapes. The concept of visual prompts can be further expanded to include video prompts, object prompts, and motion prompts, depending on the specific requirements of different target data generation scenarios. Furthermore, text-image matching models hold the potential for better alignment as multi-modal prompting in the generation. This approach could lead to more accurate and contextually relevant image generation, opening up new possibilities for the application of pre-trained vision-language models.

Generalizing Prompting Methods from Unimodal to Multimodal. Sec. 6 discusses the applications of prompt engineering in both pure vision and pure language models, which can motivate further research in multi-modality research. When combined with instruction-tuning methods, pure language models have enabled phenomenal applications such as ChatGPT [50]. The potential of these methods such as Reinforcement Learning from Human Feedback (RLHF) [180] and Harmlessness from AI Feedback [204] can be further explored in multimodal models as shown in several recent studies [203, 205]. Constitutional AI is a method proposed in [204] to train a harmless AI assistant through self-improvement, without any human labels identifying harmful outputs. Although some efforts have been put into language models [204], how to implement constitutional AI in the multimodal domain is still an open question.

Another potential direction is the adoption of in-context prompts in multimodal models. Large unimodal language models can address a specific new task given several demonstrations of the task in their text prompt without any gradient update. Flamingo [1] has also demonstrated the few-shot in-context learning ability, but how to further improve the in-context learning capacity is still under-explored.

Responsible AI Considerations of Prompting. There are already a few studies concerning ethical issues on multimodal-to-text generation in Sec. 3.6 and text-to-image generations as discussed in Sec. 5.4. Integrity and ethical issues of prompt engineering on vision-language models need much more attention. One possible direction is to prevent bias and backdoor attacks

inherited from the pre-trained model during downstream prompt adaptations [96, 206, 207, 208]. As large models are normally pre-trained on web-scale datasets which may preserve biased knowledge or sensitive privacy information, the post-deployment procedure conducted by prompt engineering should be able to control the potential risks.

Adversarial Robustness has been intensively studied in various model architectures, such as Convolutional Networks [209, 210], Vision Transformers [211, 212], and Capsule Networks [213]. It has not been fully understood how the prompting on VLMs with both a vision architecture and a language component performs under adversarial attacks. Especially, the impact of recent advances in VLM on adversarial robustness remains to be studied. E.g., do large prompts bring robustness to VLMs [57]?

Besides, transparency and controllable generation through fair prompting are also essential in generative tasks. Generative models are shown to be vulnerable to privacy leakage [170, 171, 172] and may generate biased content [164, 165, 166, 167]. Hence, constructing transparent and controllable prompts that are capable to conserve privacy and prevent unethical generation is critical for real-world applications. Last but not least, managing the accompanying risks of prompt engineering and large models requires the close collaboration of society, research institutions, and government [214, 215].

Relationship between Prompts on Different VLMs. The recent work [216] studies the relationship between concepts learned by multimodal-to-text and image-to-text and text-to-image models. They show the studied two types of models cannot fully understand each other, while they also share some concepts. Similarly, the relationship between prompts on different types of models should be explored in future work, especially the feasibility of building universal prompts across different models trained on the same data. In addition to the inter-model relationship, the interaction between prompts and model architecture should be investigated since most prompts are proposed on Transformer-based models. Concretely, how the model's self-attention changes during prompting [212].

8 Conclusion

This survey paper on prompt engineering of pre-trained vision-language models has provided valuable insights into the current state of research in this field. The main findings and trends identified through the analysis shed light on the effective utilization of prompts in adapting large pre-trained models for vision-language tasks.

One key finding is the versatility and applicability of prompt engineering across different types of vision-language models, including multimodal-to-text generation models, image-text-matching models, and text-to-image generation models. The survey explored each model type from their respective characteristics, highlighting various prompting methods on them.

The implications of these findings are significant for both academia and industry. By leveraging prompt engineering techniques, researchers can achieve remarkable performance gains in vision-language models without the need for extensive labeled data. This has the potential to reduce the burden of data annotation and accelerate the deployment of vision-language models in real-world applications.

However, it is important to acknowledge the limitations of this survey. The rapidly evolving nature of the field and the wide range of existing prompt engineering approaches make it challenging to provide an exhaustive overview. Additionally, the survey focused primarily on pre-trained vision-language models from a prompting engineering perspective and may not have covered all recent advancements in other related areas.

To address these limitations, we will maintain and release a platform to keep tracking the advance in this area. Further research should explore the integration of prompt engineering techniques with other emerging technologies, such as reinforcement learning or meta-learning, to enhance the performance and generalization capabilities of vision-language models. Additionally, investigations into the interpretability and robustness of prompt-engineered models are crucial for ensuring their practical deployment and ethical use.

Overall, this survey contributes to the existing body of knowledge by providing a comprehensive overview of prompt engineering in pre-trained vision-language models. By elucidating the current state, key trends, and implications of prompt engineering techniques, this survey serves as a valuable resource for researchers and practitioners aiming to harness the potential of vision-language models for various applications. It fills a gap in research by offering insights into the adaptation of pre-trained models in the context of vision and language, paving the way for further advancements in this exciting field.

Acknowledgements

We would like to thank Ananth Balashankar (Google Research) and Ashkan Khakzar (University of Oxford) for constructive feedback on an earlier version of this manuscript.

References

[1] J.-B. Alayrac et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

[2] A. Radford et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

[3] R. Rombach et al. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

[4] J. Wei et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[5] Z. Zhang et al. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*, 2023.

[6] B. Lester et al. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

[7] Q. Dong et al. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

[8] P. Liu et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

[9] S. Qiao et al. Reasoning with language model prompting: A survey. *arXiv preprint arXiv:2212.09597*, 2022.

[10] H. Bahng et al. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 1(3):4, 2022.

[11] X. Wang et al. SegGPT: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023.

[12] C. Wu et al. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

[13] X. Liu et al. Gpt understands, too. *arXiv preprint arXiv:2103.10385*, 2021.

[14] E. J. Hu et al. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[15] S. Huang et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023.

[16] Z. Yang et al. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[17] M. Tsimpoukelli et al. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.

[18] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[19] S. Long et al. Vision-and-language pretrained models: A survey. In L. D. Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 5530–5537. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Survey Track.

[20] J. Devlin et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[21] L. Pritchett and J. Sandefur. Learning from experiments when context matters. *American Economic Review*, 105(5):471–475, 2015.

[22] C. Raffel et al. Exploring the limits of transfer learning with a unified text-to-text transformer. In *JMLR*, 2020.

[23] T. Brown et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[24] A. Singh et al. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15638–15650, 2022.

[25] J. Lu et al. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[26] J. Cho et al. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pp. 1931–1942. PMLR, 2021.

[27] S. Ren et al. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[28] P. Wang et al. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pp. 23318–23340. PMLR, 2022.

[29] Z. Wang et al. SimVLM: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations*, 2022.

[30] X. Chen et al. PaLI: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations*, 2023.

[31] C. Eichenberg et al. MAGMA – multimodal augmentation of generative models through adapter-based finetuning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2416–2428, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[32] J. Li et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[33] A. Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[34] S. Zhang et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[35] H. W. Chung et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[36] A. Radford et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[37] A. Efrat and O. Levy. The turking test: Can language models understand instructions? *arXiv preprint arXiv:2010.11982*, 2020.

[38] O. Rubin et al. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2655–2671, Seattle, United States, July 2022. Association for Computational Linguistics.

[39] X. Li et al. Unified demonstration retriever for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4644–4668, Toronto, Canada, July 2023. Association for Computational Linguistics.

[40] J. Ye et al. Compositional exemplars for in-context learning. In *International Conference on Machine Learning*. PMLR, 2023.

[41] G. Qin and J. Eisner. Learning how to ask: Querying LMs with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5203–5212, Online, June 2021. Association for Computational Linguistics.

[42] X. L. Li and P. Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

[43] H. Yang et al. Prompt tuning for generative multimodal pretrained models. *arXiv preprint arXiv:2208.02532*, 2022.

[44] D. Noever and S. E. M. Noever. The multimodal and modular ai chef: Complex recipe generation from imagery. *arXiv preprint arXiv:2304.02016*, 2023.

[45] J. Li et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[46] Z. Peng et al. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

[47] J. B. Nici. *AP Art History: 5 Practice Tests + Comprehensive Review + Online Practice.* Barron's Educational Series, 2020.

[48] D. A. Hudson and C. 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*, pp. 6700–6709, 2019.

[49] Y. Zang et al. Unified Vision and Language Prompt Learning. *arXiv:2210.07225*, 2022.

[50] Chatgpt. https://openai.com/blog/chatgpt. Accessed: 2023-07-22.

[51] K. Zhang et al. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv preprint arXiv:2305.17100*, 2023.

[52] L. Weidinger et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

[53] S. Guo et al. Threats to pre-trained language models: Survey and taxonomy. *arXiv preprint arXiv:2202.06862*, 2022.

[54] J. Qiu et al. Benchmarking robustness under distribution shift of multimodal image-text models. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2022.

[55] Y. Zhao et al. On evaluating adversarial robustness of large vision-language models. *arXiv preprint arXiv:2305.16934*, 2023.

[56] S. Chen et al. Benchmarking robustness of adaptation methods on pre-trained vision-language models. *arXiv preprint arXiv:2306.02080*, 2023.

[57] J. Gu et al. Towards robust prompts on vision-language models. *arXiv preprint arXiv:2304.08479*, 2023.

[58] J. Li et al. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[59] G. Zhang et al. Multi-event video-text retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*

(ICCV) (to appear), 2023.

[60] T. Huang et al. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*, 2022.

[61] T. Gao et al. Making Pre-trained Language Models Better Few-shot Learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3816–3830, Online, August 2021. Association for Computational Linguistics.

[62] M. Shu et al. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022.

[63] K. Zhou et al. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[64] C. Ju et al. Prompting visual-language models for efficient video understanding. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pp. 105–124. Springer, 2022.

[65] S. Shen et al. Multitask Vision-Language Prompt Tuning. *arXiv:2211.11720*, 2022.

[66] K. Zhou et al. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16816–16825, 2022.

[67] M. Jia et al. Visual prompt tuning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pp. 709–727. Springer, 2022.

[68] J. Wu et al. Unleashing the power of visual prompting at the pixel level. *arXiv preprint arXiv:2212.10556*, 2022.

[69] Q. Huang et al. Diversity-aware meta visual prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10878–10887, 2023.

[70] Y. Yao et al. CPT: Colorful Prompt Tuning for Pre-trained Vision-Language Models. *arXiv:2109.11797*, 2022.

[71] A. Shtedritski et al. What does clip know about a red circle? visual prompt engineering for vlms. *arXiv preprint arXiv:2304.06712*, 2023.

[72] A. Bar et al. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, 35:25005–25017, 2022.

[73] M. U. Khattak et al. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19113–19122, 2023.

[74] B. Dong et al. LPT: Long-tailed Prompt Tuning for Image Classification. *arXiv preprint arXiv:2210.01033*, 2023.

[75] Z. Guo et al. Texts as images in prompt tuning for multi-label image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2808–2817, 2023.

[76] X. Sun et al. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. *Advances in Neural Information Processing Systems*, 35:30569–30582, 2022.

[77] J. Wen et al. Visual Prompt Tuning for Few-Shot Text Classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 5560–5570, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[78] X. Gu et al. Open-vocabulary Object Detection via Vision and Language Knowledge Distillation. *arXiv preprint arXiv:2104.13921*, 2022.

[79] Y. Du et al. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14084–14093, 2022.

[80] C. Feng et al. Promptdet: Towards open-vocabulary detection using uncurated images. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pp. 701–717. Springer, 2022.

[81] S. Xiao and W. Fu. Optimizing Continuous Prompts for Visual Relationship Detection by Affix-Tuning. *IEEE Access*, 10:70104–70112, 2022.

[82] T. He et al. Towards open-vocabulary scene graph generation with prompt-based finetuning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, pp. 56–73. Springer, 2022.

[83] K. Gao et al. Compositional Prompt Tuning with Motion Cues for Open-vocabulary Video Relation Detection. *arXiv preprint arXiv:2302.00268*, 2023.

[84] Y. Rao et al. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18082–18091, 2022.

[85] A. Kirillov et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[86] C. Ge et al. Domain Adaptation via Prompt Learning. *arXiv:2202.06687*, 2022.

[87] Y. Gao et al. Visual Prompt Tuning for Test-time Domain Adaptation. *arXiv:2210.04831*, 2022.

[88] Z. Wang et al. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 139–149, 2022.

[89] Z. Wang et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pp. 631–648. Springer, 2022.

[90] Z. Zheng et al. Prompt Vision Transformer for Domain Generalization. *arXiv preprint arXiv:2208.08914*, 2022.

[91] C. Mao et al. Understanding Zero-Shot Adversarial Robustness for Large-Scale Models, April 2023.

[92] A. Chen et al. Visual prompting for adversarial robustness. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

[93] A. Fang et al. Data determines distributional robustness in contrastive language image pre-training (clip). In *International Conference on Machine Learning*, pp. 6216–6234. PMLR, 2022.

[94] Z. Shi et al. Effective robustness against natural distribution shifts for models with different training data. *arXiv preprint arXiv:2302.01381*, 2023.

[95] L. Xu et al. Exploring the universal vulnerability of prompt-based learning paradigm. *arXiv preprint arXiv:2204.05239*, 2022.

[96] N. Carlini and A. Terzis. Poisoning and Backdooring Contrastive Learning, March 2022.

[97] J. Jia et al. Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 2043–2059. IEEE, 2022.

[98] H. Bansal et al. Cleanclip: Mitigating data poisoning attacks in multimodal contrastive learning. *arXiv preprint arXiv:2303.03323*, 2023.

[99] S. Agarwal et al. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*, 2021.

[100] C.-Y. Chuang et al. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*, 2023.

[101] F. Kong et al. Mitigating test-time bias for fair image retrieval. *arXiv preprint arXiv:2305.19329*, 2023.

[102] B. Smith et al. Balancing the picture: Debiasing vision-language datasets with synthetic contrast sets. *arXiv preprint arXiv:2305.15407*, 2023.

[103] K. Gregor et al. Draw: A recurrent neural network for image generation. In *International conference on machine learning*, pp. 1462–1471. PMLR, 2015.

[104] I. Goodfellow et al. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[105] S. Reed et al. Generative adversarial text to image synthesis. In

International conference on machine learning, pp. 1060–1069. PMLR, 2016.

[106] X. Pan et al. Drag your gan: Interactive point-based manipulation on the generative image manifold. *arXiv preprint arXiv:2305.10973*, 2023.

[107] T. Karras et al. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

[108] P. Isola et al. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

[109] D. P. Kingma et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

[110] Y. Pu et al. Variational autoencoder for deep learning of images, labels and captions. *Advances in neural information processing systems*, 29, 2016.

[111] A. Vahdat and J. Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679, 2020.

[112] A. Ramesh et al. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

[113] J. Yu et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.

[114] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[115] J. Sohl-Dickstein et al. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

[116] D. Jeulin. Dead leaves models: from space tesselation to random functions proc. of the symposium on the advances in the theory and applications of random sets (fontainebleau, 9-11 october 1996) ed d jeulin, 1997.

[117] R. M. Neal. Annealed importance sampling. *Statistics and computing*, 11:125–139, 2001.

[118] J. Ho et al. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[119] S. Witteveen and M. Andrews. Investigating prompt engineering in diffusion models. *arXiv preprint arXiv:2211.15462*, 2022.

[120] W. Wu et al. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. *arXiv preprint arXiv:2303.11681*, 2023.

[121] R. Beaumont. Clip retrieval: Easily compute clip embeddings and build a clip retrieval system with them, 2022.

[122] M. Ni et al. Imaginarynet: Learning object detectors without real images and annotations. In *ICLR*, 2023.

[123] R. He et al. Is synthetic data from generative models ready for image recognition? In *ICLR*, 2023.

[124] O. Avrahami et al. Blended latent diffusion. *arXiv preprint arXiv:2206.02779*, 2022.

[125] A. Nichol et al. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[126] R. Gal et al. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023.

[127] N. Ruiz et al. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

[128] N. Kumari et al. Multi-concept customization of text-to-image diffusion. *arXiv preprint arXiv:2212.04488*, 2022.

[129] W. Feng et al. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022.

[130] D. Epstein et al. Diffusion self-guidance for controllable image generation. *arXiv preprint arXiv:2306.00986*, 2023.

[131] B. Kawar et al. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6007–6017, 2023.

[132] J. Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, pp. 679–698, 1986.

[133] L. Zhang and M. Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

[134] A. Hertz et al. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[135] U. Singer et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[136] J. Ho et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[137] J. Ho et al. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23(47):1–33, 2022.

[138] C. Qi et al. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*, 2023.

[139] J. Z. Wu et al. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *arXiv preprint arXiv:2212.11565*, 2022.

[140] L. Ruan et al. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10219–10228, 2023.

[141] J. Zhu et al. Moviefactory: Automatic movie creation from text using large generative models for language and images. *arXiv preprint arXiv:2306.07257*, 2023.

[142] N. Müller et al. Diffrf: Rendering-guided 3d radiance field diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4328–4338, 2023.

[143] B. Poole et al. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[144] B. Mildenhall et al. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[145] C.-H. Lin et al. Magic3d: High-resolution text-to-3d content creation. *arXiv preprint arXiv:2211.10440*, 2022.

[146] J. Xu et al. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. *arXiv preprint arXiv:2212.14704*, 2022.

[147] H.-H. Lee and A. X. Chang. Understanding pure clip guidance for voxel grid nerf models. *arXiv preprint arXiv:2209.15172*, 2022.

[148] N. M. Khalid et al. Clip-mesh: Generating textured meshes from text using pretrained image-text models. *arXiv preprint arXiv:2203.13333*, 2022.

[149] M. Zhang et al. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022.

[150] J. Kim et al. Flame: Free-form language-based motion synthesis & editing. *arXiv preprint arXiv:2209.00349*, 2022.

[151] G. Tevet et al. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

[152] T. Yu et al. Scaling robot learning with semantically imagined experience. *arXiv preprint arXiv:2302.11550*, 2023.

[153] D. Rempe et al. Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[154] M. A. Shabani et al. Housediffusion: Vector floorplan generation via a diffusion model with discrete and continuous denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5466–5475, 2023.

[155] H. Jeong et al. Zero-shot generation of coherent storybook from plain text story using diffusion models. *arXiv preprint arXiv:2302.03900*, 2023.

[156] Y. Lu et al. Multimodal procedural planning via dual text-image prompting. *arXiv preprint arXiv:2305.01795*, 2023.

[157] A. Jobin et al. The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, 2019.

[158] J. Chen et al. Diffusion models for imperceptible and transferable adversarial attack. *arXiv preprint arXiv:2305.08192*, 2023.

[159] W. Nie et al. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022.

[160] H. Zhuang et al. A pilot study of query-free adversarial attack against stable diffusion. *arXiv preprint arXiv:2303.16378*, 2023.

[161] L. Struppek et al. Rickrolling the artist: Injecting invisible backdoors into text-guided image generation models. *arXiv preprint arXiv:2211.02408*, 2022.

[162] S. Zhai et al. Text-to-image diffusion models can be easily backdoored through multimodal data poisoning. *arXiv preprint arXiv:2305.04175*, 2023.

[163] Y. Huang et al. Zero-day backdoor attack against text-to-image diffusion models via personalization. *arXiv preprint arXiv:2305.10701*, 2023.

[164] F. Friedrich et al. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023.

[165] R. Naik and B. Nushi. Social biases through the text-to-image generation lens. *arXiv preprint arXiv:2304.06034*, 2023.

[166] J. Wang et al. T2iat: Measuring valence and stereotypical biases in text-to-image generation. *arXiv preprint arXiv:2306.00905*, 2023.

[167] A. S. Luccioni et al. Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*, 2023.

[168] A. Seth et al. Dear: Debiasing vision-language models with additive residuals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6820–6829, 2023.

[169] Y. Kim et al. Explaining visual biases as words by generating captions. *arXiv preprint arXiv:2301.11104*, 2023.

[170] Y. Wu et al. Membership inference attacks against text-to-image generation models. *arXiv preprint arXiv:2210.00968*, 2022.

[171] J. Duan et al. Are diffusion models vulnerable to membership inference attacks? *arXiv preprint arXiv:2302.01316*, 2023.

[172] R. Webster. A reproducible extraction of training images from diffusion models. *arXiv preprint arXiv:2305.08694*, 2023.

[173] X. Shen et al. Prompt stealing attacks against text-to-image generation models. *arXiv preprint arXiv:2302.09923*, 2023.

[174] D. Khashabi et al. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1896–1907, Online, November 2020. Association for Computational Linguistics.

[175] Z. Jiang et al. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021.

[176] B. Lester et al. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[177] T. Schick and H. Schütze. Few-shot text generation with natural language instructions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 390–402, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[178] X. Chen et al. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *Proceedings of the ACM Web Conference 2022*, WWW '22, pp. 2778–2788, New York, NY, USA, 2022. Association for Computing Machinery.

[179] L. Cui et al. Template-based named entity recognition using BART. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 1835–1845, Online, August 2021. Association for Computational Linguistics.

[180] L. Ouyang et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[181] R. Anil et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

[182] B. Paranjape et al. Prompting contrastive explanations for commonsense reasoning tasks. *arXiv preprint arXiv:2106.06823*, 2021.

[183] J. Liu et al. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*, 2021.

[184] Y. Fu et al. Complexity-based prompting for multi-step reasoning. *arXiv preprint arXiv:2210.00720*, 2022.

[185] P. Lewis et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[186] S. Borgeaud et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pp. 2206–2240. PMLR, 2022.

[187] G. Izacard et al. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*, 2022.

[188] O. Press et al. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*, 2022.

[189] S. M. Kazemi et al. Lambada: Backward chaining for automated reasoning in natural language. *arXiv preprint arXiv:2212.13894*, 2022.

[190] T. Khot et al. Decomposed prompting: A modular approach for solving complex tasks. *arXiv preprint arXiv:2210.02406*, 2022.

[191] Y. Yang et al. A prompting-based approach for adversarial example generation and robustness enhancement. *arXiv preprint arXiv:2203.10714*, 2022.

[192] X. Dong et al. Promptattack: Probing dialogue state trackers with adversarial prompts. *arXiv preprint arXiv:2306.04535*, 2023.

[193] P. Delobelle et al. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1693–1706, Seattle, United States, July 2022. Association for Computational Linguistics.

[194] R. Lou et al. Is prompt all you need? no. A comprehensive and broader view of instruction learning. *CoRR*, abs/2303.10475, 2023.

[195] N. Ding et al. Openprompt: An open-source framework for prompt-learning. In V. Basile et al., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022 - System Demonstrations, Dublin, Ireland, May 22-27, 2022*, pp. 105–113. Association for Computational Linguistics, 2022.

[196] X. Wang et al. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6830–6839, 2023.

[197] J. Loedeman et al. Prompt generation networks for input-based adaptation of frozen vision transformers. *arXiv preprint arXiv:2210.06466*, 2023.

[198] C.-H. Tu et al. Visual query tuning: Towards effective usage of intermediate representations for parameter and memory efficient transfer learning. *arXiv preprint arXiv:2212.03220*, 2022.

[199] S. Zhang et al. Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. *arXiv preprint arXiv:2212.05590*, 2022.

[200] H. Salman et al. Unadversarial examples: Designing objects for robust vision. *Advances in Neural Information Processing Systems*, 34:15270–15284, 2021.

[201] Y. Li et al. Exploring the benefits of visual prompting in differential privacy. *arXiv preprint arXiv:2303.12247*, 2023.

[202] R. Girdhar et al. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190, June 2023.

[203] B. Li et al. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023.

[204] Y. Bai et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[205] P. Gao et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

[206] K. Gao et al. Backdoor defense via adaptively splitting poisoned dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4005–4014, 2023.

[207] K. Huang et al. Backdoor defense via decoupling the training process. *arXiv preprint arXiv:2202.03423*, 2022.

[208] S. Yang et al. Backdoor defense via suppressing model shortcuts. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

[209] A. Madry et al. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2017.

[210] X. Jia et al. Las-at: adversarial training with learnable attack strategy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13398–13408, 2022.

[211] B. Wu et al. Towards efficient adversarial training on vision transformers. In *European Conference on Computer Vision*, pp. 307–325. Springer, 2022.

[212] J. Gu et al. Are vision transformers robust to patch perturbations? In *European Conference on Computer Vision*, pp. 404–421. Springer, 2022.

[213] J. Gu et al. Effective and efficient vote attack on capsule networks. In *International Conference on Learning Representations (ICLR)*, 2021.

[214] M. Anderljung et al. Frontier ai regulation: Managing emerging risks to public safety. *arXiv preprint arXiv:2307.03718*, 2023.

[215] P. Hacker et al. Regulating chatgpt and other large generative ai models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1112–1123, 2023.

[216] H. Li et al. Do dall-e and flamingo understand each other? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (to appear)*, 2023.