

Multimodal Foundation Models: From Specialists to General-Purpose Assistants

Chunyuan Li^{*♣}, Zhe Gan*, Zhengyuan Yang*, Jianwei Yang*, Linjie Li*,
Lijuan Wang, Jianfeng Gao

Microsoft Corporation

{chunyl,zhgan,zhengyang,jianwyan,linjli,lijuanw,jfgao}@microsoft.com

* Core Contribution ♣ Project Lead



Abstract

This paper presents a comprehensive survey of the taxonomy and evolution of multimodal foundation models that demonstrate vision and vision-language capabilities, focusing on the transition from specialist models to general-purpose assistants. The research landscape encompasses five core topics, categorized into two classes. (i) We start with a survey of well-established research areas: multimodal foundation models pre-trained for specific purposes, including two topics – methods of learning vision backbones for visual understanding and text-to-image generation. (ii) Then, we present recent advances in exploratory, open research areas: multimodal foundation models that aim to play the role of general-purpose assistants, including three topics – unified vision models inspired by large language models (LLMs), end-to-end training of multimodal LLMs, and chaining multimodal tools with LLMs. The target audiences of the paper are researchers, graduate students, and professionals in computer vision and vision-language multimodal communities who are eager to learn the basics and recent advances in multimodal foundation models.

¹ Chunyuan Li initiated the project, and took lead in the writing of Chapter 1, 5 and 7. Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li took lead in the writing of Chapter 2, 3, 4 and 6, respectively. Lijuan Wang and Jianfeng Gao provided comprehensive suggestions and edits of the entire paper. All the authors provided project advice, and contributed to paper review, editing and proofreading.

² Zhe Gan is currently with Apple AI/ML.

Contents

1	Introduction	5
1.1	What are Multimodal Foundation Models?	6
1.2	Definition and Transition from Specialists to General-Purpose Assistants	9
1.3	Who Should Read this Paper?	9
1.4	Related Materials: Slide Decks and Pre-recorded Talks	11
2	Visual Understanding	12
2.1	Overview	12
2.2	Supervised Pre-training	13
2.3	Contrastive Language-Image Pre-training	15
2.3.1	Basics of CLIP Training	15
2.3.2	CLIP Variants	16
2.4	Image-Only Self-Supervised Learning	18
2.4.1	Contrastive and Non-contrastive Learning	18
2.4.2	Masked Image Modeling	19
2.5	Synergy Among Different Learning Approaches	21
2.6	Multimodal Fusion, Region-Level and Pixel-Level Pre-training	23
2.6.1	From Multimodal Fusion to Multimodal LLM	23
2.6.2	Region-Level Pre-training	25
2.6.3	Pixel-Level Pre-training	25
3	Visual Generation	27
3.1	Overview	27
3.1.1	Human Alignments in Visual Generation	27
3.1.2	Text-to-Image Generation	29
3.2	Spatial Controllable Generation	31
3.3	Text-based Editing	33
3.4	Text Prompts Following	35
3.5	Concept Customization	38
3.6	Trends: Unified Tuning for Human Alignments	40

4 Unified Vision Models	42
4.1 Overview	42
4.2 From Closed-Set to Open-Set Models	43
4.2.1 Object Detection and Grounding	46
4.2.2 Image Segmentation and Referring	47
4.3 From Task-Specific Models to Generic Models	49
4.3.1 I/O Unification	50
4.3.2 Functionality Unification	53
4.4 From Static to Promptable Models	56
4.4.1 Multi-modal Prompting	56
4.4.2 In-context Prompting	58
4.5 Summary and Discussion	59
5 Large Multimodal Models: Training with LLM	61
5.1 Background	61
5.1.1 Image-to-Text Generative Models	61
5.1.2 Case Studies	61
5.1.3 OpenAI Multimodal GPT-4 and Research Gaps	63
5.2 Pre-requisite: Instruction Tuning in Large Language Models	64
5.2.1 Instruction Tuning	65
5.2.2 Self-Instruct and Open-Source LLMs	66
5.3 Instruction-Tuned Large Multimodal Models	68
5.4 Advanced Topics	70
5.5 How Close We Are To OpenAI Multimodal GPT-4?	76
6 Multimodal Agents: Chaining Tools with LLM	77
6.1 Overview	77
6.2 Multimodal Agent	78
6.3 Case Study: MM-REACT	80
6.3.1 System Design	81
6.3.2 Capabilities	83
6.3.3 Extensibility	84
6.4 Advanced Topics	85
6.4.1 Comparison to Training with LLM in Chapter 5	85
6.4.2 Improving Multimodal Agents	85
6.4.3 Diverse Applications of Multimodal Agents	86
6.4.4 Evaluation of Multimodal Agents	87
6.4.5 Tool Creation	88
6.4.6 Retrieval-Augmented Multimodal Agents	88

7 Conclusions and Research Trends	89
7.1 Summary and Conclusions	89
7.2 Towards Building General-Purpose AI Agents	90

Chapter 1

Introduction

Vision is one of the primary channels for humans and many living creatures to perceive and interact with the world. One of the core aspirations in artificial intelligence (AI) is to develop AI agents to mimic such an ability to effectively perceive and generate visual signals, and thus reason over and interact with the visual world. Examples include recognition of the objects and actions in the scenes, and creation of sketches and pictures for communication. Building foundational models with visual capabilities is a prevalent research field striving to accomplish this objective.

Over the last decade, the field of AI has experienced a fruitful trajectory in the development of models. We divide them into four categories, as illustrated in Figure 1.1. The categorization can be shared among different fields in AI, including language, vision and multimodality. We first use language models in NLP to illustrate the evolution process. (i) At the early years, task-specific models are developed for individual datasets and tasks, typically being trained from scratch. (ii) With large-scale pre-training, language models achieve state-of-the-art performance on many established language understanding and generation tasks, such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), T5 (Raffel et al., 2020), DeBERTa (He et al., 2021) and GPT-2 (Radford et al., 2019)). These pre-trained models serve the basis for downstream task adaptation. (iii) Exemplified by GPT-3 (Brown et al., 2020), large language models (LLMs) unify various language understanding and generation tasks into one model. With web-scale training and unification, some emerging capabilities appear, such as in-context-learning and chain-of-thoughts. (iv) With recent advances in human-AI alignment, LLMs start to play the role of general-purpose assistants to follow human intents to complete a wide range of language tasks in the wild, such as ChatGPT (OpenAI, 2022) and GPT-4 (OpenAI, 2023a). These assistants exhibit interesting capabilities, such as interaction and tool use, and lay a foundation for developing general-purpose AI agents. It is important to note that the latest iterations of foundation models build upon the noteworthy features of their earlier counterparts while also providing additional capabilities.

Inspired by the great successes of LLMs in NLP, it is natural for researchers in the computer vision and vision-language community to ask the question: what is the counterpart of ChatGPT/GPT-4 for vision, vision-language and multi-modal models? There is no doubt that vision pre-training and vision-language pre-training (VLP) have attracted a growing attention since the birth of BERT, and has become the mainstream learning paradigm for vision, with the promise to learn universal transferable visual and vision-language representations, or to generate highly plausible images. Arguably, they can be considered as the early generation of multimodal foundation models, just as BERT/GPT-2 to the language field. While the road-map to build general-purpose assistants for language such as ChatGPT is clear, it is becoming increasingly crucial for the research community to explore feasible solutions to building its counterpart for computer vision: the general-purpose visual assistants. Overall, building general-purpose agents has been a long-standing goal for AI. LLMs with emerging properties have significantly reduced the cost of building such agents for language tasks. Similarly, we foresee emerging capabilities from vision models, such as following the instructions composed by various visual prompts like user-uploaded images, human-drawn clicks, sketches and mask, in addition to text prompt. Such strong zero-shot visual task composition capabilities can significantly reduce the cost of building AI agents.

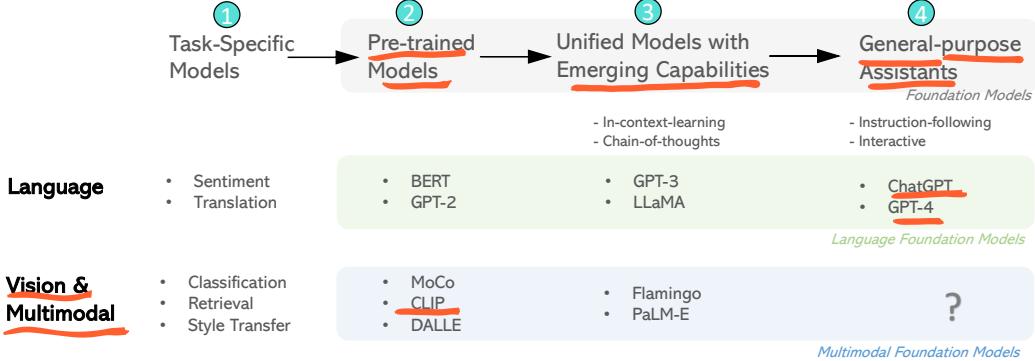


Figure 1.1: Illustration of foundation model development trajectory for language and vision/multimodality. Among the four categories, the first category is the task-specific model, and the last three categories belong to foundation models, where these foundation models for language and vision are grouped in green and blue blocks, respectively. Some prominent properties of models in each category are highlighted. By comparing the models between language and vision, we are foreseeing that the transition of multimodal foundation models follows a similar trend: from the pre-trained model for specific purpose, to unified models and general-purpose assistants. However, research exploration is needed to figure out the best recipe, which is indicated as the question mark in the figure, as multimodal GPT-4 and Gemini stay private.

In this paper, we limit the scope of multimodal foundation models to the **vision** and **vision-language** domains. Recent survey papers on related topics include (i) **image understanding models** such as self-supervised learning (Jaiswal et al., 2020; Jing and Tian, 2020; Ozbulak et al., 2023), segment anything (SAM) (Zhang et al., 2023a,c), (ii) **image generation models** (Zhang et al., 2023b; Zhou and Shimada, 2023), and (iii) **vision-language pre-training (VLP)**. Existing VLP survey papers cover VLP methods for task-specific VL problems before the era of pre-training, image-text tasks, core vision tasks, and/or video-text tasks (Zhang et al., 2020; Du et al., 2022; Li et al., 2022c; Ruan and Jin, 2022; Chen et al., 2022a; Gan et al., 2022; Zhang et al., 2023g). Two recent survey papers cover the integration of vision models with LLM (Awais et al., 2023; Yin et al., 2022).

Among them, Gan et al. (2022) is a survey on VLP that covers the CVPR tutorial series on *Recent Advances in Vision-and-Language Research* in 2022 and before. This paper summarizes the CVPR tutorial on *Recent Advances in Vision Foundation Models* in 2023. Different from the aforementioned survey papers that focus on literature review of a given research topic, this paper presents our perspectives on the role transition of multimodal foundation models from specialists to general-purpose visual assistants, in the era of large language models. The contributions of this survey paper are summarized as follows.

- We provide a comprehensive and timely survey on modern multimodal foundation models, not only covering well-established models for visual representation learning and image generation, but also summarizing emerging topics for the past 6 months inspired by LLMs, including unified vision models, training and chaining with LLMs.
- The paper is positioned to provide the audiences with the perspective to advocate a transition in developing multimodal foundation models. On top of great modeling successes for specific vision problems, we are moving towards building general-purpose assistants that can follow human intents to complete a wide range of computer vision tasks in the wild. We provide in-depth discussions on these advanced topics, demonstrating the potential of developing general-purpose visual assistants.

两大技术背景：迁移学习 + 大规模语言模型

1.1 What are Multimodal Foundation Models?

As elucidated in the Stanford foundation model paper (Bommasani et al., 2021), AI has been undergoing a paradigm shift with the rise of models (*e.g.*, BERT, GPT family, CLIP (Radford et al., 2021) and DALL-E (Ramesh et al., 2021a)) trained on broad data that can be adapted to a wide range of downstream tasks. They call these models *foundation models* to underscore their critically central

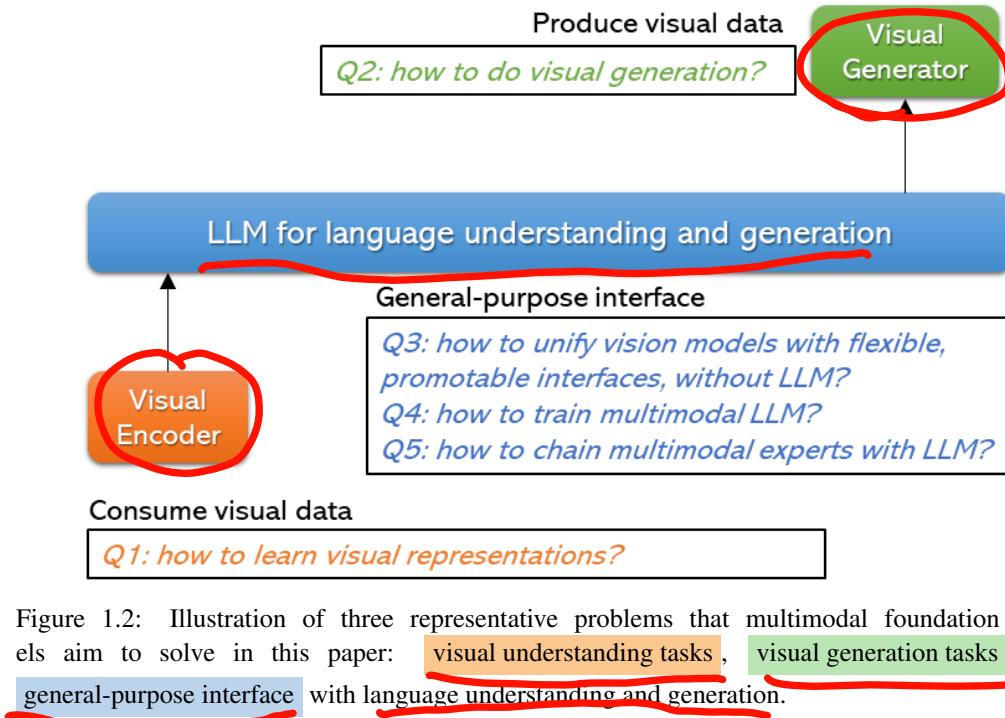


Figure 1.2: Illustration of three representative problems that multimodal foundation models aim to solve in this paper: **visual understanding tasks**, **visual generation tasks**, and **general-purpose interface** with language understanding and generation.

yet incomplete character: homogenization of the methodologies across research communities and emergence of new capabilities. From a technical perspective, it is **transfer learning** that makes foundation models possible, and it is **scale** that makes them powerful. The emergence of foundation models has been predominantly observed in the NLP domain, with examples ranging from BERT to ChatGPT. This trend has gained traction in recent years, extending to computer vision and other fields. In NLP, the introduction of **BERT** in late 2018 is considered as the **inception of the foundation model era**. The remarkable success of BERT rapidly stimulates interest in self-supervised learning in the computer vision community, giving rise to models such as SimCLR (Chen et al., 2020a), MoCo (He et al., 2020), BEiT (Bao et al., 2022), and MAE (He et al., 2022a). During the same time period, the success of pre-training also significantly promotes the vision-and-language multimodal field to an unprecedented level of attention.

In this paper, we focus on multimodal foundation models, which inherit all properties of foundation models discussed in the Stanford paper (Bommasani et al., 2021), but with an emphasis on models with the capability to deal with vision and vision-language modalities. Among the ever-growing literature, we categorize multimodal foundation models in Figure 1.2, based on their functionality and generality. For each category, we present exemplary models that demonstrate the primary capabilities inherent to these multimodal foundation models.

- **Visual Understanding Models.** (Highlighted with orange in Figure 1.2) Learning general visual representations is essential to build vision foundation models, as pre-training a strong vision backbone is fundamental to all types of computer vision downstream tasks, ranging from image-level (*e.g.*, image classification, retrieval, and captioning), region-level (*e.g.*, detection and grounding) to pixel-level tasks (*e.g.*, segmentation). We group the methods into three categories, depending on the types of supervision signals used to train the models.
 - **Label supervision.** Datasets like ImageNet (Krizhevsky et al., 2012) and ImageNet21K (Ridnik et al., 2021) have been popular for supervised learning, and larger-scale proprietary datasets are also used in industrial labs (Sun et al., 2017; Singh et al., 2022b; Zhai et al., 2022a).
 - **Language supervision.** Language is a richer form of supervision. Models like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) are pre-trained using a contrastive loss over millions or even billions of noisy image-text pairs mined from the Web. These models enable zero-shot image classification, and make traditional computer vision (CV) models to perform open-

口译声创像一本对的对比操作

vocabulary CV tasks. We advocate the concept of *computer vision in the wild*,¹ and encourage the development and evaluation of future foundation models for this.

- **Image-only self-supervision.** This line of work aims to learn image representations from supervision signals mined from the images themselves, ranging from contrastive learning (Chen et al., 2020a; He et al., 2020), non-contrastive learning (Grill et al., 2020; Chen and He, 2021; Caron et al., 2021), to masked image modeling (Bao et al., 2022; He et al., 2022a).
- **Multimodal fusion, region-level and pixel-level pre-training.** Besides the methods of pre-training image backbones, we will also discuss pre-training methods that allow multimodal fusion (e.g., CoCa (Yu et al., 2022a), Flamingo (Alayrac et al., 2022)), region-level and pixel-level image understanding, such as open-set object detection (e.g., GLIP (Li et al., 2022e)) and promptable segmentation (e.g., SAM (Kirillov et al., 2023)). These methods typically rely on a pre-trained image encoder or a pre-trained image-text encoder pair.
- **Visual Generation Models.** (Highlighted with green in Figure 1.2) Recently, foundation image generation models have been built, due to the emergence of large-scale image-text data. The techniques that make it possible include the vector-quantized VAE methods (Razavi et al., 2019), diffusion-based models (Dhariwal and Nichol, 2021) and auto-regressive models.
- **Text-conditioned visual generation.** This research area focuses on generating faithful visual content, including images, videos, and more, conditioned on open-ended text descriptions/prompts. Text-to-image generation develops generative models that synthesize images of high fidelity to follow the text prompt. Prominent examples include DALL-E (Ramesh et al., 2021a), DALL-E 2 (Ramesh et al., 2022), Stable Diffusion (Rombach et al., 2021; sta, 2022), Imagen (Saharia et al., 2022), and Parti (Yu et al., 2022b). Building on the success of text-to-image generation models, text-to-video generation models generate videos based on text prompts, such as Imagen Video (Ho et al., 2022) and Make-A-Video (Singer et al., 2022).
- **Human-aligned visual generator.** This research area focuses on improving the pre-trained visual generator to better follow human intentions. Efforts have been made to address various challenges inherent to base visual generators. These include improving spatial controllability (Zhang and Agrawala, 2023; Yang et al., 2023b), ensuring better adherence to text prompts (Black et al., 2023), supporting flexible text-based editing (Brooks et al., 2023), and facilitating visual concept customization (Ruiz et al., 2023).
- **General-purpose Interface.** (Highlighted with blue in Figure 1.2) The aforementioned multimodal foundation models are designed for specific purposes – tackling a specific set of CV problems/tasks. Recently, we see an emergence of general-purpose models that lay the basis of AI agents. Existing efforts focus on three research topics. The first topic aims to unify models for visual understanding and generation. These models are inspired by the unification spirit of LLMs in NLP, but do not explicitly leverage pre-trained LLM in modeling. In contrast, the other two topics embrace and involve LLMs in modeling, including training and chaining with LLMs, respectively.
 - **Unified vision models for understanding and generation.** In computer vision, several attempts have been made to build a general-purpose foundation model by combining the functionalities of specific-purpose multimodal models. To this end, a unified model architecture is adopted for various downstream computer vision and vision-language (VL) tasks. There are different levels of unification. First, a prevalent effort is to bridge vision and language by converting all closed-set vision tasks to open-set ones, such as CLIP (Radford et al., 2021), GLIP (Li et al., 2022f), OpenSeg (Ghiasi et al., 2022a), etc. Second, the unification of different VL understanding tasks across different granularity levels is also actively explored, such as I/O unification methods like UniTAB (Yang et al., 2021), Unified-IO (Lu et al., 2022a)), Pix2Seq-v2 (Chen et al., 2022d) and functional unification methods like GPV (Gupta et al., 2022a), GLIP-v2 (Zhang et al., 2022b)) and X-Decoder (Zou et al., 2023a). In the end, it is also necessitated to make the models more interactive and promptable like ChatGPT, and this has been recently studied in SAM (Kirillov et al., 2023) and SEEM (Zou et al., 2023b).
 - **Training with LLMs.** Similar to the behavior of LLMs, which can address a language task by following the instruction and processing examples of the task in their text prompt, it is desirable to develop a visual and text interface to steer the model towards solving a multimodal task. By extending the capability of LLMs to multimodal settings and training the model end-to-end, multimodal LLMs or large multimodal models are developed, including Flamingo (Alayrac et al., 2022) and Multimodal GPT-4 (OpenAI, 2023a).

¹Computer-Vision-in-the-Wild Readings.

- **Chaining tools with LLM.** Exploiting the tool use capabilities of LLMs, an increasing number of studies integrate LLMs such as ChatGPT with various multimodal foundation models to facilitate image understanding and generation through a conversation interface. This interdisciplinary approach combines the strengths of NLP and computer vision, enabling researchers to develop more robust and versatile AI systems that are capable of processing visual information and generating human-like responses via human-computer conversations. Representative works include Visual ChatGPT (Wu et al., 2023a) and MM-REACT (Yang* et al., 2023).

1.2 Definition and Transition from Specialists to General-Purpose Assistants

Based on the model development history and taxonomy in NLP, we group multimodal foundation models in Figure 1.2 into two categories.

- **Specific-Purpose Pre-trained Vision Models** cover most existing multimodal foundation models, including visual understanding models (e.g., CLIP (Radford et al., 2021), SimCLR (Chen et al., 2020a), BEiT (Bao et al., 2022), SAM (Kirillov et al., 2023)) and visual generation models (e.g., Stable Diffusion (Rombach et al., 2021; sta, 2022)), as they present powerful transferable ability for specific vision problems.
- **General-Purpose Assistants** refer to AI agents that can follow human intents to complete various computer vision tasks in the wild. The meanings of general-purpose assistants are two-fold: (i) generalists with unified architectures that could complete tasks across different problem types, and (ii) easy to follow human instruction, rather than replacing humans. To this end, several research topics have been actively explored, including unified vision modeling (Lu et al., 2022a; Zhang et al., 2022b; Zou et al., 2023a), training and chaining with LLMs (Liu et al., 2023c; Zhu et al., 2023a; Wu et al., 2023a; Yang* et al., 2023).

1.3 Who Should Read this Paper?

This paper is based on our CVPR 2023 tutorial,² with researchers in the computer vision and vision-language multimodal communities as our primary target audience. It reviews the literature and explains topics to those who seek to learn the basics and recent advances in multimodal foundation models. The target audiences are graduate students, researchers and professionals who are not experts of multimodal foundation models but are eager to develop perspectives and learn the trends in the field. The structure of this paper is illustrated in Figure 1.3. It consists of 7 chapters.

- Chapter 1 introduces the landscape of multimodal foundation model research, and presents a historical view on the transition of research from specialists to general-purpose assistants.
- Chapter 2 introduces different ways to consume visual data, with a focus on how to learn a strong image backbone.
- Chapter 3 describes how to produce visual data that aligns with human intents.
- Chapter 4 describes how to design unified vision models, with an interface that is interactive and promptable, especially when LLMs are not employed.
- Chapter 5 describes how to train an LLM in an end-to-end manner to consume visual input for understanding and reasoning.
- Chapter 6 describes how to chain multimodal tools with an LLM to enable new capabilities.
- Chapter 7 concludes the paper and discusses research trends.

Relations among Chapters 2-6. Chapter 2-6 are the core chapters of this survey paper. An overview of the structure for these chapters are provided in Figure 1.2. We start with a discussion of two typical multimodal foundation models for specific tasks, including visual understanding in Chapter 2 and visual generation in Chapter 3. As the notion of multimodal foundation models are originally based on visual backbone/representation learning for understanding tasks, we first present a comprehensive review to the transition of image backbone learning methods, evolving from early

²<https://vlp-tutorial.github.io/2023/index.html>

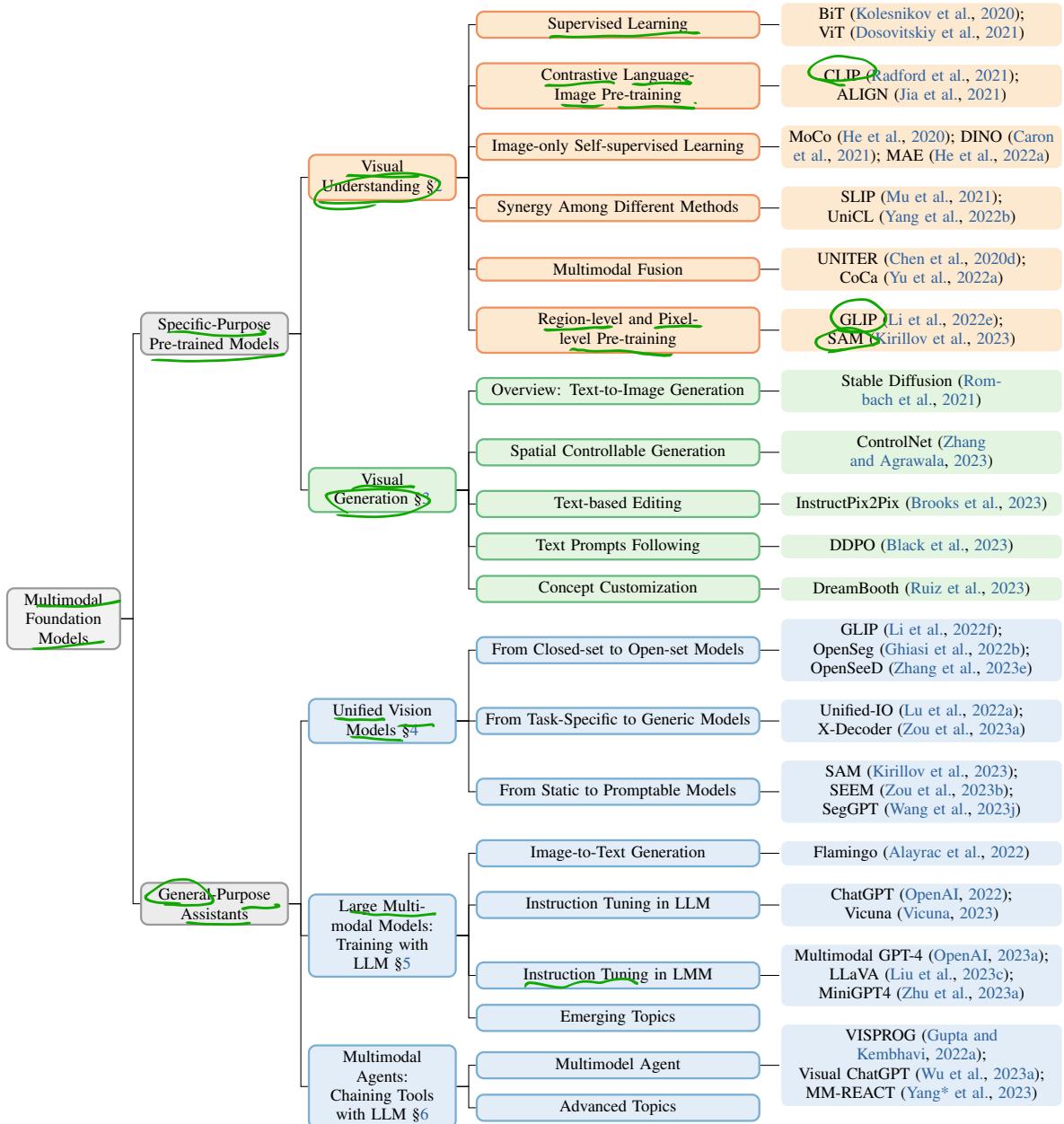


Figure 1.3: An overview of the paper’s structure, detailing Chapters 2-6.

supervised methods to the recent language-image contrastive methods, and extend the discussion on image representations from image-level to region-level and pixel-level (Chapter 2). Recently, generative AI is becoming increasingly popular, where vision generative foundation models have been developed. In Chapter 3, we discuss large pre-trained text-to-image models, and various ways that the community leverage the generative foundation models to develop new techniques to make them better aligned with human intents. Inspired by the recent advances in NLP that LLMs serve as general-purpose assistants for a wide range of language tasks in daily life, the computer vision community has been anticipating and attempting to build general-purpose visual assistants. We discuss three different ways to build general-purpose assistants. Inspired by the spirit of LLMs, Chapter 4 focuses on unifying different vision models of understanding and generation without explicitly incorporating LLMs in modeling. In contrast, Chapter 5 and Chapter 6 focus on embracing LLMs to build general-purpose visual assistants, by explicitly augmenting LLMs in modeling. Specifically, Chapter 5 describes end-to-end training methods, and Chapter 6 focuses on training-free approaches that chain various vision models to LLMs.

How to read the paper. Different readers have different backgrounds, and may have different purposes of reading this paper. Here, we provide a few guidance.

- Each chapter is mostly self-contained. If you have a clear goal and a clear research direction that you want to focus on, then just jump to the corresponding chapter. For example, if you are interested in building a mini prototype using OpenAI’s multimodal GPT-4, then you can directly jump to Chapter 5.
- If you are a beginner of multimodal foundation models, and are interested in getting a glimpse of the cutting-edge research, we highly recommend that you read the whole paper chapter by chapter in order, as the early chapters serve as the building blocks of later chapters, and each chapter provides the description of the key concepts to help you understand the basic ideas, and a comprehensive literature review that to help you grasp the landscape and state of the art.
- If you already have rich experience in multimodal foundation models and are familiar with the literature, feel free to jump to specific chapters you want to read. In particular, we include in most chapters a section to discuss advanced topics and sometimes provide our own perspectives, based on the up-to-date literature. For example, in Chapter 6, we discuss several important aspects of multimodal agents in tool use, including tool creation and its connection to retrieval-augmented methods.

1.4 Related Materials: Slide Decks and Pre-recorded Talks

This survey paper extends what we present in the CVPR 2023 tutorial by covering the most recent advances in the field. Below, we provide a list of slide decks and pre-recorded talks, which are related to the topics in each chapter, for references.

- **Chapter 2:** [Visual and Vision-Language Pre-training](#) (Youtube, Bilibili)
- **Chapter 3:** [Alignments in Text-to-Image Generation](#) (Youtube, Bilibili)
- **Chapter 4:** [From Representation to Interface: The Evolution of Foundation for Vision Understanding](#) (Youtube, Bilibili)
- **Chapter 5:** [Large Multimodal Models](#) (Youtube, Bilibili)
- **Chapter 6:** [Multimodal Agents: Chaining Multimodal Experts with LLMs](#) (Youtube, Bilibili)

Chapter 2

Visual Understanding



Over the past decade, the research community has devoted significant efforts to study the acquisition of high-quality, general-purpose image representations. This is essential to build vision foundation models, as pre-training a strong vision backbone to learn image representations is fundamental to all types of computer vision downstream tasks, ranging from image-level (e.g., image classification (Krizhevsky et al., 2012), image-text retrieval (Frome et al., 2013), image captioning (Chen et al., 2015)), region-level (e.g., object detection (Girshick, 2015), phrase grounding (Plummer et al., 2015)), to pixel-level (e.g., semantic/instance/panoptic segmentation (Long et al., 2015; Hafiz and Bhat, 2020; Kirillov et al., 2019)) tasks.

In this chapter, we present how image representations can be learned, either using supervision signals mined inside the images, or through using language supervision of image-text datasets mined from the Web. Specifically, Section 2.1 presents an overview of different learning paradigms, including supervised pre-training, contrastive language-image pre-training (CLIP), and image-only self-supervised learning. Section 2.2 discusses supervised pre-training. Section 2.3 focuses on CLIP. Section 2.4 discusses image-only self-supervised learning, including contrastive learning, non-contrastive learning, and masked image modeling. Given the various learning approaches to training vision foundation models, Section 2.5 reviews how they can be incorporated for better performance. Lastly, Section 2.6 discusses how vision foundation models can be used for finer-grained visual understanding tasks, such as fusion-encoder-based pre-training for image captioning and visual question answering that require multimodal fusion, region-level pre-training for grounding, and pixel-level pre-training for segmentation.

2.1 Overview

There is a vast amount of literature on various methods of learning general-purpose vision backbones. As illustrated in Figure 2.1, we group these methods into three categories, depending on the types of supervision signals used to train the models, including:

- **Label supervision:** Arguably, the most well-studied image representation learning methods are based on label supervisions (typically in the form of image classification) (Sun et al., 2017), where datasets like ImageNet (Krizhevsky et al., 2012) and ImageNet21K (Ridnik et al., 2021) have been popular, and larger-scale proprietary datasets are also used in industrial labs (Sun et al., 2017; Singh et al., 2022b; Zhai et al., 2022a; Wu et al., 2023d).
- **Language supervision:** Another popular approach to learning image representations leverages weakly supervised signals from text, which is easy to acquire in large scale. For instance, CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) are pre-trained using a contrastive loss and billions of image-text pairs mined from the internet. The resultant models achieve strong zero-shot performance on image classification and image-text retrieval, and the learned image and text encoders have been widely used for various downstream tasks and allow traditional computer vision models to perform open-vocabulary CV tasks (Gu et al., 2021; Ghiasi et al., 2022a; Qian et al., 2022; Ding et al., 2022b; Liang et al., 2023a; Zhang et al., 2023e; Zou et al., 2023a; Minderer et al., 2022).

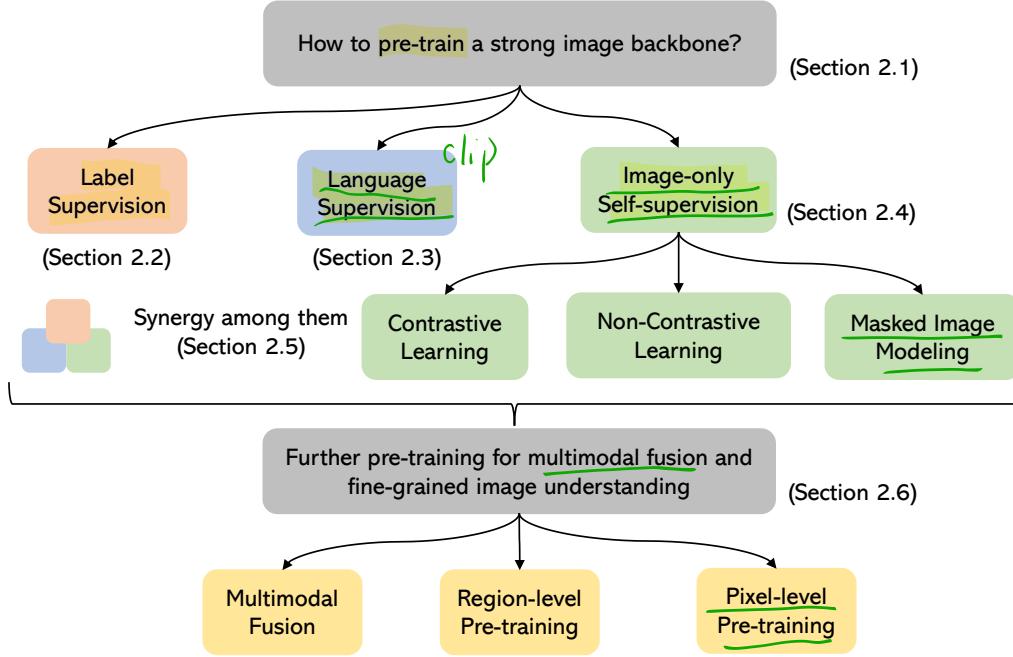


Figure 2.1: An overview of the structure of Chapter 2.

- **Image-only self-supervision:** There is also a vast amount of literature on exploring image-only self-supervised learning methods to learn image representations. As the name indicates, the supervision signals are mined from the images themselves, and popular methods range from contrastive learning (Chen et al., 2020a; He et al., 2020), non-contrastive learning (Grill et al., 2020; Chen and He, 2021; Caron et al., 2021), to masked image modeling (Bao et al., 2022; He et al., 2022a).

An illustration of these learning methods is shown in Figure 2.2. Besides the methods of pre-training image backbones, we will also discuss pre-training methods that allow multimodal fusion (*e.g.*, CoCa (Yu et al., 2022a), Flamingo (Alayrac et al., 2022)), region-level and pixel-level image understanding (*e.g.*, GLIP (Li et al., 2022e) and SAM (Kirillov et al., 2023)). These methods typically rely on a pre-trained image encoder or a pre-trained image-text encoder pair. Figure 2.3 shows an overview of the topics covered in this chapter and some representative works in each topic.

2.2 Supervised Pre-training

Supervised pre-training on large-scale human-labeled datasets, such as ImageNet (Krizhevsky et al., 2012) and ImageNet21K (Ridnik et al., 2021), has emerged as a widely adopted approach to acquiring transferable visual representations. It aims to map an image to a discrete label, which is associated with a visual concept. This approach has greatly expedited progress in designing various vision backbone architectures (*e.g.*, AlexNet (Krizhevsky et al., 2012), ResNet (He et al., 2016), vision transformer (Dosovitskiy et al., 2021), and Swin transformer (Liu et al., 2021)), and is the testbed for all the modern vision backbones. It also powered computer vision tasks across the whole spectrum, ranging from image classification, object detection/segmentation, visual question answering, image captioning, to video action recognition. However, the effectiveness of learned representations is often limited by the scale and diversity of supervisions in pre-training datasets, as human annotation is expensive.

Large-scale datasets. For larger-scale pre-training, noisy labels can be derived in large quantities from image-text pairs crawled from the Web. Using noisy labels, many industrial labs have successfully constructed comprehensive classification datasets using semi-automatic pipelines, such as JFT (Sun et al., 2017; Zhai et al., 2022a) and I2E (Wu et al., 2023d), or by leveraging proprietary data like Instagram hashtags (Singh et al., 2022b). The statistics of existing large-scale image clas-

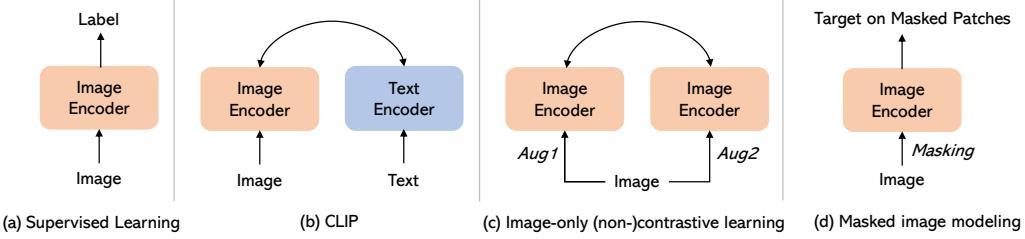


Figure 2.2: A high-level overview of different approaches to learn general image representations, including supervised learning (Krizhevsky et al., 2012), contrastive language-image pre-training (Radford et al., 2021; Jia et al., 2021), and image-only self-supervised learning, including contrastive learning (Chen et al., 2020a; He et al., 2020), non-contrastive learning (Grill et al., 2020; Chen and He, 2021), and masked image modeling (Bao et al., 2022; He et al., 2022a).

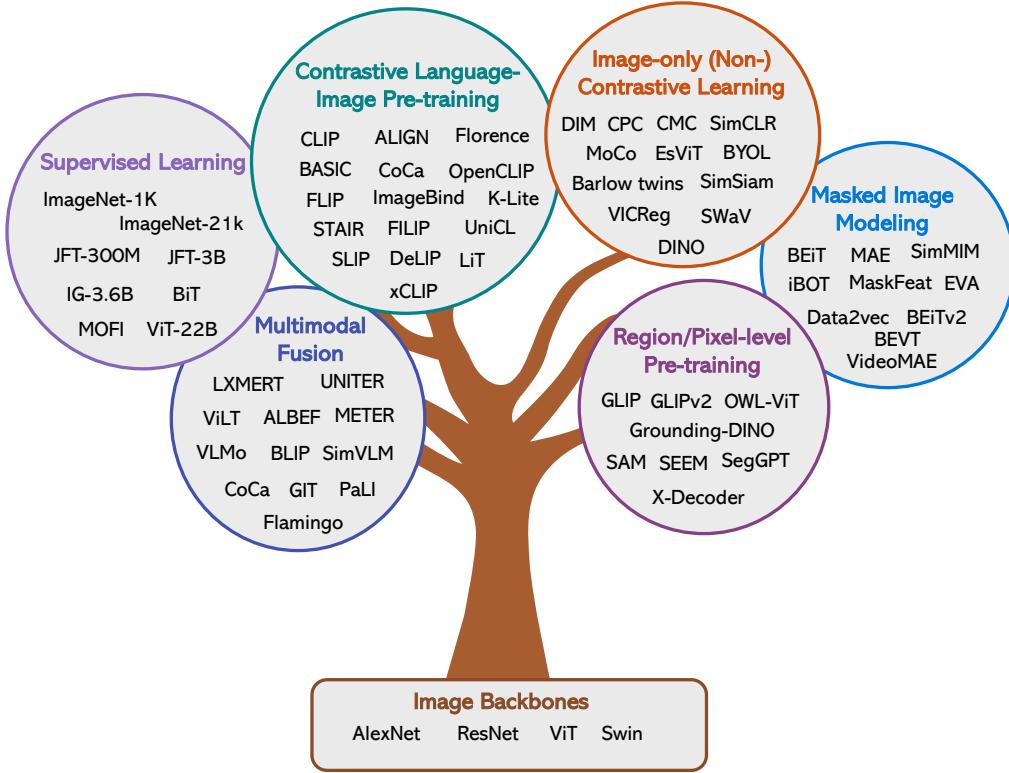


Figure 2.3: An overview of the topics covered in this chapter and representative works in each topic. We start from supervised learning and CLIP, and then move on to image-only self-supervised learning, including contrastive learning, non-contrastive learning, and masked image modeling. Lastly, we discuss pre-training methods that empower multimodal fusion, region-level and pixel-level image understanding.

sification datasets are shown in Table 2.1. The labels are typically in the form of fine-grained image entities with a long-tailed distribution. Though classical, this approach has been very powerful for learning universal image representations. For example, JFT-300M (Sun et al., 2017) has been used for training the BiT (“Big Transfer”) models (Kolesnikov et al., 2020), and JFT-3B (Zhai et al., 2022a) has been used to scale up the training of a plain vision transformer (Dosovitskiy et al., 2021) to 22B in model size. LiT (Zhai et al., 2022b) proposes to first learn the image backbone on JFT-3B (Zhai et al., 2022a), and keep it frozen and learn another text tower to align the image and text embedding space to make the model open-vocabulary and is capable of performing zero-shot image classification.

Dataset	# Images	# Classes
ImageNet-1K (Russakovsky et al., 2015)	1.2M	1K
ImageNet-21K (Ridnik et al., 2021)	14M	21K
JFT-300M (Sun et al., 2017)	300M	18K
JFT-3B (Zhai et al., 2022a)	3B	30K
IG-3.6B (Singh et al., 2022b)	3.6B	27K
I2E (Wu et al., 2023d)	1.1B	2M

Table 2.1: Statistics of existing large-scale image classification datasets.

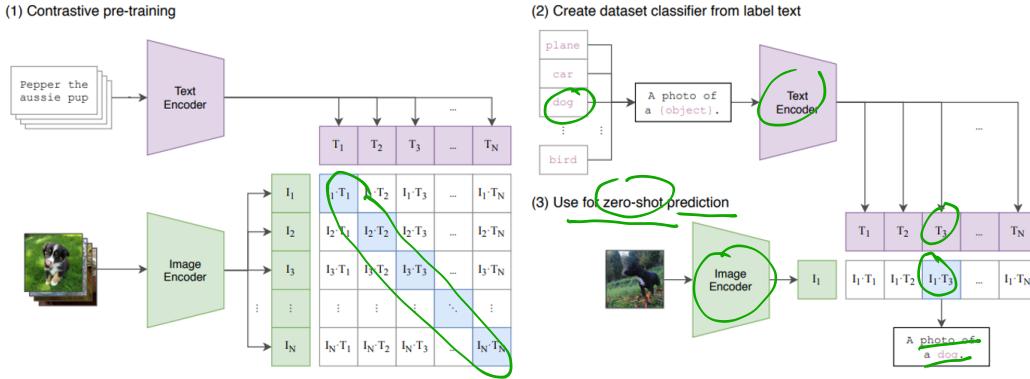


Figure 2.4: Illustration of contrastive language-image pre-training, and how the learned model can be used for zero-shot image classification. Image credit: Radford et al. (2021).

Model training. There are many loss functions that can be used to promote embedding properties (*e.g.*, separability) (Musgrave et al., 2020). For example, the large margin loss (Wang et al., 2018) is used for MOFI training (Wu et al., 2023d). Furthermore, if the datasets have an immense number of labels (can potentially be over 2 million as in MOFI (Wu et al., 2023d)), predicting all the labels in each batch becomes computationally costly. In this case, a fixed number of labels is typically used for each batch, similar to sampled softmax (Gutmann and Hyvärinen, 2010).

2.3 Contrastive Language-Image Pre-training

2.3.1 Basics of CLIP Training

Language is a richer form of supervision than classical closed-set labels. Rather than deriving noisy label supervision from web-crawled image-text datasets, the alt-text can be directly used for learning transferable image representations, which is the spirit of contrastive language-image pre-training (CLIP) (Radford et al., 2021). In particular, models trained in this way, such as ALIGN (Jia et al., 2021), Florence (Yuan et al., 2021), BASIC (Pham et al., 2021), and OpenCLIP (Ilharco et al., 2021), have showcased impressive zero-shot image classification and image-text retrieval capabilities by mapping images and text into a shared embedding space. Below, we discuss how the CLIP model is pre-trained and used for zero-shot prediction.

- **Training:** As shown in Figure 2.4(1), CLIP is trained via simple contrastive learning. CLIP is an outstanding example of “*simple algorithms that scale well*” (Li et al., 2023m). To achieve satisfactory performance, model training needs to be scaled along three dimensions: batch size, data size, and model size (Pham et al., 2021). Specifically, the typical batch size used for CLIP training can be 16k or 32k. The number of image-text pairs in the pre-training datasets is frequently measured in billions rather than millions. A vision transformer trained in this fashion can typically vary from 300M (Large) to 1B (giant) in model size.
- **Zero-shot prediction:** As shown in Figure 2.4 (2) and (3), CLIP empowers zero-shot image classification via reformatting it as a retrieval task and considering the semantics behind labels. It

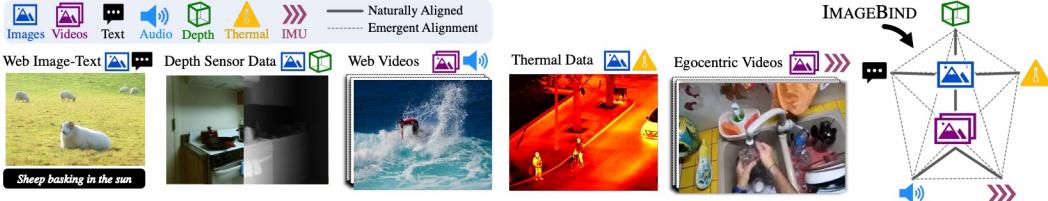


Figure 2.5: ImageBind (Girdhar et al., 2023) proposes to link a total of six modalities into a common embedding space via leveraging pre-trained CLIP models, enabling new emergent alignments and capabilities. Image credit: Girdhar et al. (2023).

can also be used for zero-shot image-text retrieval by its design. Besides this, the aligned image-text embedding space makes it possible to make all the traditional vision models open vocabulary and has inspired a rich line of work on open-vocabulary object detection and segmentation (Li et al., 2022e; Zhang et al., 2022b; Zou et al., 2023a; Zhang et al., 2023e).

2.3.2 CLIP Variants

Since the birth of CLIP, there have been tons of follow-up works to improve CLIP models, as to be discussed below. We do not aim to provide a comprehensive literature review of all the methods, but focus on a selected set of topics.

Data scaling up. Data is the fuel for CLIP training. For example, OpenAI’s CLIP was trained on 400M image-text pairs mined from the web, while ALIGN used a proprietary dataset consisting of 1.8B image-text pairs. In BASIC (Pham et al., 2021), the authors have carefully studied the scaling among three dimensions: batch size, data size, and model size. However, most of these large-scale datasets are not publicly available, and training such models requires massive computing resources.

In academic settings, researchers (Li et al., 2022b) have advocated the use of a few millions of image-text pairs for model pre-training, such as CC3M (Sharma et al., 2018), CC12M (Changpinyo et al., 2021), YFCC (Thomee et al., 2016). Relatively small-scale image-text datasets that are publicly available include SBU (Ordonez et al., 2011), RedCaps (Desai et al., 2021), and WIT (Srinivasan et al., 2021). Large-scale public available image-text datasets include Shutterstock (Nguyen et al., 2022), LAION-400M (Schuhmann et al., 2021), COYO-700M (Byeon et al., 2022), and LAION-2B (Schuhmann et al., 2022), to name a few. For example, LAION-2B (Schuhmann et al., 2022) has been used by researchers to study the reproducible scaling laws for CLIP training (Cherti et al., 2023).

Interestingly, in search of the next-generation image-text datasets, in DataComp (Gadre et al., 2023), instead of fixing the dataset and designing different algorithms, the authors propose to select and rank datasets using the fixed CLIP training method. Besides paired image-text data mined from the Web for CLIP training, inspired by the interleaved image-text dataset M3W introduced in Flamingo (Alayrac et al., 2022), there have been recent efforts of collecting interleaved image-text datasets, such as MMC4 (Zhu et al., 2023b) and OBELISC (Laurençon et al., 2023).

Model design and training methods. CLIP training has been significantly improved. Below, we review some representative works.

- **Image tower:** On the image encoder side, FLIP (Li et al., 2023m) proposes to scale CLIP training via masking. By randomly masking out image patches with a high masking ratio, and only encoding the visible patches as in MAE (He et al., 2022a), the authors demonstrate that masking can improve training efficiency without hurting the performance. The method can be adopted for all CLIP training. Cao et al. (2023) found that filtering out samples that contain text regions in the image improves CLIP training efficiency and robustness.
- **Language tower:** On the language encoder side, K-Lite (Shen et al., 2022a) proposes to use external knowledge in the form of Wiki definition of entities together with the original alt-text for contrastive pre-training. Empirically, the use of enriched text descriptions improves the CLIP

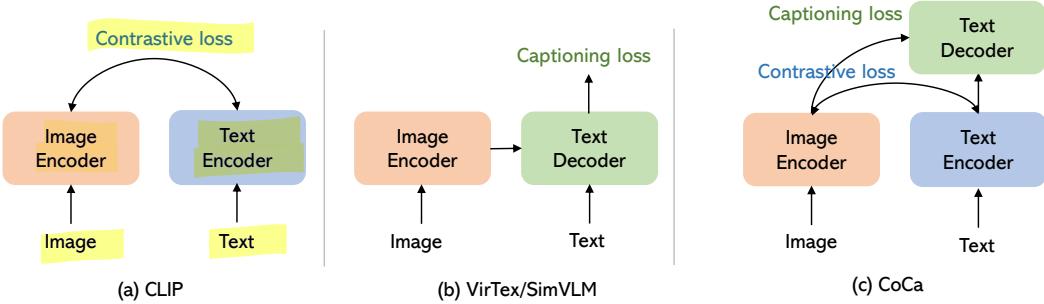


Figure 2.6: A high-level comparison of contrastive loss and captioning loss for image encoder pre-training. (a) CLIP (Radford et al., 2021) uses contrastive loss alone for pre-training, which enables zero-shot image classification and has demonstrated strong scaling behavior. (b) VirTex (Desai and Johnson, 2021) uses captioning loss alone for pre-training. SimVLM (Wang et al., 2022g) uses prefix language modeling for pre-training in a much larger scale. The model architecture is similar to multimodal language models (*e.g.*, GIT (Wang et al., 2022a) and Flamingo (Alayrac et al., 2022)), but VirTex and SimVLM aim to pre-train the image encoder from scratch. (c) CoCa (Yu et al., 2022a) uses both contrastive and captioning losses for pre-training. The model architecture is similar to ALBEF (Li et al., 2021b), but CoCa aims to pre-train the image encoder from scratch, instead of using a pre-trained one.

performance. LaCLIP (Fan et al., 2023a) shows that CLIP can be improved via rewriting the noisy and short alt-text using large language models such as ChatGPT.

- **Interpretability:** The image representation is typically a dense feature vector. In order to improve the interpretability of the shared image-text embedding space, STAIR (Chen et al., 2023a) proposes to map images and text to a high-dimensional, sparse, embedding space, where each dimension in the sparse embedding is a (sub-)word in a large dictionary in which the predicted non-negative scalar corresponds to the weight associated with the token. The authors show that STAIR achieves better performance than the vanilla CLIP with improved interpretability.
- **More modalities:** The idea of contrastive learning is general, and can go beyond just image and text modalities. For example, as shown in Figure 2.5, ImageBind (Girdhar et al., 2023) proposes to encode six modalities into a common embedding space, including images, text, audio, depth, thermal, and IMU modalities. In practice, a pre-trained CLIP model is used and kept frozen during training, which indicates that other modality encoders are learned to align to the CLIP embedding space, so that the trained model can be applied to new applications such as audio-to-image generation and multimodal LLMs (*e.g.*, PandaGPT (Su et al., 2023)).

Objective function. The use of contrastive loss alone is powerful, especially when the model is scaled up. However, other objective functions can also be applied.

- **Fine-grained supervision:** Instead of using a simple dot-product to calculate the similarity of an image-text pair, the supervision can be made more fine-grained via learning word-patch alignment. In FILIP (Yao et al., 2022b), the authors propose to first compute the loss by calculating the token-wise similarity, and then aggregating the matrix by max-pooling for word-patch alignment.
- **Contrastive captioner:** Besides the contrastive learning branch, CoCa (Yu et al., 2022a) (shown in Figure 2.6(c)) adds a generative loss to improve performance and allow new capabilities that require multimodal fusion (*e.g.*, image captioning and VQA). This is similar to many fusion-encoder-based vision-language models such as ALBEF (Li et al., 2021b), but with the key difference in that CoCa aims to learn a better image encoder from scratch. A detailed discussion on multimodal fusion is in Section 2.6.1.
- **Captioning loss alone:** How about using the captioning loss alone to pre-train an image encoder? Actually, before CLIP was invented, VirTex (Desai and Johnson, 2021) (shown in Figure 2.6(b)) and ICMLM (Sariyildiz et al., 2020) learn encoders using a single image captioning loss, but the scale is very small (restricted to COCO images) and the performance is poor. CLIP also shows that contrastive pre-training is a much better choice. In SimVLM (Wang et al., 2022g), the authors

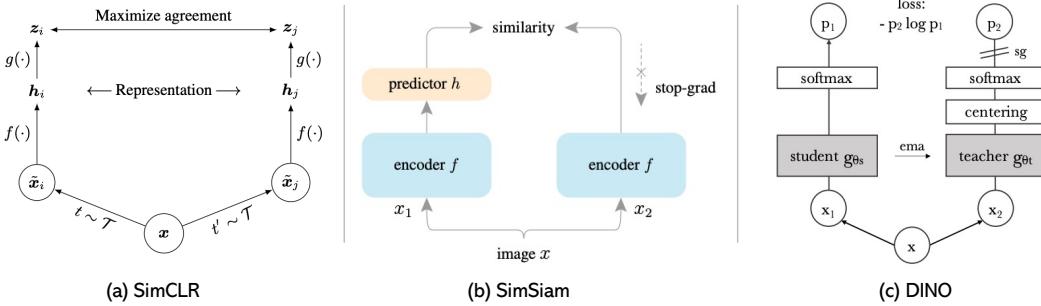


Figure 2.7: Overview of SimCLR (Chen et al., 2020a), SimSiam (Chen and He, 2021), and DINO (Caron et al., 2021) for self-supervised image representation learning. SimCLR uses contrastive learning for model training, while SimSiam and DINO explores non-contrastive learning methods. Image credit: Chen et al. (2020a), Chen and He (2021), Caron et al. (2021).

found that the learned image encoder was not as competitive as CLIP. However, in the recent work Cap/CapPa (Tschanne et al., 2023), the authors argue that image captioners are scalable vision learners, too. Captioning can exhibit the same or even better scaling behaviors.

- **Sigmoid loss for language-image pre-training:** Unlike standard contrastive learning with softmax normalization, Zhai et al. (2023) uses a simple pairwise sigmoid loss for image-text pre-training, which operates on image-text pairs and does not require a global view of the pairwise similarities for normalization. The authors show that the use of simple sigmoid loss can also achieve strong performance on zero-shot image classification.

2.4 Image-Only Self-Supervised Learning

Now, we shift our focus to image-only self-supervised learning, and divide the discussion into three parts: (i) contrastive learning, (ii) non-contrastive learning, and (iii) masked image modeling.

2.4.1 Contrastive and Non-contrastive Learning

Contrastive learning. The core idea of contrastive learning (Gutmann and Hyvärinen, 2010; Arora et al., 2019) is to promote the positive sample pairs and repulse the negative sample pairs. Besides being used in CLIP, contrastive learning has also been a popular concept in self-supervised image representation learning (Wu et al., 2018; Ye et al., 2019b; Tian et al., 2020a; Chen et al., 2020a; He et al., 2020; Misra and Maaten, 2020; Chen et al., 2020c). It has been shown that the contrastive objective, known as the InfoNCE loss (Oord et al., 2018), can be interpreted as maximizing the lower bound of mutual information between different views of the data (Hjelm et al., 2018; Bachman et al., 2019; Henaff, 2020).

In a nutshell, all the image-only contrastive learning methods (*e.g.*, SimCLR (Chen et al., 2020a), see Figure 2.7(a), MoCo (He et al., 2020), SimCLR-v2 (Chen et al., 2020b), MoCo-v2 (Chen et al., 2020c)) share the same high-level framework, detailed below.

- Given one image, two separate data augmentations are applied;
- A base encoder is followed by a project head, which is trained to maximize agreement using a contrastive loss (*i.e.*, they are from the same image or not);
- The project head is thrown away for downstream tasks.

However, a caveat of contrastive learning is the requirement of a large number of negative samples. These samples can be maintained in a memory bank (Wu et al., 2018), or directly from the current batch (Chen et al., 2020a), which suggests the requirement of a large batch size. MoCo (He et al., 2020) maintains a queue of negative samples and turns one branch into a momentum encoder to improve the consistency of the queue. Initially, contrastive learning was primarily studied for pre-training convolutional networks. However, with the rising popularity of vision transformers (ViT),

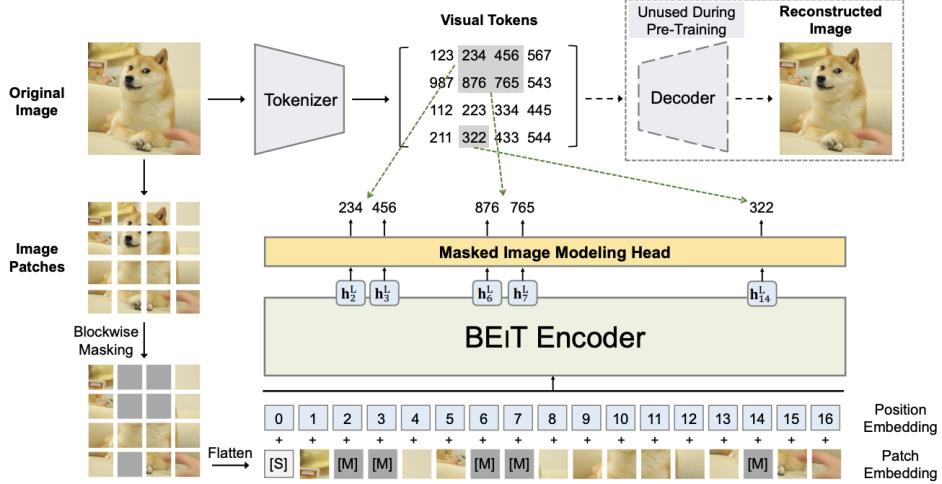


Figure 2.8: Overview of BEiT pre-training for image transformers. Image credit: [Bao et al. \(2022\)](#).

researchers have also explored its application in the context of ViT. ([Chen et al., 2021b](#); [Li et al., 2021a](#); [Xie et al., 2021](#)).

Non-contrastive learning. Recent self-supervised learning methods do not depend on negative samples. The use of negatives is replaced by asymmetric architectures (*e.g.*, BYOL ([Grill et al., 2020](#)), SimSiam ([Chen and He, 2021](#))), dimension de-correlation (*e.g.*, Barlow twins ([Zbontar et al., 2021](#)), VICReg ([Bardes et al., 2021](#)), Whitening ([Ermolov et al., 2021](#))), and clustering (*e.g.*, SWaV ([Caron et al., 2020](#)), DINO ([Caron et al., 2021](#)), [Caron et al. \(2018\)](#); [Amrani et al. \(2022\)](#); [Assran et al. \(2022\)](#); [Wang et al. \(2023b\)](#)), *etc.*

For example, as illustrated in Figure 2.7(b), in SimSiam ([Chen and He, 2021](#)), two augmented views of a single image are processed by an identical encoder network. Subsequently, a prediction MLP is applied to one view, while a stop-gradient operation is employed on the other. The primary objective of this model is to maximize the similarity between the two views. It is noteworthy that SimSiam relies on neither negative pairs nor a momentum encoder.

Another noteworthy method, known as DINO ([Caron et al., 2021](#)) and illustrated in Figure 2.7(c), takes a distinct approach. DINO involves feeding two distinct random transformations of an input image into both the student and teacher networks. Both networks share the same architecture but have different parameters. The output of the teacher network is centered by computing the mean over the batch. Each network outputs a feature vector that is normalized with a temperature softmax applied to the feature dimension. The similarity between these features is quantified using a cross-entropy loss. Additionally, a stop-gradient operator is applied to the teacher network to ensure that gradients propagate exclusively through the student network. Moreover, DINO updates the teacher's parameters using an exponential moving average of the student's parameters.

2.4.2 Masked Image Modeling

Masked language modeling ([Devlin et al., 2019](#)) is a powerful pre-training task that has revolutionized the NLP research. To mimic the success of BERT pre-training for NLP, the pioneering work BEiT ([Bao et al., 2022](#)), as illustrated in Figure 2.8, proposes to perform masked image modeling (MIM) to pre-train image transformers. Specifically,

- **Image tokenizer:** In order to perform masked token prediction, an image tokenizer is required to tokenize an image into discrete visual tokens, so that these tokens can be treated just like an additional set of language tokens. Some well-known learning methods for image tokenizers include VQ-VAE ([van den Oord et al., 2017](#)), VQ-VAE-2 ([Razavi et al., 2019](#)), VQ-GAN ([Esser et al., 2021](#)), ViT-VQGAN ([Yu et al., 2021](#)), *etc.* These image tokenizers have also been widely used for

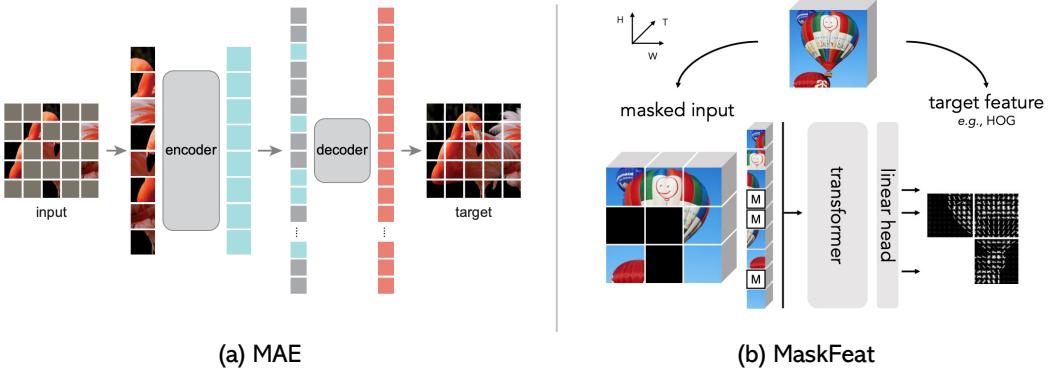


Figure 2.9: Illustration of Masked Autoencoder (MAE) (He et al., 2022a) that uses raw pixel values for MIM training, and MaskFeat (Wei et al., 2021) that uses different features as the targets. HOG, a hand-crafted feature descriptor, was found to work particularly well in terms of both performance and efficiency. Image credit: He et al. (2022a) and Wei et al. (2021).

autoregressive image generation, such as DALLE (Ramesh et al., 2021a), Make-A-Scene (Gafni et al., 2022), Parti (Yu et al., 2022b), to name a few.

- **Mask-then-predict:** The idea of MIM is conceptually simple: models accept the corrupted input image (*e.g.*, via random masking of image patches), and then predict the target of the masked content (*e.g.*, discrete visual tokens in BEiT). As discussed in iBOT (Zhou et al., 2021), this training procedure can be understood as knowledge distillation between the image tokenizer (which serves as the teacher) and the BEiT encoder (which serves as the student), while the student only sees partial of the image.

Targets. In Peng et al. (2022b), the authors have provided a unified view of MIM: a teacher model, a normalization layer, a student model, an MIM head, and a proper loss function. The most significant difference among all these models lies in the reconstruction targets, which can be pixels, discrete image tokens, features from pre-trained models, and outputs from the momentum updated teacher. Specifically, the targets can be roughly grouped into two categories.

- **Low-level pixels/features as targets:** MAE (He et al., 2022a), SimMIM (Xie et al., 2022b), ConvMAE (Gao et al., 2022), HiViT (Zhang et al., 2022d), and GreenMIM (Huang et al., 2022a) leverage either original or normalized pixel values as the target for MIM. These methods have typically explored the use of a plain Vision Transformer (Dosovitskiy et al., 2021) or the Swin Transformer (Liu et al., 2021) as the backbone architecture. MaskFeat (Wei et al., 2021) introduced the Histogram of Oriented Gradients (HOG) feature descriptor as the target for MIM (see Figure 2.9(b)). Meanwhile, Ge²-AE (Liu et al., 2023b) employed both pixel values and frequency information obtained from the 2D discrete Fourier transform as the target. Taking MAE (He et al., 2022a) as an example (Figure 2.9(a)), the authors show that using pixel values as targets works particularly well. Specifically, a large random subset of images (*e.g.*, 75%) is masked out; then, the image encoder is only applied to visible patches, while mask tokens are introduced after the encoder. It was shown that such pre-training is especially effective for object detection and segmentation tasks, which require finer-grained image understanding.
- **High-level features as targets:** BEiT (Bao et al., 2022), CAE (Chen et al., 2022g), SplitMask (El-Nouby et al., 2021), and PeCo (Dong et al., 2023) involve the prediction of discrete tokens using learned image tokenizers. MaskFeat (Wei et al., 2021) takes a different approach by proposing direct regression of high-level features extracted from models like DINO (Caron et al., 2021) and DeiT (Touvron et al., 2021). Expanding this idea, MVP (Wei et al., 2022b) and EVA (Fang et al., 2023) make feature prediction using image features from CLIP as target features. Additionally, other methods such as data2vec (Baevski et al., 2022), MSN (Assran et al., 2022), ConMIM (Yi et al., 2022), SIM (Tao et al., 2023), and BootMAE (Dong et al., 2022) propose to construct regression feature targets by leveraging momentum-updated teacher models to enhance online learning. The choice of loss functions depends on the nature of the targets: cross-entropy loss is

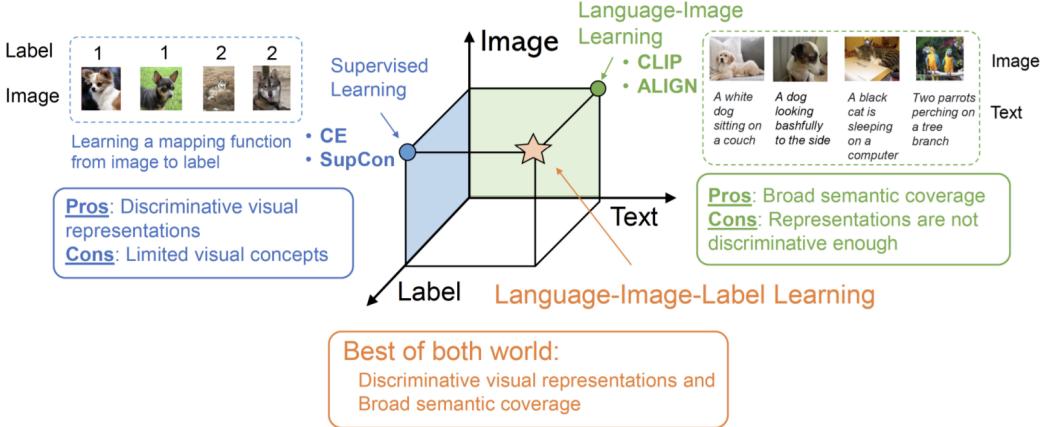


Figure 2.10: Overview of UniCL (Yang et al., 2022a) that performs unified contrastive pre-training on image-text and image-label data. Image credit: Yang et al. (2022a).

typically used when the targets are discrete tokens, while ℓ_1 , ℓ_2 , or cosine similarity losses are common choices for pixel values or continuous-valued features.

MIM for video pre-training. Naturally, there are recent works on extending MIM to video pre-training. Prominent examples include BEVT (Wang et al., 2022c), MAE as spatiotemporal learner (Feichtenhofer et al., 2022), VideoMAE (Tong et al., 2022), and VideoMAEv2 (Wang et al., 2023e). Taking Feichtenhofer et al. (2022) as an example. This paper studies a conceptually simple extension of MAE to video pre-training via randomly masking out space-time patches in videos and learns an autoencoder to reconstruct them in pixels. Interestingly, the authors found that MAE learns strong video representations with almost no inductive bias on space-time, and spacetime-agnostic random masking performs the best, with an optimal masking ratio as high as 90%.

Lack of learning global image representations. MIM is an effective pre-training method that provides a good parameter initialization for further model finetuning. However, the vanilla MIM pre-trained model does not learn a global image representation. In iBOT (Zhou et al., 2021), the authors propose to enhance BEiT (Bao et al., 2022) with a DINO-like self-distillation loss (Caron et al., 2021) to force the [CLS] token to learn global image representations. The same idea has been extended to DINoV2 (Quab et al., 2023).

Scaling properties of MIM. MIM is scalable in terms of model size. For example, we can perform MIM pre-training of a vision transformer with billions of parameters. However, the scaling property with regard to data size is less clear. There are some recent works that aim to understand the data scaling of MIM (Xie et al., 2023b; Lu et al., 2023a); however, the data scale is limited to millions of images, rather than billions, except Singh et al. (2023) that studies the effectiveness of MAE as a so-called “pre-pretraining” method for billion-scale data. Generally, MIM can be considered an effective regularization method that helps initialize a billion-scale vision transformer for downstream tasks; however, whether or not scaling the MIM pre-training to billion-scale image-only data requires further exploration.

2.5 Synergy Among Different Learning Approaches

Till now, we have reviewed different approaches to pre-training image backbones, especially for vision transformers. Below, we use CLIP as the anchor point, and discuss how CLIP can be combined with other learning methods.

Combining CLIP with label supervision. Noisy labels and text supervision can be jointly used for image backbone pre-training. Some representative works are discussed below.

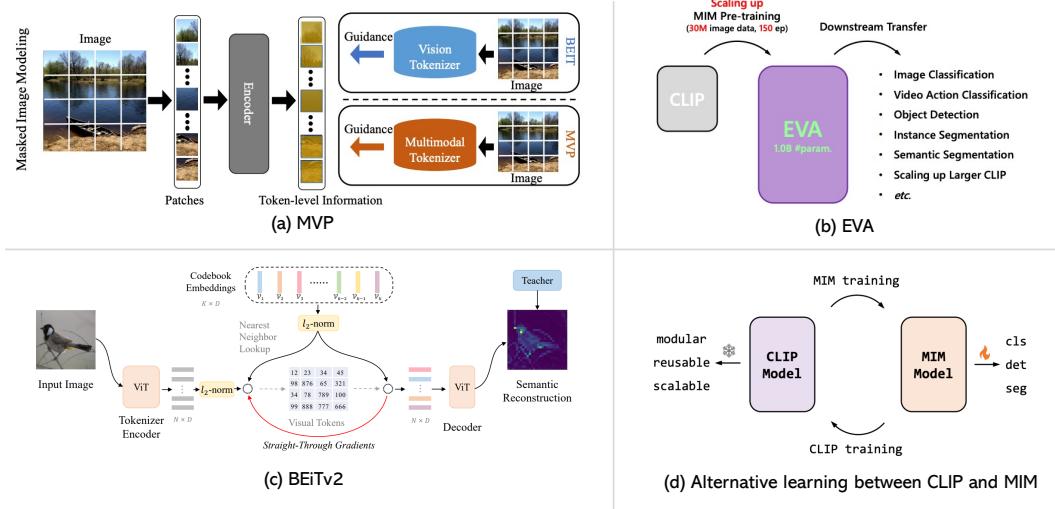


Figure 2.11: Illustration of MVP (Wei et al., 2022b), EVA (Fang et al., 2023) and BEiTv2 (Peng et al., 2022a). (a) & (b) MVP and EVA directly regress CLIP features for MIM pre-training. (c) BEiTv2 compresses the information inside CLIP features into discrete visual tokens, and then performing regular BEiT training. (d) Alternative learning between CLIP and MIM. Image credit: Wei et al. (2022b), Fang et al. (2023), Peng et al. (2022a), Fang et al. (2023).

- UniCL (Yang et al., 2022a) proposes a principled way to use image-label and image-text data together in a joint image-text-label space for unified contrastive learning, and Florence (Yuan et al., 2021) is a scaled-up version of UniCL. See Figure 2.10 for an illustration of the framework.
- LiT (Zhai et al., 2022b) uses a pre-trained ViT-g/14 image encoder learned from supervised pre-training on the JFT-3B dataset, and then makes the image encoder open-vocabulary by learning an additional text tower via contrastive pre-training on image-text data. Essentially, LiT teaches a text model to read out good representations from a pre-trained image model for new tasks.
- MOFI (Wu et al., 2023d) proposes to learn image representations from 1 billion noisy entity-annotated images, and uses both image classification and contrastive losses for model training. For image classification, entities associated with each image are considered as labels, and supervised pre-training on a large number of entities is conducted; for contrastive pre-training, entity names are treated as free-form text, and are further enriched with entity descriptions.

Combining CLIP with image-only (non-)contrastive learning. CLIP can also be enhanced with image-only self-supervision. Specifically,

- SLIP (Mu et al., 2021) proposes a conceptually simple idea to combine SimCLR (Chen et al., 2020a) and CLIP for model training, and shows that SLIP outperforms CLIP on both zero-shot transfer and linear probe settings. DeCLIP (Li et al., 2022g) mines self-supervised learning signals on each modality to make CLIP training data-efficient. In terms of image supervision, the SimSam framework (Chen and He, 2021) is used.
- xCLIP (Zhou et al., 2023c) makes CLIP non-contrastive via introducing additional sharpness and smoothness regularization terms borrowed from the image-only non-contrastive learning literature. However, the authors show that only non-contrastive pre-training (nCLIP) is not sufficient to achieve strong performance on zero-shot image classification, and it needs to be combined with the original CLIP for enhanced performance.

Combining CLIP with MIM. There are recent works that aim to combine CLIP and MIM for model training. We group them into two categories.

- **Shallow interaction.** It turns out that image features extracted from CLIP are a good target for MIM training, as the CLIP image features potentially capture the semantics that are missing in MIM training. Along this line of work, as shown in Figure 2.11, MVP (Wei et al., 2022b)

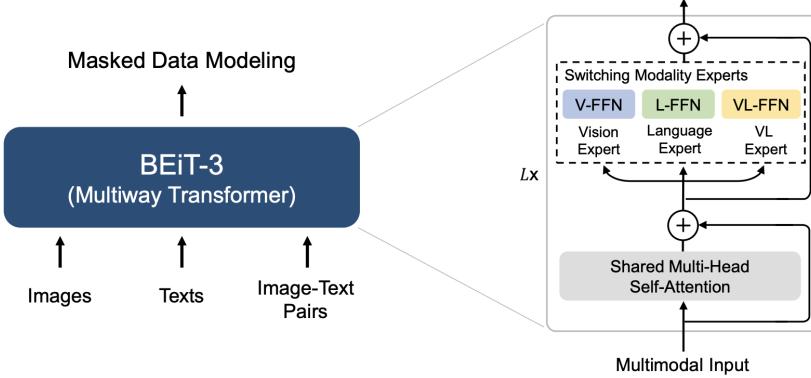


Figure 2.12: Overview of BEiT-3 that performs masked data modeling on both image/text and joint image-text data via a multiway transformer. Image credit: Wang et al. (2022d).

proposes to regress CLIP features directly, while BEiTv2 (Peng et al., 2022a) first compresses the information inside CLIP features into discrete visual tokens, and then performs regular BEiT training. Similar use of CLIP features as MIM training target has also been investigated in EVA (Fang et al., 2023), CAEv2 (Zhang et al., 2022c), and MaskDistill (Peng et al., 2022b). In EVA-02 (Fang et al., 2023), the authors advocate alternative learning of MIM and CLIP representations. Specifically, an off-the-shelf CLIP model is used to provide a feature target for MIM training; while the MIM pre-trained image backbone is used to initialize CLIP training. The MIM representations are used to finetune various downstream tasks while the learned frozen CLIP embedding enables zero-shot image classification and other applications.

- **Deeper integration.** However, instead of using CLIP as targets for MIM training, if one aims to combine CLIP and MIM for joint model training, MIM does not seem to improve a CLIP model at scale (Weers et al., 2023; Li et al., 2023m).
- Although the combination of CLIP and MIM does not lead to a promising result at the current stage, the combination of BERT and BEiT is very promising, as evidenced in BEiT-3 (Wang et al., 2022d) (see Figure 2.12), where the authors show that masked data modeling can be performed on both image/text and joint image-text data via the design of a multiway transformer, and state-of-the-art performance can be achieved on a wide range of vision and vision-language tasks.

2.6 Multimodal Fusion, Region-Level and Pixel-Level Pre-training

Till now, we have focused on the methods of pre-training image backbones from scratch, but not on pre-training methods that power multimodal fusion, region-level and pixel-level image understanding. These methods typically use a pre-trained image encoder at the first hand to perform a second-stage pre-training. Below, we briefly discuss these topics.

2.6.1 From Multimodal Fusion to Multimodal LLM

For dual encoders such as CLIP (Radford et al., 2021), image and text are encoded separately, and modality interaction is only handled via a simple dot product of image and text feature vectors. This can be very effective for zero-shot image classification and image-text retrieval. However, due to the lack of deep multimodal fusion, CLIP alone performs poorly on the image captioning (Vinyals et al., 2015) and visual question answering (Antol et al., 2015) tasks. This requires the pre-training of a fusion encoder, where additional transformer layers are typically employed to model the deep interaction between image and text representations. Below, we review how these fusion-encoder pre-training methods are developed over time.

OD-based models. Most early methods use pre-trained object detectors (ODs) to extract visual features. Among them, ViLBERT (Lu et al., 2019) and LXMERT (Tan and Bansal, 2019) use co-attention for multimodal fusion, while methods like VisualBERT (Li et al., 2019b), Unicoder-VL (Li et al., 2020a), VL-BERT (Su et al., 2019), UNITER (Chen et al., 2020d), OSCAR (Li et al., 2020b),

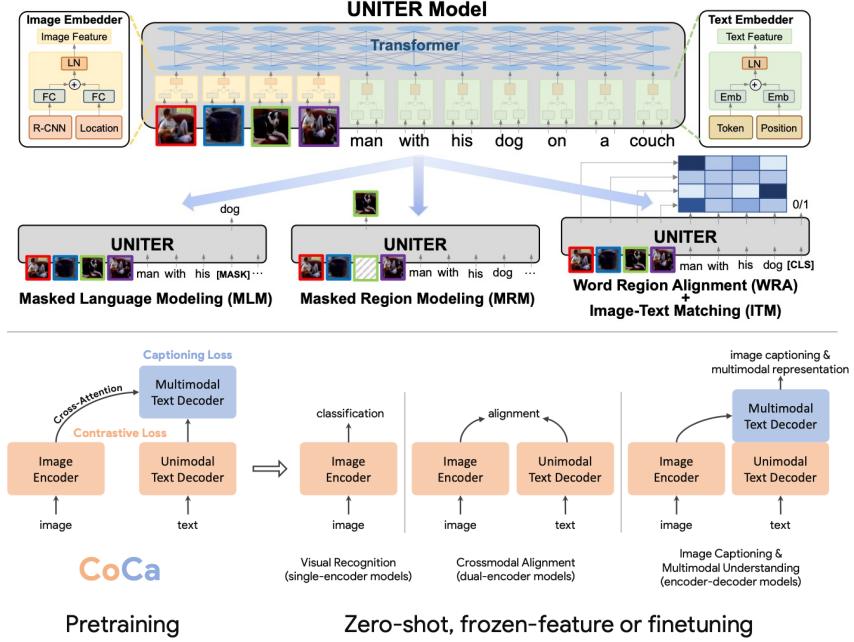


Figure 2.13: Illustration of UNITER (Chen et al., 2020d) and CoCa (Yu et al., 2022a), which serve as a classical and a modern model that performs pre-training on multimodal fusion. CoCa also pre-trains the image backbone from scratch. Specifically, UNITER extracts image features via an off-the-shelf object detector and treat image features as soft prompts of the text input to be sent into a multimodal transformer. The model is pre-trained over a few millions of image-text pairs. For CoCa, an image encoder and a text encoder is used, with a multimodal transformer stacked on top. Both contrastive loss and captioning loss are used for model training, and the model is trained over billions of image-text pairs and JFT data. Image credit: Chen et al. (2020d), Yu et al. (2022a).

VILLA (Gan et al., 2020) and VinVL (Zhang et al., 2021) treat image features as soft prompts of the text input to be sent into a multimodal transformer.

End-to-end models. Now, end-to-end pre-training methods become the mainstream. Some early methods use CNNs to extract image features, such as PixelBERT (Huang et al., 2020), SOHO (Huang et al., 2021), and CLIP-ViT (Shen et al., 2022b), while ViLT (Kim et al., 2021) and ViTCAP (Fang et al., 2022) directly feed image patch features and text token embeddings into a multimodal transformer. Due to the popularity of vision transformer (ViT), now most methods simply use ViT as the image encoder (*e.g.*, plain ViT (Dosovitskiy et al., 2021) and Swin transformer (Liu et al., 2021)). Prominent examples include ALBEF (Li et al., 2021b), METER (Dou et al., 2022b), VLMo (Wang et al., 2021b), X-VLM (Zeng et al., 2022), BLIP (Li et al., 2022d), SimVLM (Wang et al., 2022g), FLAVA (Singh et al., 2022a) and CoCa (Yu et al., 2022a).

An illustration of UNITER (Chen et al., 2020d) and CoCa (Yu et al., 2022a) is shown in Figure 2.13. They serve as two examples of a classical model and a modern model, respectively, which performs pre-training on multimodal fusion. CoCa also performs image backbone pre-training directly, as all the model components are trained from scratch. Please refer to Chapter 3 of Gan et al. (2022) for a comprehensive literature review.

Trend to multimodal LLM. Instead of using masked language modeling, image-text matching and image-text contrastive learning, SimVLM (Wang et al., 2022g) uses a simple PrefixLM loss for pre-training. Since then, multimodal language models have become popular. Early models focus on large-scale pre-training, such as Flamingo (Alayrac et al., 2022), GIT (Wang et al., 2022a), PaLI (Chen et al., 2022h), PaLI-X (Chen et al., 2023g), while recent works focus on using pre-trained LLMs for instruction tuning, such as LLaVA (Liu et al., 2023c) and MiniGPT-4 (Zhu et al., 2023a). A detailed discussion on this topic is provided in Chapter 5.

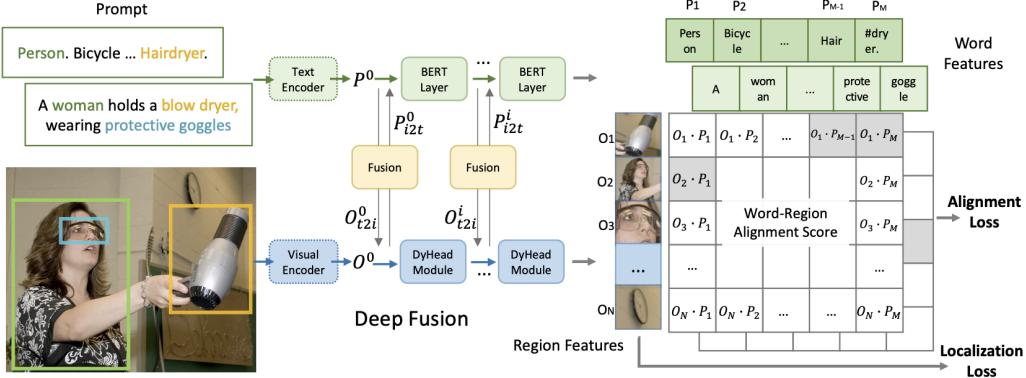


Figure 2.14: Overview of GLIP that performs grounded language-image pre-training for open-set object detection. Image credit: Li et al. (2022f).

2.6.2 Region-Level Pre-training

CLIP learns global image representations via contrastive pre-training. However, for tasks that require fine-grained image understanding such as object detection, CLIP is not enough. Object detection contains two sub-tasks: localization and recognition. (i) Localization aims to locate the presence of objects in an image and indicate the position with a bounding box, while (ii) recognition determines what object categories are present in the bounding box. By following the reformulation that converts image classification to image retrieval used in CLIP, generic open-set object detection can be achieved.

Specifically, ViLD (Gu et al., 2021) and RegionCLIP (Zhong et al., 2022a) distill knowledge from CLIP with a two-stage detector for zero-shot object detection. In MDETR (Kamath et al., 2021) and GLIP (Li et al., 2022e) (as shown in Figure 2.14), the authors propose to reformulate detection as a phrase grounding problem, and perform grounded language-image pre-training. GLIPv2 (Zhang et al., 2022b) and FIBER (Dou et al., 2022a) further perform unified pre-training for both grounding and vision-language understanding tasks. OVR-CNN (Zareian et al., 2021) finetunes an image-text model to detection on a limited vocabulary and relies on image-text pre-training for generalization to an open vocabulary setting. Detic (Zhou et al., 2022b) improves long-tail detection performance with weak supervision by training only the classification head on the examples where only image-level annotations are available. Other works include OV-DETR (Zang et al., 2022), X-DETR (Cai et al., 2022), FindIT (Kuo et al., 2022), PromptDet (Feng et al., 2022a), OWL-ViT (Minderer et al., 2022), GReT (Wu et al., 2022b), to name a few. Recently, Grounding DINO (Liu et al., 2023h) is proposed to marry DINO (Zhang et al., 2022a) with grounded pre-training for open-set object detection. Please refer to Section 4.2 for a detailed review of this topic.

2.6.3 Pixel-Level Pre-training

The Segment Anything Model (SAM) (Kirillov et al., 2023) is a recent vision foundation model for image segmentation that aims to perform pixel-level pre-training. Since its birth, it has attracted wide attention and spurred tons of follow-up works and applications. Below, we briefly review SAM, as a representative work for pixel-level visual pre-training.

As depicted in Figure 2.15, the objective of the Segment Anything project is to develop a foundational vision model for segmentation. This model is designed to be readily adaptable to a wide range of both existing and novel segmentation tasks, such as edge detection, object proposal generation, instance segmentation, open-vocabulary segmentation, and more. This adaptability is seamlessly accomplished through a highly efficient and user-friendly approach, facilitated by the integration of three interconnected components. Specifically,

- **Task.** The authors propose the promptable segmentation task, where the goal is to return a valid segmentation mask given any segmentation prompt, such as a set of points, a rough box or mask, or free-form text.

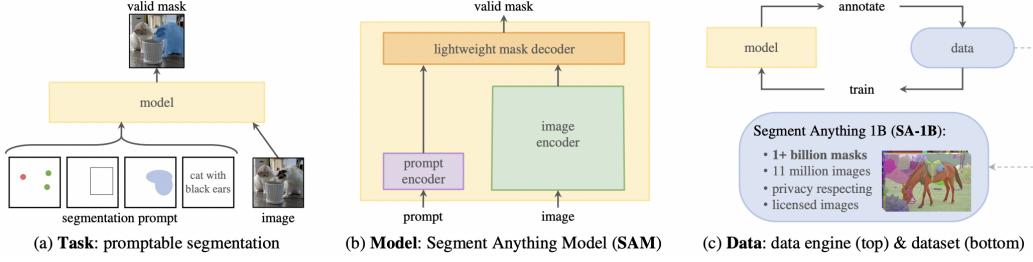


Figure 2.15: Overview of the Segment Anything project, which aims to build a vision foundation model for segmentation by introducing three interconnected components: a promptable segmentation task, a segmentation model, and a data engine. Image credit: Kirillov et al. (2023).

- **Model.** The architecture of SAM is conceptually simple. It is composed of three main components: (i) a powerful image encoder (MAE (He et al., 2022a) pre-trained ViT); (ii) a prompt encoder (for sparse input such as points, boxes, and free-form text, the CLIP text encoder is used; for dense input such as masks, a convolution operator is used); and (iii) a lightweight mask decoder based on transformer.
- **Data.** To acquire large-scale data for pre-training, the authors develop a *data engine* that performs model-in-the-loop dataset annotation.

Concurrent to SAM. Parallel to SAM, many efforts have been made to develop general-purpose segmentation models as well. For example, OneFormer (Jain et al., 2023) develops a universal image segmentation framework; SegGPT (Wang et al., 2023j) proposes a generalist in-context learning framework that unifies different segmentation data formats; SEEM (Zou et al., 2023b) further expands the types of supported prompts that a single segmentation model can handle, including points, boxes, scribbles, masks, texts, and referred regions of another image.

Extensions of SAM. SAM has spurred tons of follow-up works that extend SAM to a wide range of applications, *e.g.*, Inpaint Anything (Yu et al., 2023c), Edit Everything (Xie et al., 2023a), Any-to-Any Style Transfer (Liu et al., 2023g), Caption Anything (Wang et al., 2023g), Track Anything (Yang et al., 2023b), Recognize Anything (Zhang et al., 2023n; Li et al., 2023f), Count Anything (Ma et al., 2023), 3D reconstruction (Shen et al., 2023a), medical image analysis (Ma and Wang, 2023; Zhou et al., 2023d; Shi et al., 2023b; Zhang and Jiao, 2023), *etc.* Additionally, recent works have attempted to develop models for detecting and segmenting anything in the open-vocabulary scenarios, such as Grounding DINO (Liu et al., 2023h) and Grounding-SAM¹. For a comprehensive review, please refer to Zhang et al. (2023a) and some GitHub repos.²

¹<https://github.com/IDEA-Research/Grounded-Segment-Anything>

²<https://github.com/Hedlen/awesome-segment-anything>

Chapter 3

Visual Generation



Visual generation aims to generate high-fidelity visual content, including images, videos, neural radiance fields, 3D point clouds, etc.. This topic is at the core of recently popular artificial intelligence generated content (AIGC), and this ability is crucial in supporting creative applications such as design, arts, and multimodal content creation. It is also instrumental in synthesizing training data to help understand models, leading to the closed loop of multimodal content understanding and generation. To make use of visual generation, it is critical to produce visual data that is strictly aligned with human intents. These intentions are fed into the generation model as input conditions, such as class labels, texts, bounding boxes, layout masks, among others. Given the flexibility offered by open-ended text descriptions, text conditions (including text-to-image/video/3D) have emerged as a pivotal theme in conditional visual generation.

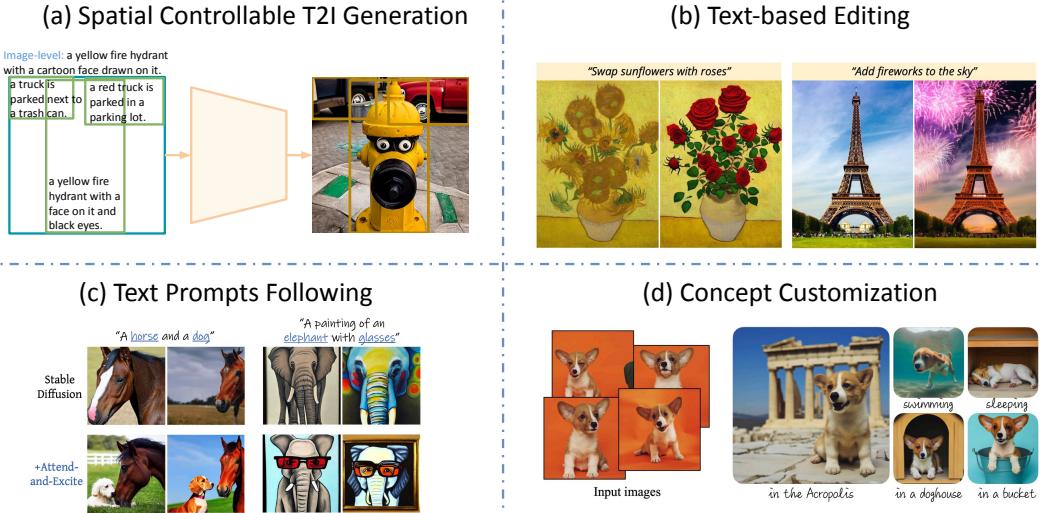
In this chapter, we describe how to align with human intents in visual generation, with a focus on image generation. We start with the overview of the current state of text-to-image (T2I) generation in Section 3.1, highlighting its limitations concerning alignment with human intents. The core of this chapter is dedicated to reviewing the literature on four targeted areas that aim at enhancing alignments in T2I generation, *i.e.*, spatial controllable T2I generation in Section 3.2, text-based image editing in Section 3.3, better following text prompts in Section 3.4, and concept customization in T2I generation in Section 3.5. At the end of each subsection, we share our observations on the current research trends and short-term future research directions. These discussions coalesce in Section 3.6, where we conclude the chapter by considering future trends. Specifically, we envision the development of a generalist T2I generation model, which can better follow human intents, to unify and replace the four separate categories of alignment works.

3.1 Overview

3.1.1 Human Alignments in Visual Generation

AI Alignment research in the context of T2I generation is the field of study dedicated to developing image generation models that can easily follow human intents to synthesize the desired generated visual content. Current literature typically focuses on one particular weakness of vanilla T2I models that prevents them from accurately producing images that align with human intents. This chapter delves into four commonly studied issues, as summarized in Figure 3.1 (a) and follows.

- **Spatial controllable T2I generation.** Text serves as a powerful medium for human-computer interaction, making it a focal point in conditional visual generation. However, text alone falls short in providing precise spatial references, such as specifying open-ended descriptions for arbitrary image regions with precise spatial configurations. Spatial controllable T2I generation (Yang et al., 2023b; Li et al., 2023n; Zhang and Agrawala, 2023) aims to combine text inputs with other conditions for better controllability, thereby facilitating users to generate the desired images.
- **Text-based image editing.** Editing is another important means for acquiring human-intended visual content. Users might possess near-perfect images, whether generated by a model or naturally captured by a camera, but these might require specific adjustments to meet their intent. Editing



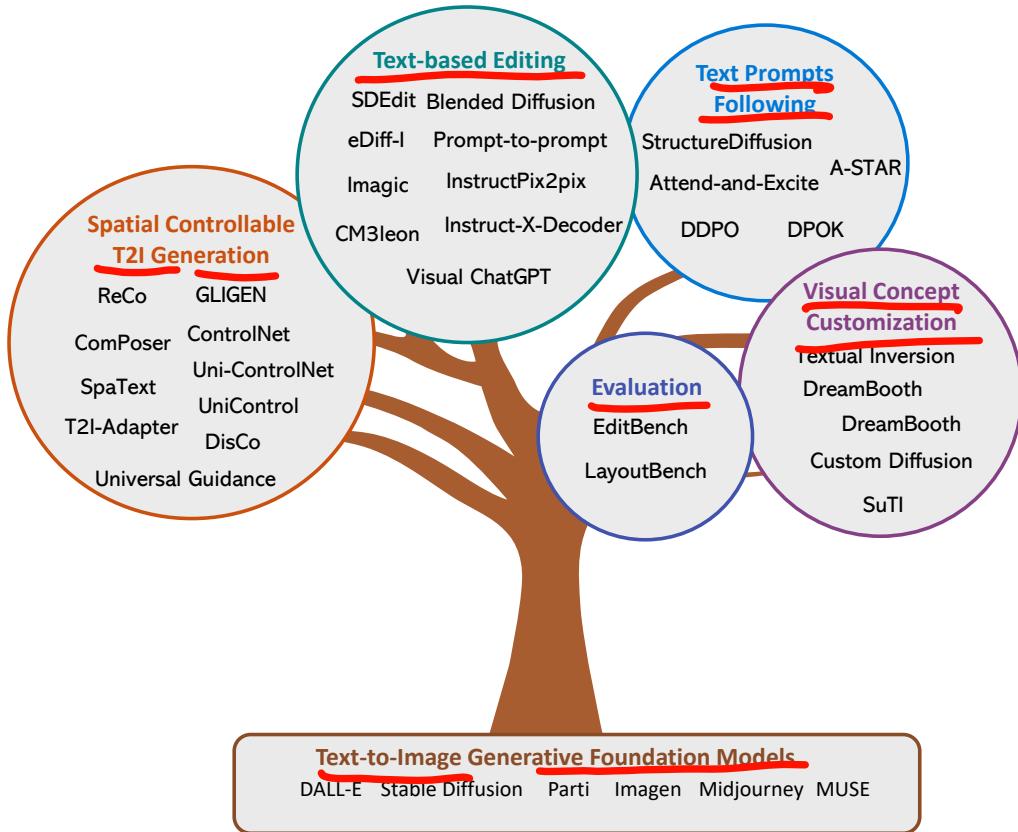
(c) Text Prompts Following



(d) Concept Customization



(a) An overview of topics on human alignment for generative foundation models. Image credit: Yang et al. (2023b); Brooks et al. (2023); Chefer et al. (2023); Ruiz et al. (2023).



(b) Summary and categorization of papers on “Human Alignments in Visual Generation.”

Figure 3.1: An overview of improving human intent alignments in T2I generation.

has diverse objectives, ranging from locally modifying an object to globally adjusting the image style. Text-based image editing (Brooks et al., 2023) explores effective ways to create a versatile editing tool.

- **Better following text prompts.** Despite T2I models being trained to reconstruct images conditioned on the paired text input, the training objective does not necessarily ensure or directly

optimize for a strict adherence to text prompts during image generation. Studies (Yu et al., 2022b; Rombach et al., 2022) have shown that vanilla T2I models might overlook certain text descriptions and generate images that do not fully correspond to the input text. Research (Feng et al., 2022b; Black et al., 2023) along this line explores improvements to have T2I models better following text prompts, thereby facilitating the easier use of T2I models.

- **Visual concept customization.** Incorporating visual concepts into textual inputs is crucial for various applications, such as generating images of one’s pet dog or family members in diverse settings, or crafting visual narratives featuring a specific character. These visual elements often encompass intricate details that are difficult to articulate in words. Alternatively, studies (Ruiz et al., 2023; Chen et al., 2023f) explore if T2I models can be customized to draw those visual concepts with specialized token embeddings or conditioned images.

Before introducing the alignment works in detail, we first review the basics of text-to-image generation in the next section.

3.1.2 Text-to-Image Generation

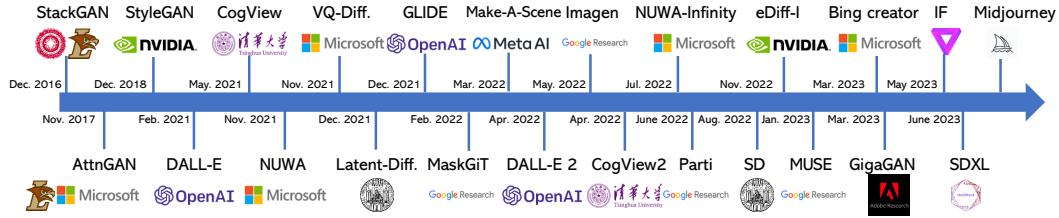


Figure 3.2: An overview of representative text-to-image generation models until July 2023.

T2I generation aims to generate images that are not only of high visual quality but also semantically correspond to the input text. T2I models are usually trained with image-text pairs, where text is taken as input conditions, with the paired image being the targeted output. Abstracted from the wide range of T2I models shown in Figure 3.2, we give a high-level overview of the representative image generation techniques.

- **Generative adversarial networks (GAN).** GANs (Goodfellow et al., 2020; Creswell et al., 2018; Kang et al., 2023) consist of two key components: a generator and a discriminator. The generator is tasked with creating synthetic images from random noise inputs, and it is trained to adjust these noise inputs based on input text conditions to generate semantically relevant images. In this adversarial process, the discriminator competes with the generator, attempting to differentiate between the synthetically generated images and real ones, thus guiding the generator to improve its image creation capabilities.
- **Variational autoencoder (VAE)** Variational Autoencoder (VAE) (Kingma and Welling, 2013; van den Oord et al., 2017; Vahdat and Kautz, 2020) is a probabilistic model that can generate images by employing paired encoder and decoder network modules. The encoder network optimizes the encoding of an image into a latent representation, while the decoder refines the process of converting the sampled latent representations back into a new image. VAEs are trained by minimizing the reconstruction error between the original and decoded images, while regularizing the encoded latent space using the Kullback-Leibler (KL) divergence. Vector Quantised-VAE (VQ-VAE) (van den Oord et al., 2017) further improves VAEs by leveraging the discrete latent space through vector quantization, enabling improved reconstruction quality and generative capabilities.
- **Discrete image token prediction.** At the core of this approach lies a combination of a paired image tokenizer and detokenizer, like Vector Quantized Generative Adversarial Networks (VQ-GAN) (esser et al., 2021), which efficiently transform continuous visual signals into a finite set of discrete tokens. In this way, the image generation problem is converted to a discrete token prediction task. A widely employed strategy for token prediction is to use an auto-regressive Transformer (Ramesh et al., 2021b; Yu et al., 2022b) to sequentially generates visual tokens, typically starting from the top left corner and moving row-by-row towards the bottom right, conditioned on the text inputs. Alternatively, studies (Chang et al., 2022, 2023) also explore the parallel decoding to speed up the token prediction process. Finally, the predicted visual tokens are detokenized, culminating in the final image prediction.

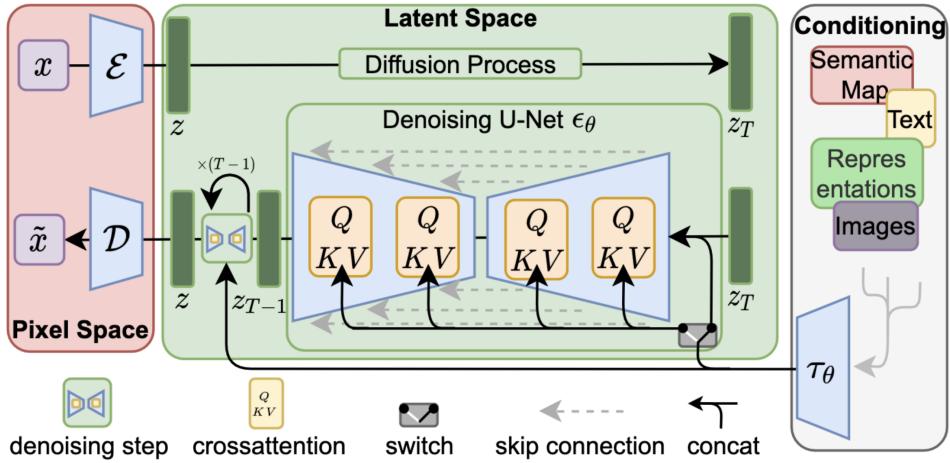


Figure 3.3: An overview of the latent diffusion model architecture. Image credit: Rombach et al. (2022).

- **Diffusion model.** Diffusion models (Sohl-Dickstein et al., 2015; Song and Ermon, 2020; Ho et al., 2020) employ stochastic differential equations to evolve random noises into images. A diffusion model works by initiating the process with a completely random image, and then gradually refining it over multiple iterations in a denoising process. Each iteration predicts and subsequently removes an element of noise, leading to a continuous evolution of the image, conditioned on the input texts.

We use Stable Diffusion (SD) (Rombach et al., 2022) as an example to explain in detail how T2I models work. We choose this model for a variety of reasons. Firstly, SD is one of the most widely used open-source T2I models, which makes it a solid foundation for many alignment techniques we discuss in this chapter. Additionally, as a diffusion-based generation model, it serves as an excellent case study for introducing diffusion models. Finally, its cross-attention-based image-text fusion mechanism is a classic example of various text-conditioned methods, such as auto-regressive T2I generation (Yu et al., 2022b), helping us gain an in-depth understand of the image-text interaction in T2I generation.

Stable Diffusion (SD)¹, and its academic version latent diffusion (Rombach et al., 2022), contains mainly three modules, *i.e.*, an image VAE, a denoising U-Net, and a condition encoder, as shown in the left, center, and right part of Figure 3.3, respectively. We will introduce each module and the inference flow for image generation, following the notations in the original latent diffusion paper (Rombach et al., 2022).

- **VAE.** As introduced in the image generation technique overview, the VAE module contains a paired encoder \mathcal{E} and decoder \mathcal{D} , trained to encode RGB image x into a latent random variable z and then decode the latent to best reconstruct the image. Given an RGB image $x \in \mathbb{R}^{H \times W \times 3}$, the encoder \mathcal{E} encodes it into a continuous latent representation $z \in \mathbb{R}^{h \times w \times c}$. With the parameters of $H = W = 512$, $h = w = 64$, and $c = 4$ in SD, latent z is 48 times smaller than image x , thereby significantly improving the computational efficiency by performing the denoising process in this compressed compact latent space.
- **Text encoder.** SD is a conditional image generation model, where the input text condition is encoded using a condition encoder τ . Specifically, SD uses the ViT-L/14 CLIP text encoder (Radford et al., 2021) that encodes the tokenized input text query y into text feature $\tau(y) \in \mathbb{R}^{N \times d_\tau}$, where the maximum length N is 77 and text feature dimension d_τ is 768.
- **Denoising U-Net.** The denoising U-Net is the core module for the diffusion image generation process. The module is trained to predict the noise $\hat{\epsilon}(z_t, t)$ to subtract in the latent space at each denoising timestep t , such that it can step-by-step evolve the initial random noise into a meaningful image latent. The module is trained with the L2 loss between the predicted noise $\hat{\epsilon}(z_t, t)$ and the

¹We use Stable Diffusion v1 for the introduction. Later versions such as SD2 and SDXL share the same method but may have different detailed model configurations, such as a larger text encoder, U-Net, and latent dimension.

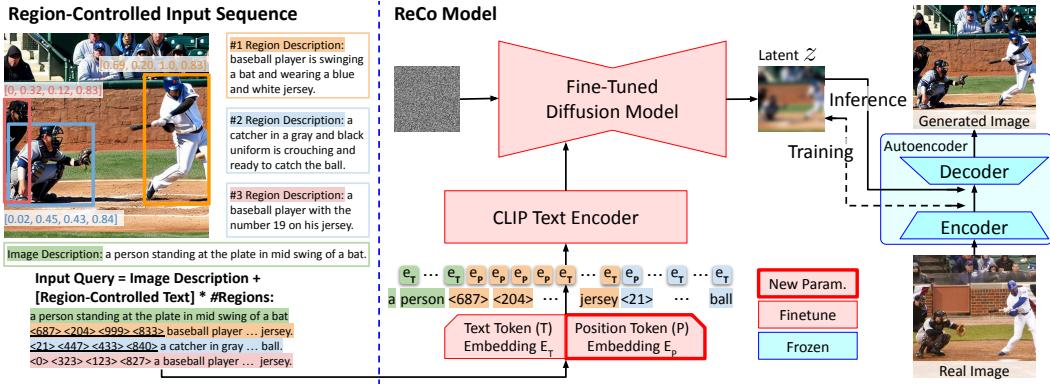


Figure 3.4: Overview of the ReCo model architecture. Image credit: Yang et al. (2023b).

target noise ϵ , which is added to the target image latent encoded by VAE encoder \mathcal{E} . At inference, the iteratively denoised latent z , started from the random noise, is sent through the VAE decoder \mathcal{D} for the final generated image.

In each denoising step, the U-Net takes the text condition as input to generate images that are semantically relevant to the text query. We next detail how the visual stream $z \in \mathbb{R}^{h \times w \times c}$ interacts with the text stream $\tau(y) \in \mathbb{R}^{N \times d_\tau}$. The denoising U-Net, similar to a classic U-Net (Ronneberger et al., 2015; Long et al., 2015), consists of a series of spatial downsampling and upsampling blocks with skip connections in between. In SD’s U-Net, each down/upsampling block has a cross-attention layer and a 2D convolutional down/upsampling layer. Each block takes the visual latent feature, text feature, and denoising step as input and generates the next visual latent feature. The image-text interaction happens in the image-text cross-attention layer.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V, \quad (3.1)$$

where K, V are projected from the text stream $\tau(y)$ and Q is projected from the visual stream z to share the same hidden dimension d . Therefore, the softmax between Q and K produces an attention map M of size $(hw \times d) \cdot (N \times d)^T = hw \times N$. The cross-attention map M indicates the fine-grained image-text interaction among each one of the N text words in all spatial positions hw . The attention map M then products V to yield the output of a down/upsampling block.

3.2 Spatial Controllable Generation

T2I generation takes open-ended text for users to describe their intended images. However, text alone is ineffective in certain descriptions, such as spatial referring. Studies in spatial controllable T2I generation explore extending T2I models to take extra spatial input conditions to guide image generation process.

We categorize related studies into three topics. (i) We start with works (Yang et al., 2023b; Li et al., 2023n; Avrahami et al., 2023b; Cho et al., 2023) that extend the image-level text description in vanilla T2I models to the region-grounded text description, such that open-ended text descriptions can precisely operate on a particularly spatial region. (ii) We then extend from boxes to dense spatial conditions represented as 2D arrays, such as segmentation masks, edge maps, depth maps, key points. We review representative works ControlNet (Zhang and Agrawala, 2023) and many others (Mou et al., 2023; Zeng et al., 2023; Zhao et al., 2023b; Qin et al., 2023a). (iii) The previous two threads of work require finetuning T2I models to understand the extended spatial condition. We next review techniques of inference-time guidance (Bansal et al., 2023; Chen et al., 2023e) that achieve spatial control without model finetuning.

Region-controlled T2I generation. Large-scale T2I models have demonstrated remarkable efficacy in generating high-resolution images. However, the models lack spatial controllability, e.g., precisely specifying content in a specified area using a free-form text description. This limitation motivates the studies on region-controlled T2I generation. As shown in the left side of Figure 3.4, these studies explore the extra input condition of open-ended text descriptions on arbitrary regions (*i.e.*, region-

controlled text), augmenting the global image description in T2I models. This new input condition requires T2I models to understand spatial inputs, and associate them with grounded texts.

ReCo ([Yang et al., 2023b](#)) is among the most representative works along this direction. The core idea is to extend the text vocabulary of the text encoder \mathcal{E} and arrange different tokens to represent the grounded text inputs. The study augments text tokens understood using pre-trained T2I models with an extra set of position tokens, which represent the quantized spatial coordinates. As shown in Figure 3.4, the position tokens (e.g., <687>, <204>, <999>, <833>) are seamlessly mixed with the text tokens and operate as a spatial modifier, indicating that the text to follow only operates on the specified spatial region, such as the “baseball player … jersey.” The pre-trained T2I model is then finetuned to support such a new input interface, thereby facilitating region-controlled T2I generation.

Shared by other approaches along this direction, ReCo discusses several advantages of region-controlled T2I generation in improving the alignment with human intents. (i) The grounded texts provide an extra input condition that allows users to specify the desired image easily, i.e., having a free-form regional description precisely at a specific location. The box token and the input sequence design allow users to generate grounded text with the same user interface as query a T2I model with text, making the extension easy to use. (ii) The additional region-level controlled texts help better generate images with correct object count, spatial relationship, and region attributes such as color/size, which may otherwise confuse the vanilla T2I model ([Rombach et al., 2022](#)). (iii) Studies also observe a better image generation quality, with the conjecture that the grounded text provides object-level image-text association and therefore simplifies the learning process.

GLIGEN ([Li et al., 2023n](#)) is another representative work. Alternate to generating grounded descriptions through the expansion of input tokens and finetuning the entire T2I model, GLIGEN uses a plug-and-play recipe: freezing the original T2I model and training extra gated self-attention layers to learn the new grounding skills. The grounding tokens carry two types of information: the semantic representation of text words that need to be grounded in and their spatial configurations. These grounding tokens are then added to the pre-trained T2I model via a newly added gated self-attention layer, with all remaining pre-trained parameters frozen. This layer is equipped with a gating parameter, which is initialized to zero, allowing the pre-trained model to incrementally incorporate the grounded text inputs. GLIGEN facilitates various types of grounded controls, including bounding box grounding, keypoint grounding, image prompting, as well as other types of spatially-aligned dense conditions.

T2I generation with dense conditions. In addition to spatial coordinates, there exist other spatial conditions often represented as 2D arrays, such as segmentation masks, edge maps and depth maps. ControlNet ([Zhang and Agrawala, 2023](#)) is a prominent example of incorporating these dense spatial controls into T2I models. ControlNet is built upon Stable Diffusion, and introduces an additional trainable ControlNet branch that adds an extra input condition to the text prompt. This extra condition can be a canny edge map, hough line, HED boundary, under sketching, human pose maps, segmentation masks, depth images, normal maps, or line drawing, each enabled with its distinct model copy. The added branch is initialized from the pre-trained downsampling blocks in the SD’s U-Net. This branch takes the added visual latent and the extra dense condition as input. Before combining input dense conditions with visual latent in the input and merging the ControlNet branch’s output back to SD’s upsampling blocks, there is a unique zero-initialized 1×1 convolutional layer. This layer serves as a gated connector to gradually inject the extra condition into the pre-trained Stable Diffusion model. With the extra dense spatial control, ControlNet provides an effective channel of generation controllability.

Follow-up studies such as Uni-ControlNet ([Zhao et al., 2023b](#)) and UniControl ([Qin et al., 2023a](#)) further improve ControlNet by unifying the input condition, such that a single model can understand multiple input condition types or even take a combination of two conditions. Examples of the dense controls and the corresponding generated images are shown in Figure 3.5. Moreover, Disco ([Wang et al., 2023f](#)) exemplifies the efficiency of ControlNet in the generation of human dancing videos, which aims to generate videos with controllable elements such as human subjects, video backgrounds, and motion pose sequences. The study successfully separates the background and human pose conditions, which are fed into two distinct branches of ControlNet, which condition on image frames and pose maps, respectively. This disentanglement of control from all three conditions allows Disco to accomplish high fidelity in both the human foregrounds and backgrounds.

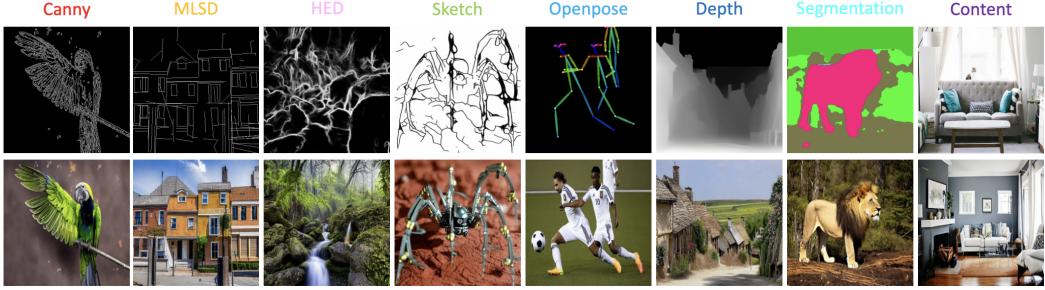


Figure 3.5: Examples of the dense controls and the corresponding generated images. Image credit: Zhao et al. (2023b).



Figure 3.6: Qualitative results of inference-time spatial guidance. Image credit: Bansal et al. (2023).

More importantly, it enables the arbitrary compositionality of human subjects, backgrounds, and dance movements.

Inference-time spatial guidance. The aforementioned works require model training, either the T2I models or additional modules to understand the extra spatial conditions. Alternatively, studies (Bansal et al., 2023; Chen et al., 2023e) explore providing the inference-time spatial guidance to T2I models without extra model training. The core idea is similar to classifier guidance (Dhariwal and Nichol, 2021), which takes a discriminator loss to guide the diffusion process as follows:

$$\hat{\epsilon}'(z_t, t) = \hat{\epsilon}(z_t, t) + s(t) \cdot \nabla_{z_t} \ell(c, f(\hat{z}_0)). \quad (3.2)$$

Taking spatial control as an example, the discriminator can be a Faster-RCNN object detector (Ren et al., 2015) indicated by f , which operates on the intermediate estimated image \hat{z}_0 , and compute the object detection loss ℓ with the desired layout c , to guide the generation $\hat{\epsilon}(z_t, t)$. $s(t)$ is the guidance strength. This approach enables the spatial control in T2I generation without extra training, with qualitative results shown in Figure 3.6. However, it may not yield results as precise as those from finetuning methods (Yang et al., 2023b; Li et al., 2023n; Zhang and Agrawala, 2023).

Summary and trends. Early research on spatial controllable generation, such as layout-to-image and mask-to-image generation, was often treated in parallel with T2I generation. However, with the emergence of advanced large-scale T2I models, recent studies, as discussed in this subsection, are now leaning towards integrating spatial conditions with textual conditions. We identify two primary trends in integrating spatial conditions into T2I models. First, region-controllable T2I generation, such as ReCo, merges spatial coordinate control into text inputs by enlarging the text vocabulary with position tokens. Second, studies extended from ControlNet integrate an additional “image-like” condition to T2I frameworks, thereby capturing a broad spectrum of dense conditions. Moving forward, T2I models may have a finetuning stage that allows them to comprehend both image and text inputs. In such a scenario, box coordinates could be incorporated through text, while dense controls could be provided as image inputs. We will explore and elaborate on this idea in Section 3.5.

3.3 Text-based Editing

Text-to-image editing synthesizes new images from an given image and input text descriptions. Instead of producing an image entirely from scratch, users might already possess a satisfactory starting point; this could be an image previously generated from a T2I model or a natural image. The objective is to retain the majority of the visual content, only modifying specific components. This could involve altering a local object or the overall image style to precisely match the user’s intentions. This text-based editing approach offers users a tool to generate fresh images based on a predecessor, playing a crucial role in creating visual content that accurately follows human intent.

There are various definitions and task setups in text-based editing. We introduce the following representative threads. (i) One classic editing scenario is to change a local image region, such as remov-



Figure 3.7: Three types of editing (word swap, adding new phrases, attention re-weighting) on synthetically generated images, enabled by attention map manipulation. Image credit: Hertz et al. (2022).

ing or changing an object or adding an object in a certain region. Spatially manipulating the latent in image generation according to the user-generated masks is a simple but effective method (Avrahami et al., 2022b,a; Meng et al., 2021). Studies (Balaji et al., 2022; Hertz et al., 2022) also show that manipulating the image-text cross-attention mask is effective for spatial editing. (ii) Extended from spatial editing where the language inputs describe the desired appearance in the spatial region, language can also be used as editing instruction to tell the machine what to do (Kawar et al., 2023; Brooks et al., 2023), such as “change object A in the image to object B.” (iii) Instead of extending a single T2I model for editing, editing systems (Wu et al., 2023a) integrate different specialized modules such as segmentation models (Kirillov et al., 2023; Zou et al., 2023b) and large language models (Brown et al., 2020; OpenAI, 2023a).

Diffusion process manipulations. The multi-step denoising process in diffusion image generation naturally supports a certain extent of image editing. Stochastic Differential Editing (SDEdit) (Meng et al., 2021) shows that first adding noises to the input image to edit and then subsequently denoising the sample, could produce a meaningful edit. Blended Latent Diffusion (Avrahami et al., 2022a) shows that the diffusion process manipulation can achieve local object editing with a user-generated object mask m_{latent} . In each diffusion step, the latent z is a spatial blend of the foreground and background latent: $z = z_{fg} \odot m_{latent} + z_{bg} \odot (1 - m_{latent})$, where z_{fg} is the edited object generated from the text description and z_{bg} is the original background image with noises added.

However, there are certain limitations on blending spatial latents. Firstly, it may not always be feasible to require human-generated masks. Secondly, the generation process can sometimes result in artifacts at the edges. Instead of simply blending the latent in a spatial manner, researchers delve into image-text cross-attention maps to unearth clues for object editing. Specifically, Prompt2Prompt (Hertz et al., 2022) discovers that cross-attention layers control the interaction among visual regions and text words. Based on this observation, the study enables three types of editing for images generated by a diffusion T2I model, including word swap, adding new phrases, and attention re-weighting, each of which is enabled with corresponding manipulation on the image-text cross-attention map. Specifically, the Prompt2Prompt tracks both cross-attention maps generated by the original prompt (namely M_t) and the edited prompt (namely M_t^*), and merges the attention maps with pre-defined rules into the new attention maps \widehat{M}_t , which is used for latent computing. For example, while adding a new phrase, attention map \widehat{M}_t remains unaltered for words present in the original prompt. It only incorporates the modified attention maps M_t^* for words that did not exist in the original prompt. Qualitative results of the edits are shown in Figure 3.7.

Going beyond editing synthetically generated images, Imagic (Kawar et al., 2023) explores editing real natural images. The core idea is to represent the image to be edited as text embedding, and blend this embedding with the target text embedding describing the desired image. This blend ensures that the resulting image retains elements from the original while aligning with the aesthetics detailed in the target textual prompt. In practice, test-time finetuning is needed to generate high-quality images.

Text instruction editing. Instead of repeating the visual contents of the image to edit in the text prompts, it might be more efficient for users to directly specify editing instructions using language, such as “swap sunflowers with roses” in Figure 3.8. The desired text instruction editing model should work on both model-generated and natural images, and across different types of editing instructions.

InstructPix2Pix (Brooks et al., 2023) is designed to accept an image and a text editing instruction to produce an edited version of the input image. The goal is to train an image-to-image model that can understand such editing text instructions. To achieve this, T2I models can be adapted to accept the additional image input by incorporating more input channels into the SD’s convolutional



Figure 3.8: Examples of text instruction editing. Image credit: [Brooks et al. \(2023\)](#).



Figure 3.9: The editing data generation pipeline proposed in InstructPix2Pix. Image credit: [Brooks et al. \(2023\)](#).

layer. The main challenge is how to generate paired editing data. As shown in Figure 3.9, InstructPix2Pix ([Brooks et al., 2023](#)) proposes to use a LMM ([Brown et al., 2020](#)) to generate a pair of an editing instruction and an edited caption from the original input caption, e.g., “have her ride a dragon,” “photograph of a girl riding a dragon,” and “photograph of a girl riding a horse.” The study then uses Prompt2Prompt ([Hertz et al., 2022](#)) to convert the original and edited caption pair to a pair of images before and after editing, corresponding to the GPT-generated editing instruction. The study generates over 450K samples to train the editing model. This data generation method has also been adopted in subsequent research, such as CM3Leon ([Ge et al., 2023](#)) for training general-purpose image-text-to-image models.

Editing with external pre-trained models. Furthermore, recent studies show the efficacy of incorporating external language and vision models for editing, as opposed to relying solely on a single model. Advancements in generalist segmentation models, such as SAM ([Kirillov et al., 2023](#)) and SEEM ([Zou et al., 2023b](#)), have paved the way for using segmentation models to ground the region for text-based editing. Representative works include Instruct X-Decoder ([Zou et al., 2023a](#)), Grounded SAM inpainting ([Liu et al., 2023h](#)), Inpaint anything ([Yu et al., 2023c](#)), etc.. Another emerging trend is the allocation of various generation and editing tools through LMM. Studies such as VisualChatGPT ([Wu et al., 2023a](#)) can solve complicated visual editing that requires the collaboration of multiple generation and editing models in multiple steps.

Summary and trends. Text-based editing models have made significant progress in their capabilities, leading to improved editing quality, expanded domain coverage, and more flexible user interface. For example, early studies require user-generated masks for object editing, while recent models can work on synthetically generated images without mask inputs, or even directly understand general text editing instructions. As we look to the future, we anticipate an all-encompassing generative foundation model that is capable of processing both image and text inputs. Within this framework, editing instructions would be a specialized form of text input, seamlessly integrated with the image description in T2I generation.

3.4 Text Prompts Following

Training with image-text pairs encourages T2I models to generate images that semantically correspond to the input text condition. However, the image generation training objective does not directly enforce generated images to exactly follow text prompts. Studies ([Feng et al., 2022b; Chefer et al.,](#)

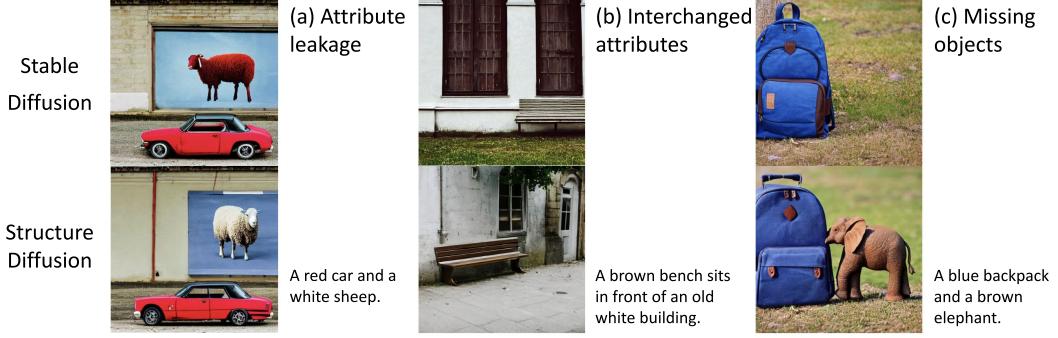


Figure 3.10: Failure cases of vanilla T2I model in text prompt following. Image credit: Feng et al. (2022b).

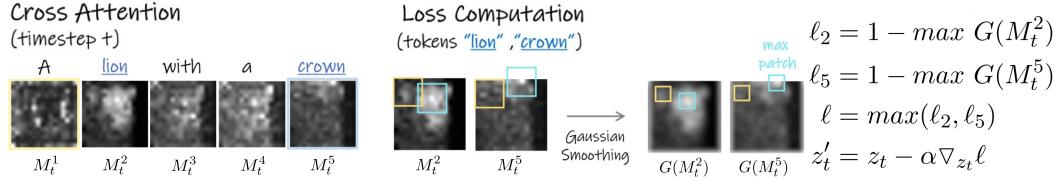


Figure 3.11: Inference time guidance proposed in Attend-and-Excite. Image credit: Chefer et al. (2023).

2023) show that T2I models may fail to follow text prompts, especially when the image description becomes complicated. For example, certain noun phrases may get omitted, attributes may apply to incorrect objects, and generated images may have the wrong object count, relationship, styles, etc.. These limitations motivate work on improving T2I models to better follow text prompts.

The related literature can be broadly categorized into two main groups. (i) *Inference-time manipulation*. In the inference stage, the latent and attention adjustment (Liu et al., 2022a; Feng et al., 2022b; Chefer et al., 2023; Agarwal et al., 2023) design various methods to redistribute the visual latent or image-text cross-attention, such that all noun phrases in the text prompts are represented in the generated image. (ii) *Alignment tuning*. An extra model learning stage is learned (Black et al., 2023; Fan et al., 2023b), typically with the image-text similarity as rewards, such that the tuned T2I model can better follow text prompts.

Inference-time manipulation. Training with image-text pairs does not guarantee that T2I models consistently adhere to the text prompts. There can be multiple discrepancies, particularly when the text descriptions are lengthy and intricate. For instance, T2I models may apply attributes to the wrong entity or miss certain objects, as shown in Figure 3.10. Intuitively, parsing the text query at inference time and explicitly enforcing T2I models to pay closer attention to each noun phrase may generate images that better follow text prompts.

Building upon this intuition, StructureDiffusion (Feng et al., 2022b) employs a parsing tree to extract noun phrases and the text prompt's linguistic structure. The study then enforces the model to "look at" all extracted noun phrases. This is implemented by modifying SD's cross-attention mechanism introduced in (3.1), written as $O = M \cdot V$ where M is the softmax cross-attention map. Instead of producing M with the sentence feature V , which may result in words getting overlooked, the study computes the $O = \frac{1}{k+1} \sum_{i=0}^k (M \cdot V_i)$, where V_0 is the sentence feature V , and $V_i, i = 1, \dots, k$ is the phrase feature in the parsing tree. This approach ensures that the visual stream maintains a balanced attention across all identified noun phrases, fostering more accurate image generation.

Motivated by the same objective to ensure that no object is overlooked, Attend-and-Excite (Chefer et al., 2023) manipulates the attention map. As shown in the right side equations in Figure 3.11, a regularization loss ℓ is computed to amplify the maximal attention towards the most neglected subject token:

$$\ell = \max_{n=1, \dots, N_{\text{sub}}} (1 - \max G(M_t^n))$$

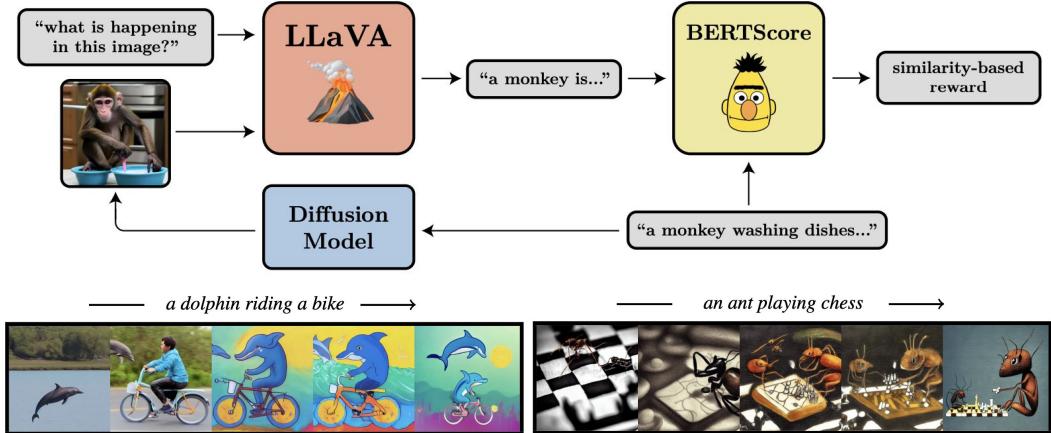


Figure 3.12: DDPO with vision-language-model-based reward function for image-text alignment tuning. Image credit: [Black et al. \(2023\)](#).

where G is a Gaussian kernel to smooth the attention map and N_{sub} is the number of subject tokens. The loss is then used to update the latent z_t at inference time:

$$z'_t = z_t - \alpha \nabla_{z_t} \ell,$$

where α is a scalar for the step size. Results show that this inference-time guidance enables T2I models to focus more on objects described in the text prompt, resulting in superior image generation. Follow-up studies ([Agarwal et al., 2023](#)) further refine the guidance loss to optimize prompt-following performance.

Model tuning to follow text prompt. Instead of inference-time manipulation, one may wonder if we can refine a pre-trained T2I model to better follow text prompts. One promising way to achieve this is via reinforcement learning, using image-text similarity as reward instead of the image generation objective used in the main T2I training. This allows the model to be optimized towards a better image-text alignment.

One work along this direction is the denoising diffusion policy optimization (DDPO) ([Black et al., 2023](#)), with the tuning framework shown in Figure 3.12. Specifically, a vision-language model ([Li et al., 2023e](#)) is used to convert the generated image into a text description. This generated caption is compared with the input text prompt, deriving a similarity reward through the use of BERTScore ([Zhang et al., 2019](#)). The similarity reward is then used to finetune the pre-trained T2I model, such that the model can better follow the text prompts. The bottom of Figure 3.12 shows the progression of the generated sample during this similarity-based training. Further, it is worth noting that other human intent may also be formatted as rewards for model tuning, such as compressibility, aesthetic quality, etc.

Summary and trends. In this section, we present studies aimed at enhancing the capability of T2I models to better adhere to text prompts. Despite the good performance achieved by the inference-time manipulation, the alignment tuning provides a more intuitive user experience, eliminating the need for extra modifications. In parallel to instruction tuning in LLMs to align human intent for text generation, the T2I model tuning shares a similar goal, but focuses on image generation. We foresee a similar paradigm emerging in the near future for generative T2I foundational model development. Specifically, the initial training phase still relies on the existing image generation objective on large-scale data, while the subsequent alignment-tuning phase enhances the model's prompt adherence and other aspects of human intent, such as diminishing harmful content. Current RL-based methods show potentials, but they typically zero in on a single optimization goal. Future research could extend these methods for more adaptable alignment tuning, amalgamated with features like accommodating diverse image and text scenarios.



Figure 3.13: The problem setup and result visualization of the visual concept customization task. Image credit: Ruiz et al. (2023).

3.5 Concept Customization

Though language is a powerful medium to express human intent, it is inefficient in comprehensively describing all details of a visual concept for reconstruction. For example, it is challenging to use texts to describe my pet dog or family members with sufficient details, so that they can be generated in different visual scenes. In such applications, directly extending T2I models to understand visual concepts via image inputs is a better option.

We examine relevant research on visual concept customization, which offers users the ability to generate these personalized concepts. (i) Pioneer studies (Gal et al., 2022; Ruiz et al., 2023; Wei et al., 2023) start with single-concept customization that involves test-time finetuning to encode multiple images of the visual concept into a new token embedding, such that the learned embedding can be used to refer to the concept during T2I generation. (ii) Multi-concept customization (Kumari et al., 2023; Avrahami et al., 2023a) allows multiple concept tokens to be expanded from the T2I model’s token vocabulary, enabling multiple concepts to interact with each other and the remaining visual scene during generation. (iii) Test-time finetuning requires users to tune T2I models for each new concept to customize. To simplify the usage, studies (Chen et al., 2022f; Shi et al., 2023a; Chen et al., 2023f; Yang et al., 2023a) explore customization without test-time finetuning and uses a unified finetuning stage to extend T2I models for accepting image condition inputs. The models take images of the visual concept as an extra input condition, and generate images with the visual concept following the text descriptions.

Single-concept customization. The goal of visual concept customization is to enable T2I models to comprehend additional visual concepts tailored to very specific cases. The problem setup, studied in Textual Inversion (Gal et al., 2022), involves translating visual concepts from a handful of images into unique token embeddings. As illustrated in the left side of Figure 3.13, the T2I model processes four images of a distinct dog breed, subsequently learning the embedding for a new token, denoted as [V]. This [V] token can be used as a text token to represent this specific dog. The [V] token can be seamlessly integrated with other textual descriptions to render the specific dog in various contexts, such as swimming, in a bucket, and getting a haircut.

Textual Inversion (Gal et al., 2022) learns the [V] token embedding via prefix tuning, *i.e.*, freezing all T2I model’s parameters and training the [V] token embedding to generate the input images. Later studies observe that tuning more model parameters leads to significantly better image generation quality. However, adjusting only the input image may lead to the risk of overfitting the T2I model for a particular concept, and losing the capability to generate diverse images. For instance, the model might become unable to generate various dog types. To address this, Dreambooth (Ruiz et al., 2023) proposes the class-specific prior preservation loss. Central to this approach is using the pre-trained T2I model to produce images of the same class as the targeted customization concept. The model is then jointly finetuned on both the input image (with the [V] token) and the model-generated images (without the [V] token). This ensures that the model can differentiate between the unique “[V] dog” and other general dogs it was initially trained, thus maintaining its overall T2I capability. Dreambooth then finetunes all T2I model parameters and achieves better image generation quality.

Multi-concept customization. Building on studies that focused on learning a single visual concept [V], recent research has delved into the possibility of integrating multiple visual concepts into a single Text-to-Image (T2I) model, represented as [V₁], [V₂], and so on. Custom Diffusion (Kumari et al., 2023) employs a selective subset of model weights, specifically the key and value mappings



Figure 3.14: Illustration of in-context concept customization without test-time finetuning. Image credit: [Chen et al. \(2023f\)](#).

from text to latent features in the cross-attention layers for concept customization, learned from multiple sets of concept images. The study facilitates the ability to embed multiple customized visual concepts in a single text prompt. Instead of learning from multiple sets of input images, Break-A-Scene ([Avrahami et al., 2023a](#)) explores extracting multiple visual concepts in a single image. The study augments input images with segmentation masks to pinpoint the intended target concepts and subsequently transforms them into a series of concept embeddings denoted as $[V_i]$.

Customization without test-time finetuning. While the concept customization studies, as described above, have achieved good visual quality, the necessity for test-time finetuning hinders its application in real-world settings. Most end users and application platforms lack the compute resources required for finetuning, not to mention the complexities of finetuning process. This naturally leads to the question: can we take concept images as input conditions, and achieve concept customization without finetuning?

The input/output format of the imagined system is similar to the retrieval-augmented generation ([Chen et al., 2022f](#)), which aims to ease the image generation by conditioning on a retrieved similar image. The system supports extra image inputs that contain relevant information for the generation process. By altering the conditioning images during the training phase, the model can potentially achieve a broad in-context learning capability, producing images that align with the given input examples. In line with this framework, SuTI ([Chen et al., 2023f](#)) trains a single model to imitate the finetuned subject-specific experts, and generates images conditioning on both text and subject input images, as shown in Figure 3.14. As a result, the trained model can perform in-context concept customization without test-time finetuning, and generalize to unseen subjects and descriptions. Another concurrent work, InstantBooth ([Shi et al., 2023a](#)), also shows remarkable results in generating images that are not only aligned with language but also preserve identities, with a single forward pass.

Summary and trends. The field of visual concept customization has advanced from finetuning embeddings during the testing stage, to directly performing in-context image generation with a frozen model. The in-context generation pipeline, which incorporates additional image inputs, shows remarkable potentials in real-world applications. In this subsection, we have explored two applications of this approach: facilitating generation through the retrieval of pertinent images ([Chen et al., 2022f](#)), and personalizing visual concepts by conditioning them on subject images ([Chen et al., 2023f; Shi et al., 2023a](#)). An intriguing direction is to unify the diverse uses of image inputs, directed by descriptive textual instructions. We elaborate on this idea in the following sub-section.

Related topics	Instruction text input	Content text input	Image input
T2I models (Sec. 3.1.2; SD)	None	Image description	None
Region-controlled T2I (Sec. 3.2; ReCo)	None	Image description + Box Tokens	None
T2I with dense conditions (Sec. 3.2; ControlNet)	None	Image description	Dense conditions (segmentation, edgemap, depth, keypoints, etc.)
Text instruction editing (Sec. 3.3; InstructPix2Pix)	Editing instruction ("change the dog's color to blue")	Contents for editing instruction	Image 
Concept customization (Sec. 3.5; SuTI)	Customization instruction ("generate a dog looks like this one")	Image description	Image 
Alignment tuned T2I models (Sec. 3.6)	Arbitrary instruction ("generate a dog looks like the left one but in blue")	Image description + Box Tokens ("in <687>, <204>, <999>, <833>")	Image or dense conditions 

Figure 3.15: Overview of the unified image and text input interface for human alignments, and its connection to previous sub-sections.

3.6 Trends: Unified Tuning for Human Alignments

In previous subsections, we presented the literature related to tuning T2I models to more accurately align with human intent. This includes enhancing spatial controllability, editing existing images for improved alignment, more effectively following text prompts, and personalizing T2I models for new visual concepts. A trend observed across these subtopics is the shift towards integrated alignment solutions that require minimal problem-specific adjustments. Along this direction, we envision a future T2I model having a unified alignment tuning stage, which transforms a pre-trained T2I model into one that resonates more intimately with human intent. Such a model would seamlessly process both text and image inputs, generating the intended visual content without the need for multiple models tailored to different alignment challenges.

Drawing parallels to the established practice of human-alignment tuning in LLM development, we anticipate that the techniques reviewed in this section will merge into a holistic second-stage tuning for generative foundation model development. This alignment tuning phase serves two primary purposes. First, it extends the T2I's text inputs to incorporate interleaved image-text inputs, as illustrated in Figure 3.15. Second, it finetunes the base T2I model, which has been trained using image generation loss, by the employing data, loss, and rewards that aim to align with human expectations.

Unified image and text inputs. We begin with the discussion on interface unification. Specifically, we aim to evolve the textual inputs of T2I models into a multimodal interface that seamlessly integrates both image and text inputs. As shown in Figure 3.15, we consider three types of inputs to begin with: "content text input" characterizes the visual scene to be produced; the "image input" accommodates dense 2D inputs such as images and dense conditions; and the "instruction text input" explains how the input content texts and images should be collectively composed as the condition for generation.

Vanilla T2I models, as shown in the first row of Figure 3.15, take the "content text input" of the image description and generate the corresponding image. For the spatial controllable generation in Section 3.2, the extra spatial condition can be specified via text inputs by expanding text words with extra box tokens, or via image input by feeding the dense spatial conditions as an image input. For the text-based editing in Section 3.3, we examine the efficacy of text instruction editing, a task that finetunes the T2I model to comprehend editing instruction texts that manipulate the image input, altering its pixel values accordingly. For visual concept customization in Section 3.5, the training-

free models can now understand customization instructions to extract visual concepts from the image inputs, and combine the concept with context text inputs for image generation.

Incorporating the three elements of the input interface, the envisioned alignment-tuned T2I model can handle all previous tasks described in Section 3.2-3.5. Its behavior is steered by specific text instructions that dictate how the image and text inputs should be jointly processed as the generation condition. Given the same image input, different text instructions can invoke different tasks: “generate a cat image with the same layout” for spatial control, “change the dog’s color” for editing, “generate the same dog sleeping” for concept customization, and the arbitrary mixture of the existing modes. Achieving such a unified interface in generative foundational models may be possible through training on a consolidated dataset encompassing data from various tasks, drawing similarities to the success of supervised instruction tuning observed in LLMs. Furthermore, transitioning from processing a single image-text pair to handling interleaved image-text pairs could enable more intriguing capabilities like in-context visual demonstrations (Sun et al., 2023b). Another interesting direction is to build a generative model that is capable of generating any combination of output modalities, such as language, image, video, or audio, from any combination of input modalities, as demonstrated in Composable Diffusion (CoDi) (Tang et al., 2023b).

Tuning with alignment-focused loss and rewards. In addition to the unified input interface, another noteworthy element deserving consideration is the alignment-focused loss and rewards. As mentioned in Section 3.4, the image generation loss based on image-text pairs enables models to produce images that match the target data distribution. Yet, it doesn’t always perfectly align with human intent. This is reminiscent of the language model loss in LLM training, which necessitates a separate alignment tuning phase (Ouyang et al., 2022). The recent success in supervised instruction tuning and reinforcement learning from human feedback methods (Black et al., 2023) on image generation provides effective tools for similar alignment tuning in generative foundation models. An intriguing topic left for future exploration is how to balance the different target losses and rewards, such as jointly optimizing for higher aesthetic scores, better image-text alignment, fewer harmful contents, stronger instruction adherence, along with many other desired properties.

Closed-loop of multimodal content understanding and generation. As we look ahead, one promising avenue of research is the closed-loop integration of multimodal content understanding and generation. Preliminary studies have shown the benefit of using synthesized data to benefit generation from understanding (Li et al., 2023a; He et al., 2022b), and vice versa. An exciting prospect would be the development of an image-text-input, image-text-output foundational model for both understanding and generation tasks. The ideal balance in combining these two dimensions, and the most efficient approach to achieve it, are left for future explorations.

Chapter 4

Unified Vision Models



In this chapter, we discuss the unification of vision models. We start with an overview of the challenges in the unification of vision models and the most recent efforts towards this goal in Section 4.1. What follows are detailed discussions on (*i*) how to transform closed-set models to open-set ones in Section 4.2; (*ii*) how to unify different granularities of vision tasks in Section 4.3; and (*iii*) how to build a more promptable interface for vision in Section 4.4. Finally, we summarize the chapter and discuss future trends in Section 4.5.

4.1 Overview

Before talking about general-purpose unified vision systems, we revisit how language models and natural language processing (NLP) have evolved in the past years. Before 2018, different NLP tasks are addressed with different task-specific models, such as translation (Bahdanau et al., 2015), semantic parsing (Berant et al., 2013), summarization (Allahyari et al., 2017), and so on. With the emergence of the transformer architecture (Vaswani et al., 2017), language models for different NLP tasks are unified with a decoder-only architecture, e.g., the GPT models (Brown et al., 2020). Afterwards, the GPT models learned using the next word prediction task are further finetuned to follow human instructions. This leads to ChatGPT¹, which fundamentally changes our expectations on what AI systems can do. The evolution as depicted in Figure 1.1 motivates us to wonder whether we can build a general-purpose vision system in a similar manner.

Challenges. That computer vision tasks vary greatly presents a great challenge to build a unified vision model. First, vision tasks have different types of inputs, ranging from static images (Russakovsky et al., 2015) to sequential videos (Miech et al., 2019), from pure vision inputs such as image dehazing (He et al., 2010) to multi-modality inputs that include e.g., vision and language Antol et al. (2015). Second, different granularities are required for different tasks, such as image-level tasks like image classification (He et al., 2016) and captioning (Vinyals et al., 2016), region-level tasks like object detection (Girshick, 2015) and grounding (Plummer et al., 2015), and pixel-level tasks like image segmentation (He et al., 2017), super-resolution (Wang et al., 2020), etc. As a result, the outputs of vision systems are also of different formats, such as spatial information like edges, boxes, and masks, semantic information like class labels, multi-label tags, or detailed descriptions. In addition to the challenges in modeling, there are also challenges with data. First, the cost of annotation varies greatly among different types of labels. As shown in Figure 4.6, these labels are at different levels of granularity and semantic richness, ranging from whole images, regions (box annotations), to masks (pixel annotations). Second, it is in general much more costly to collect image data than text data. So, the scale of vision data is often much smaller than that of text corpora.

Towards a unified vision model. Despite these challenges, there is a growing interest in the computer vision community to develop a general-purpose, unified vision system, in particular for visual understanding tasks. As illustrated in Figure 4.1, we group these efforts in three categories:

¹<https://chat.openai.com/>

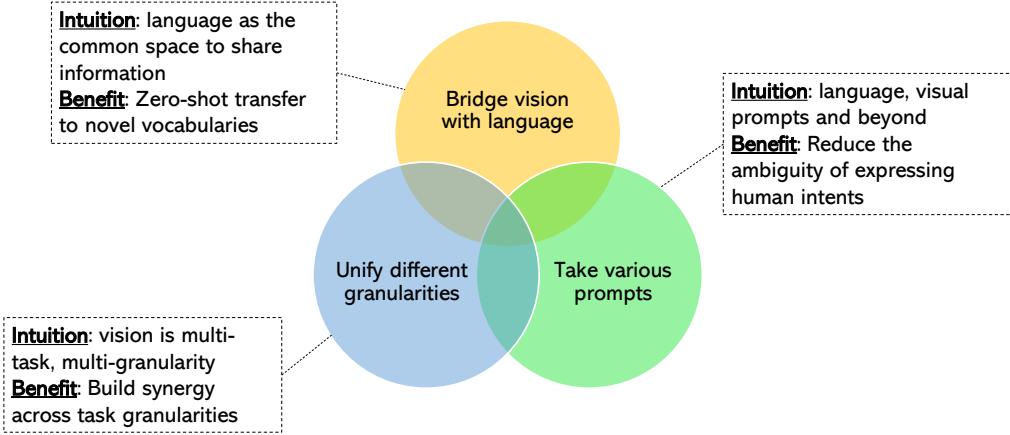


Figure 4.1: In NLP, we have witnessed a clear trend to build a unified model such as GPT-3 (Brown et al., 2020) and then the sophisticated Human-AI interaction system ChatGPT, which has ignited the interests of the whole community and society in AI. A natural question for computer vision (CV) is whether we can unify all different types of vision tasks such as image classification, object detection, segmentation and visual question answering, etc., and likewise build an interaction interface between CV models and humans. Aspired by this, a lot of attempts have been made recently to crack the problem from different angles including but not limited to (a) making vision models open-set; (b) unifying different granularities; and (c) turning the models more promptable.

- **Bridging vision and language.** By extending closed-set classification to open-world recognition, the contrastive language-image models like CLIP (Radford et al., 2021) demonstrate impressive zero-shot transferability for different vision tasks. These models learn the mapping between raw visual signals and rich semantics and can power various open-vocabulary vision recognition tasks (Zhong et al., 2022b; Gu et al., 2022; Li et al., 2022f; Ghiasi et al., 2022b).
- **Unified multi-task modeling.** Traditional task-specific vision models are trained using task-specific data. It is often prohibitively expensive to develop a model for a new task. Thus, it is desirable to develop a unified vision model that can perform well across many vision tasks (Yang et al., 2022c; Lu et al., 2022a; Zou et al., 2023a; Chen et al., 2022c).
- **LLM-like promptable interface.** LLMs can take different language and in-context prompts as inputs and produce user-desired outputs without finetuning. A general-purpose vision model should have possessed the same in-context learning capability to align the output to various user intents without changing its model parameters (Bar et al., 2022; Kirillov et al., 2023; Zou et al., 2023b; Wang et al., 2023j; Balažević et al., 2023).

In what follows, we will elaborate the detailed techniques and methods in each category.

4.2 From Closed-Set to Open-Set Models

Traditionally, visual recognition is formulated as a classification problem that maps raw visual data (e.g., images) to discrete text labels. For example, image classification predicts a label from a pre-defined close set for a whole image (Deng et al., 2009), and object detection identifies the objects, defined in a close set, within an image (Lin et al., 2014). However, such *closed-set* models can hardly transfer to other tasks where the close set (or vocabulary) is insufficient. For example, it is difficult to apply an object detector trained using the Microsoft COCO object set² to detect Minecraft objects. Recently, CLIP (Radford et al., 2021) addresses the limitation of closed-set models by introducing a contrastive language-image pre-training method to train an *open-set* model. As illustrated in Figure 4.2 (a), instead of learning the mapping from input to labels, CLIP learns an aligned visual-semantic space using hundreds of millions of image-text pairs. Mathematically,

²<https://cocodataset.org/>

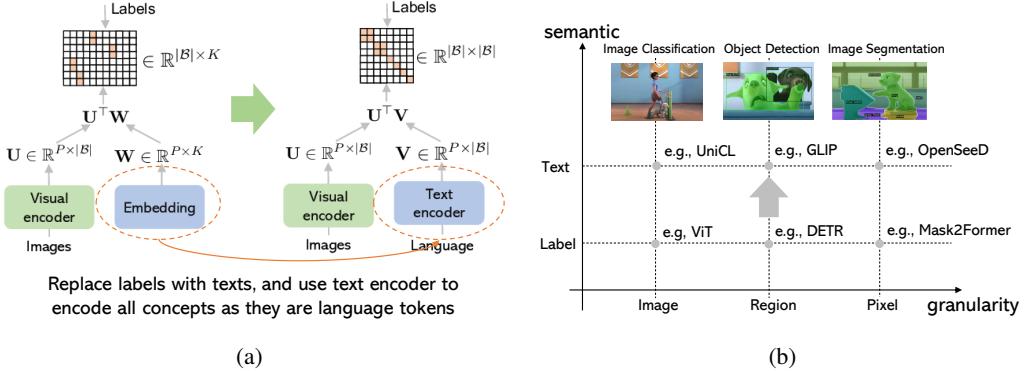


Figure 4.2: (a) As proposed in CLIP (Radford et al., 2021), replacing labels with textual descriptions and using a text encoder to encode them can feasibly convert closed-set problems to open-set ones. Image credit from Yang et al. (2022b). (b) A number of works have been proposed to transform different computer vision tasks by replacing the label space with language space, such as UniCL (Yang et al., 2022b), GLIP (Li et al., 2022f) and OpenSeeD (Zhang et al., 2023e).

the traditional vision tasks optimize the log-likelihood of assigning label $y = c$ to an image, often represented as a feature vector $u \in \mathcal{R}^P$:

$$\log \mathcal{P}(y = c|u) = \log \frac{\exp^{w_c \cdot u}}{\sum_{i=1}^K \exp^{w_i \cdot u}}, \quad (4.1)$$

where $w \in \mathcal{R}^{K \times P}$ is the projection matrix. Instead of using a pre-determined project matrix w , the CLIP method uses a text encoder Enc_{text} to for the projection:

$$v_i = Enc_{text}(Concept_i), \quad (4.2)$$

where v plays the role of w in Eq. (4.1). The reason why a text encoder can help achieve open-set recognition is that all textual concepts are embedded in the same feature space through large-scale pre-training, and the feature distributions are coherent to the semantic meanings without the need of a pre-defined vocabulary. As such, the aligned visual-semantic space can be easily transferred to a wide range of image recognition tasks in a zero-shot manner. Please refer to Chapter 2 for a detailed discussion. In the following, we focus our discussion on the region-level and pixel-level models.

After the release of the CLIP model (Radford et al., 2021), a number of open-set vision models have been developed using large amounts of text-image pairs for visual understanding at different levels of granularity (Yang et al., 2022b; Zhang et al., 2023e; Li et al., 2022f; Ghiasi et al., 2022a), ranging from image-level tasks (*e.g.*, image classification Deng et al. (2009), image-text retrieval, image captioning Chen et al. (2015)), region-level localization (*e.g.*, object detection and phrase grounding Plummer et al. (2015)), to pixel-level grouping tasks (*e.g.*, image segmentation and referring segmentation Long et al. (2015); Kirillov et al. (2019); Hafiz and Bhat (2020)). These models can be categorized along the following three dimensions: model initialization, design and training.

Model initialization. There are different initialization methods for open-set model training.

- **CLIP initialized.** Many recent open-set models are trained by using a pre-trained model such as CLIP for initialization since a pre-trained model already provides a well-aligned (but often coarse-grained) visual-semantic feature space. For example, OVR-CNN (Zareian et al., 2021) and RegionCLIP (Zhong et al., 2022b) use a CLIP-style pre-trained ResNet (He et al., 2016) as the vision encoder and a pre-trained RPN (Ren et al., 2015) to extract regional features. Likewise, MaskCLIP (Zhou et al., 2022a) and FreeSeg (Qin et al., 2023b) exploit the CLIP model to extract dense labels for pixels. FC-CLIP (Yu et al., 2023a) uses a frozen convolution network ConvNeXt (Liu et al., 2022b) in CLIP to encode input images of various resolutions.
 - **CLIP augmented.** Instead of initializing a model with CLIP parameters, other methods initialize the model parameters as usually (e.g., setting random values to model parameters), but use the pre-trained CLIP to help model training. For example, ViLD (Gu et al., 2022) augments the model

with aligned CLIP features via knowledge-distillation. MaskCLIP (Ding et al., 2022b) and Mask-Adapted CLIP Liang et al. (2023a) rely on the pre-trained CLIP model to provide features and scores, respectively, during the course of model training.

- Other works learn a visual-semantic feature space using supervised pre-trained models or from scratch. For example, GLIP (Li et al., 2022f) and OpenSeeD (Zhang et al., 2023e) use a pre-trained BERT model (Devlin et al., 2019) and the CLIP text encoder, respectively, and use a vision backbone pre-trained on ImageNet for image encoding. Though these separately pre-trained image and text encoders do not explicitly learn the alignment between image and language, it turns out that these models still give good representations for images and texts, and are instrumental to efficient model training. Differently, GroupViT (Xu et al., 2022a) is trained jointly using an open-set semantic segmentation task and a global image-text alignment task from scratch. ODISE (Xu et al., 2023a) exploits pre-trained Stable Diffusion models (Rombach et al., 2022) to extract compact masks.

Model design. Open-set models can be either multi-stage or end-to-end.

- **Two-stage models.** These models usually follow the design of the pre-DETR based models (Ren et al., 2015; He et al., 2017), which decouples localization and recognition. For object detection, a region proposal network is typically pre-trained for localizing the object of interest (Zhong et al., 2022b; Gu et al., 2021), and a mask proposal network for extracting masks (Ghiasi et al., 2022a; Yao et al., 2022a). Given the localization results, a pre-trained CLIP model is used to measure the similarity between visual contents and language concepts. A clear advantage for two-stage models is that they can inherit the open-set semantic understanding capacity without additional training so as to devote modeling training to requiring a well-performed localization network.
- **End-to-end models.** Different from two-stage models, the end-to-end models follow the DETR-based methods (Carion et al., 2020; Cheng et al., 2022) or other one-stage models (Dai et al., 2021). GLIP (Li et al., 2022f) is one of the representative works. GLIP formulates object detection as textual grounding and is trained end-to-end on image-text pairs with detection and grounding labels. Follow-up works enhance GLIP by enabling deeper vision-language interactions (Liu et al., 2023h) or using DETR-like model design (Zang et al., 2022; Minderer et al., 2022). For segmentation, both ZegFormer (Ding et al., 2022a) and OpenSeeD (Zhang et al., 2023e) exploit a DETR-like architecture and predict the masks and categories based on the outputs of their decoders.

Model pre-training. There are mainly three learning methods for pre-training open-set vision models.

- **Supervised learning.** By converting label supervision to language supervision, many works directly leverage the existing supervised annotations for training open-set models. For example, OVR-CNN (Zareian et al., 2021) trains a model with COCO categories and then evaluates its performance on novel categories. Likewise, ViLD (Gu et al., 2021) trains and evaluates two separate models on COCO and LVIS datasets, respectively. Following a similar protocol, many works train the open-set segmentation models on a subset of annotated segmentation data and evaluate the generalization ability on held-out data (Ding et al., 2022a,b; Zhang et al., 2023e; Xu et al., 2023a).
- **Semi-supervised learning.** One might use both annotated data and unlabeled or weakly-labeled data. For example, both RegionCLIP (Zhong et al., 2022b) and GLIP (Li et al., 2022f) use a teacher model to extract fine-grained region-text alignments from image-text pairs to augment the training data for better open-set detection performance. Differently, OpenSeg (Ghiasi et al., 2022b) exploits Localized Narrative datasets (Pont-Tuset et al., 2020) as weakly-labeled data, which provides coarse correspondence between language phrases and strokes in images. Empirically, such semi-supervised learning methods often help improve models’ generalization ability because they can effectively leverage rich semantics from noisy data.
- **Weakly-supervised learning.** Some works solely use weakly-labeled data for modeling. For example, GroupViT (Xu et al., 2022a) uses a contrastive learning method where all supervisions for model training are from positive and negative image-text pairs. Following the same contrastive learning method, SegCLIP (Luo et al., 2023b) uses a gathering mechanism to learn to merge image patches through the training on image-text pairs.

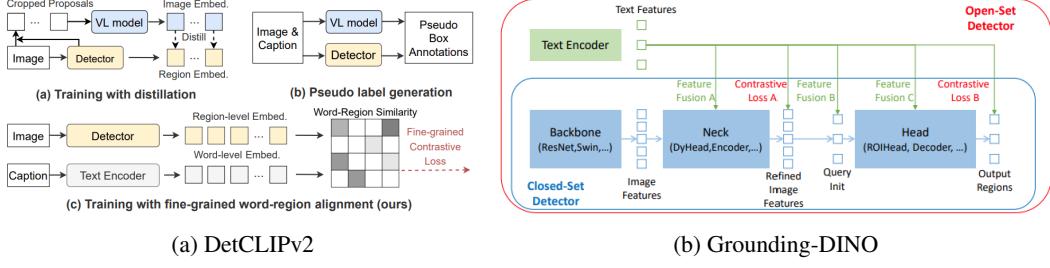


Figure 4.3: (a) DetCLIPv2 (Yao et al., 2023) learns fine-grained word-region alignment from object detection and grounding data and large-scale image-text pairs. (b) Grounding-DINO (Liu et al., 2023h) injects text conditions into different stages of the Transformer encoder-decoder, which significantly improves the grounding performance. Image credit: Yao et al. (2023) and Liu et al. (2023h).

Below, we review recent models developed for region-level and pixel-level tasks.

4.2.1 Object Detection and Grounding

Object detection is a fundamental task in computer vision that involves identifying and localizing objects of interest within an image or a video sequence (Viola and Jones, 2001). Over the years, various techniques and algorithms have been developed to improve the accuracy and efficiency of object detection. In the past, region-based approaches such as R-CNN Girshick et al. (2015), Fast R-CNN (Girshick, 2015) and Faster R-CNN (Ren et al., 2015) have been fostering the development of advanced techniques for object detection. To improve real-time performance, YOLO (Redmon et al., 2016) proposes a single neural network that simultaneously predicts object classes and bounding box coordinates. Some improvements are made by either using multiple feature maps at different scales (Liu et al., 2016) or introducing a focal loss to address the class imbalance problem in dense object detection scenarios (Lin et al., 2017). After the emergence of Transformer (Vaswani et al., 2017), DETR (Carion et al., 2020) applies the transformer architecture to object detection, treating it as a set prediction problem. Since DETR, a number of methods have been proposed to improve transformer-based detection models from various aspects, such as DINO (Zhang et al., 2022a), Group DETR (Chen et al., 2022b), and Co-DETR (Zong et al., 2023).

Open-set object detection models aim to detect arbitrary concepts beyond the vocabulary provided in training data. Three main evaluation settings have been developed in the literature:

- **Zero-shot object detection.** Similar to zero-shot image classification (Xian et al., 2018), zero-shot object detection restricts the object classes used for training, and evaluates models' transferability to novel classes. Methods falling in this category mainly focus on evaluating how a model leverages pre-trained concept embeddings (*e.g.*, word2vec (Mikolov et al., 2013)) and learns good visual-semantic alignments (Bansal et al., 2018; Rahman et al., 2020; Zhu et al., 2019, 2020).
- **Strict open-vocabulary object detection.** First introduced in OV-RCNN (Zareian et al., 2021), this setting differs from zero-shot object detection in that there is no limit on the training vocabulary as long as it does not cover any target classes. Under this protocol, some representative works are ViLD (Gu et al., 2021), RegionCLIP (Zhong et al., 2022a) which leverage large-scale language-image models (Radford et al., 2021; Jia et al., 2021), and Detic (Zhou et al., 2022b) that learns from image-label data.
- **Generalized open-vocabulary object detection.** Some recent works like GLIP (Li et al., 2022f), and OWL-VIT (Minderer et al., 2022) advocate a more flexible setting to evaluate the dataset or task transferability for object detection models. This setting allows vocabulary overlap between training and test sets, *e.g.*, Objects365 for training while COCO for evaluation. This is arguably a more practical setting than the two settings described above in that models can be trained using any arbitrary set of training data and their detection performance evaluated in the wild (Li et al., 2022b).

Object grounding can be considered as a generalized open-set object detection task (Plummer et al., 2015; Kazemzadeh et al., 2014; Chen et al., 2019; Deng et al., 2018). In this task, models take a sentence and an image as input and localize objects that are associated with the noun phrases.

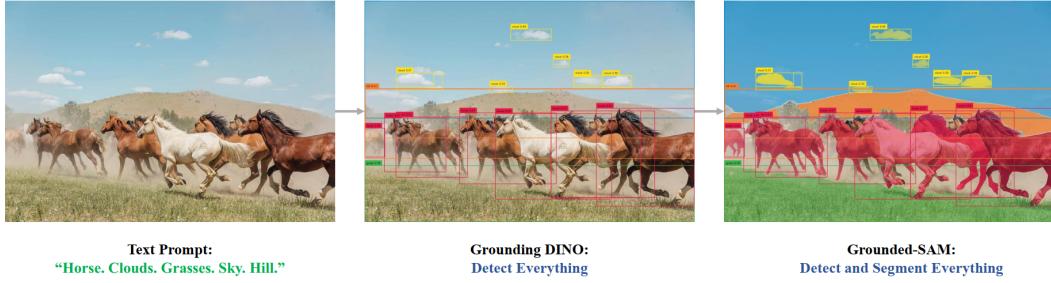


Figure 4.4: Grounding-SAM consisting of Grounding-DINO (Liu et al., 2023h) and SAM (Kirillov et al., 2023). Image credit: Liu et al. (2023h).

Recently, M-DETR (Kamath et al., 2021) employs a transformer-based architecture to build an end-to-end modulated detector to detect objects in an image given a raw text query. Unlike previous works where models are trained on specific datasets, the network is pre-trained with 1.3M pairs of text and images, sourced from multi-modal datasets where the connections between text phrases and corresponding image objects are labeled. Inspired by M-DETR, GLIP (Li et al., 2022f) casts object detection as a grounding problem, and jointly learns a model using object detection and grounding data for open-set scenarios. Following this line of research, DetCLIPV2 (Yao et al., 2023) proposes a simple joint learning method where multiple tasks are converted into a word-region alignment task, and then a model is trained end-to-end on a corpus consisting of object detection data, grounding data and image-text pairs. Grounding-DINO (Liu et al., 2023h) is a state-of-the-art grounded object detection method, where the object detector is composed of components: a backbone, a neck, and a head, and inject language conditions at every stage. A combined text and image backbone is employed to extract features at multiple scales, which are then passed on to the neck. The text and image characteristics generated by the neck are subsequently used for language-driven query selection. Grounding-SAM is developed by combining Grounding-DINO with SAM (Kirillov et al., 2023). As shown in Figure 4.4, an image and a group of concepts are first fed into Grounding-DINO to produce the boxes, and then the boxes are used as prompts for SAM to predict masks for each box.

4.2.2 Image Segmentation and Referring

Image segmentation is a long-standing and challenging vision problem. There are mainly three sub-tasks, including semantic (Long et al., 2015), instance (Hafiz and Bhat, 2020), and panoptic (Kirillov et al., 2019) segmentation. Semantic segmentation cares about the per-pixel semantic within an image (Long et al., 2015; Chen et al., 2017, 2022j), whereas instance segmentation groups pixels of the same semantic meaning into objects. Models for both tasks have evolved from CNN-based architectures (Long et al., 2015) to transformer-based ones (Chen et al., 2022j), and from two-stage models (He et al., 2017) and one-stage models (Bolya et al., 2019; Tian et al., 2020b) to the recent query-based approaches (Dong et al., 2021; Zou et al., 2022). With the capability of per-pixel and instance-level understanding, a natural step was taken to formulate panoptic segmentation (Kirillov et al., 2019; Wang et al., 2021a; Cheng et al., 2022). Most recently, Mask2Former (Cheng et al., 2022) proposed to address all three tasks with a unified encoder-decoder architecture. Nevertheless, all these works cope with a limited number of categories. In the following, we will review the most recent works on open-set image segmentation and referring segmentation.

Open-Vocabulary Segmentation. Recently, a number of methods have been proposed to transfer or distill the rich visual-semantic knowledge from foundation models (Radford et al., 2021; Jia et al., 2021) to specific segmentation tasks. Prominent examples include LSeg (Li et al., 2022a), OpenSeg (Ghiasi et al., 2022a), and Huynh et al. (2022). Instead of using existing models, GroupViT Xu et al. (2022a) performs language-image pre-training from scratch with a bottom-up grouping ViT (Dosovitskiy et al., 2021), while DenseCLIP (Rao et al., 2022) demonstrates the superiority of foundation models in finetuning settings compared with supervised models. Recently, MaskCLIP (Ding et al., 2022b) is proposed to tackle open-vocabulary panoptic and semantic segmentation simultaneously by leveraging CLIP, and achieves impressive performance on ADE20K (Zhou et al., 2017) and PASCAL (Mottaghi et al., 2014; Everingham and Winn, 2011).

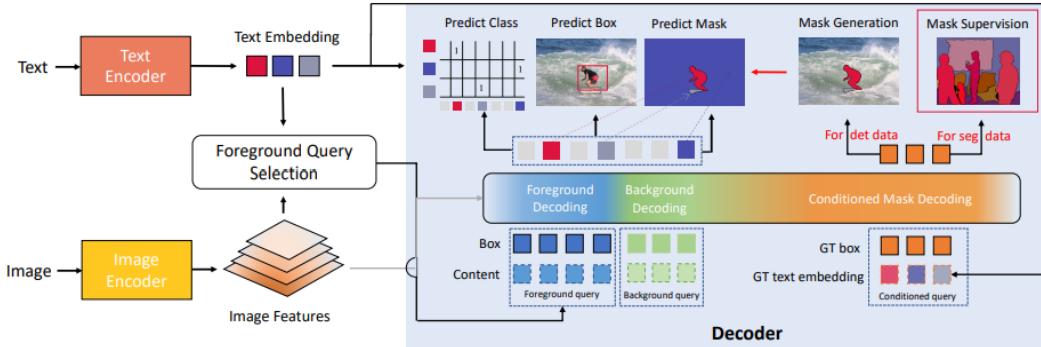


Figure 4.5: OpenSeeD (Zhang et al., 2023e) leverages both mask and box supervision for learning a universal open-vocabulary image segmentation model. Image credit: Zhang et al. (2023e).

Instead of using the ViT backbone, a recent work called FC-CLIP (Yu et al., 2023a) exploits a convolutional CLIP backbone (*i.e.*, ConvNeXt trained by OpenCLIP (Ilharco et al., 2021)) as both a feature extractor and a vision encoder. Based on a simplified pipeline, FC-CLIP shows plausible efficiency and lefts the state of the art on various open-vocabulary segmentation benchmarks. Rather than only using CLIP, a recent work ODISE (Xu et al., 2023a) leverages text-to-image diffusion models, and shows that the latent features in the pre-trained UNet can provide useful compact segmentation information for open-vocabulary segmentation.

A big challenge in open-vocabulary segmentation is the lack of segmentation data annotated with semantic labels. Thus far, most of the works are still using COCO segmentation annotations. A few recent works attempt to leverage object detection data as the extra supervision to augment the training of segmentation models, such as OpenSeeD (Zhang et al., 2023e) (shown in Figure 4.5) and DataSeg (Gu et al., 2023). In addition to these new modeling techniques, new datasets have been developed to mitigate this problem, including curating multi-domain segmentation datasets (Lambert et al., 2020), collecting high-quality annotations (Lu et al., 2023c) or scaling up to billions of masks (Kirillov et al., 2023).

Referring Segmentation by design is open-vocabulary. Models are usually designed specifically to learn from target datasets using various multimodal fusion strategies (Hu et al., 2016; Liu et al., 2017; Margffoy-Tuay et al., 2018; Ye et al., 2019a; Yu et al., 2016; Wu et al., 2022a). CLIPSeg (Lüdecke and Ecker, 2022) extends a textual query to a visual query and shows superior performance not only on referring segmentation but also on semantic segmentation. Since the emergence of vision transformers, works like LAVT (Yang et al., 2022e) enhance the cross-modal interactions from the very beginning, which leads to a decent performance on RefCOCO (Yu et al., 2016), RefCOCO+ (Yu et al., 2016) and G-Ref (Mao et al., 2016; Nagaraja et al., 2016). Differently, PolyFormer (Liu et al., 2023e) converts masks into polygons and asks the transformer decoder to decode a sequence of polygon coordinates. Inspired by Pix2Seq (Chen et al., 2022c), a similar method in object detection, PolyFormer presents an alternative way to represent masks for state-of-the-art referring segmentation. As we discussed earlier, one can also compose Grounding DINO (Liu et al., 2023h) with SAM (Kirillov et al., 2023) for referring segmentation.

Unified Segmentation. Given the above methods for open-vocabulary and referring segmentation, an open question is how to unify all segmentation tasks in a single framework. Recently, X-Decoder (Zou et al., 2023a) uses a generalized encoder-decoder architecture to unify all these segmentation tasks. The referring segmentation task is reformulated as a conditioned panoptic segmentation that takes some textual phrases as input to the decoder. UNINEXT (Yan et al., 2023) is another work that attempts to unify all instance-level segmentation in images and videos. Different from X-Decoder, UNINEXT uses early fusion to fuse the various prompts and vision features, which are then fed to the transformer encoder-decoder.

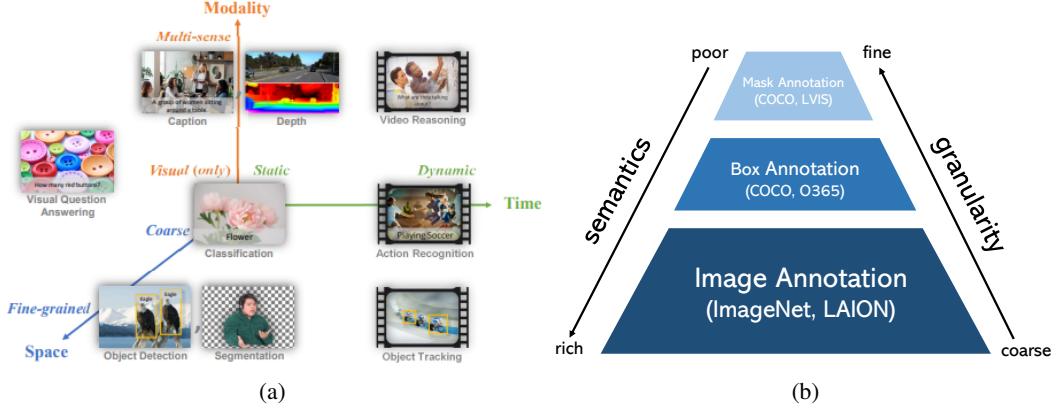


Figure 4.6: (a) CV task landscape: CV tasks can span different axes, including modality, space and time, which renders significant challenges to unify all of them in a single model. Image credit: Yuan et al. (2021). (b) The data scale pyramid: In particular, datasets in different tasks usually contain different types of supervision. Image-level datasets like ImageNet (Deng et al., 2009) and LAION Schuhmann et al. (2021) have annotations that have rich semantics coverage but are coarse-grained, while pixel-level datasets like COCO panoptic segmentation (Chen et al., 2015) provides fine-grained annotations but with limited concepts.

4.3 From Task-Specific Models to Generic Models

Above we have discussed the recent efforts of transforming closed-set models to open-set ones for detection and segmentation. Until recently, however, most vision tasks have been separately tackled with specialized model designs, preventing the synergy of tasks across different granularities or domains from being exploited. This is arguably due to two reasons:

- **Vision tasks are fragmented.** As shown in Figure 4.6 (a), computer vision tasks span across different axes including space, time, and modality. From the space aspect, it can be image-level, region-level and pixel-level tasks as we discussed before. Along the time axis, we need to tackle not only static images but also temporal video sequences. Regarding the modality, the inputs and outputs can be images, texts, or other types (*e.g.*, human pose, depth map). Such diverse task formats significantly impede the development of a unified model for all tasks.
- **Data scales are different.** In addition to the complicated task landscape, the scarcity of human annotations and their different scales for different tasks also make building a unified model challenging. In Figure 4.6 (b), we can see a clear pyramid of data scale, where different layers of human annotations have different semantics. More specifically, image-text datasets like LAION Schuhmann et al. (2021) contain up to 2B samples, while object detection datasets like Objects365 (Shao et al., 2019) have 1.7M images in total. More significant gap is observed in segmentation datasets due to the high cost of annotating masks.

Despite the aforementioned challenges, we are now witnessing a growing interest in building unified, general-purpose models that can learn from and be applied to a diverse set of vision and vision-language tasks, thanks to the versatility of transformers (Vaswani et al., 2017). These attempts can be grouped into two main categories:

- **I/O Unification.** Following the development of unified LLMs, a number of recent works reformulate many vision tasks as a sequence-to-sequence problem (Wang et al., 2022b; Yang et al., 2022c; Chen et al., 2022d; Lu et al., 2022a). They typically use a tokenizer to tokenize the original inputs and outputs (I/O) in different modalities used in various tasks into a coherent sequence (visual or text) tokens and then exploit a unified, sequence-to-sequence model.
- **Functionality Unification.** In addition to I/O unification, one might built a generic model via functionality unification. Extending multi-task learning methods (Lu et al., 2020; Gupta et al., 2022a; Hu and Singh, 2021a), many recent use a coherent encoder-decoder architectures (Yu et al., 2022a; Zhang et al., 2022b; Zou et al., 2023a). This line of work usually does not need task-

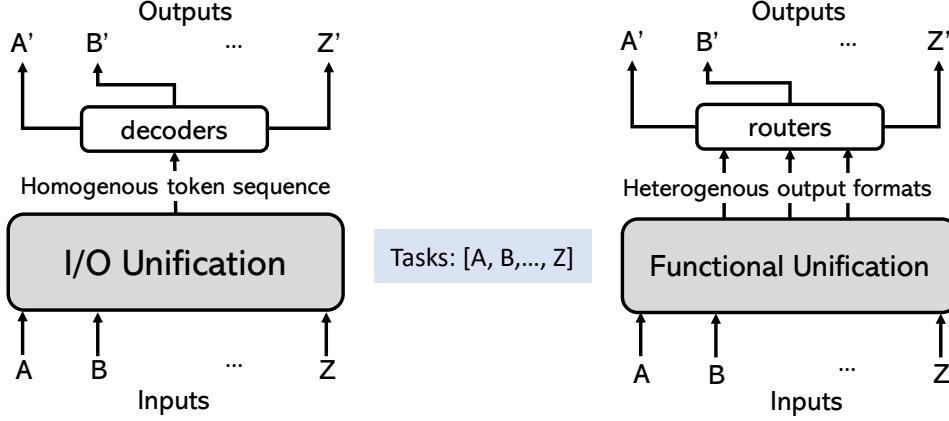


Figure 4.7: Side-by-side comparison between I/O unification and functionality unification. I/O unification is aimed at utilizing a decoder to decode homogeneous token sequences, which are then decoded by task-specific decoders. In contrast, functionality unification predicts heterogeneous outputs and then uses different routers or headers to produce the final outputs for various tasks.

specific or modality-specific tokenizers but requires a sophisticated model design to accommodate various tasks.

Figure 4.7 illustrates the difference between the two categories of unification methods. For I/O unification, the I/O unification module always generates a sequence of tokens, and exploits a separate decoder to decode the final outputs for different tasks. For functionality unification, the functional unification module generates heterogeneous outputs for different task, *e.g.*, semantic outputs and spatial outputs. Then, these different types of outputs are combined to produce the final task-specific outputs. Both unification methods strive to make use of synergy across tasks with different levels of granularity. For example, coarse-grained data is expected to contribute to rich semantic understanding required by fine-grained tasks, while fine-trained data to enhance the grounding ability for coarse-grained tasks. In the following, we review some recent works of these two categories.

4.3.1 I/O Unification

This line of work is mainly inspired by LLMs that unify many NLP tasks as sequential modeling. In the vision domain, the methods of building generic models via I/O unification can be grouped into two categories depending on the tasks of interest and output formats.

Sparse and discrete outputs

For vision tasks that produce sparse or discrete token outputs, we can easily exploit a language tokenizer, such as byte-pair encoding (BPE) (Sennrich et al., 2016), for I/O unification. In contrast, spatial outputs like boxes, masks, or human skeletons can be formulated as a sequence of numeric coordinates which are then tokenized into discrete tokens (Cho et al., 2021; Yang et al., 2022c; Liu et al., 2023e). As a result, the decoded output tokens are interleaved with organic textual tokens and numeric textual tokens to support a wide range of tasks. Without the loss of generality, the decoding process is formulated as auto-regressive generation and the model trained with the objective function defined as:

$$L(\theta) = - \sum_{t=1}^T \log p(s_t | s_{<t}, v; \theta), \quad (4.3)$$

where $\{s\}_{t=1}^T$ is the discrete token sequence of length T , and v is the visual feature. Below, we review some representative works.

UniTab (Yang et al., 2022c) unifies text and box output in a sequence decoding manner. As shown in Figure 4.8 (a), the box coordinates are represented by numerical numbers with $<\rangle$ and then a special token $<\text{obj}>$ is used to encompass the location information. In this way, the model can unify a variety of tasks that require textual and location outputs, including image captioning (Chen



Figure 4.8: (a) UniTab (Yang et al., 2022c) is proposed to unify grounded captioning, visual grounding, image captioning, VQA, and object localization. (b) Pix2Seqv2 (Chen et al., 2022d) is proposed to unify object detection, referring segmentation, keypoint detection, and image captioning. Image credit: Yang et al. (2022c) and Chen et al. (2022d).

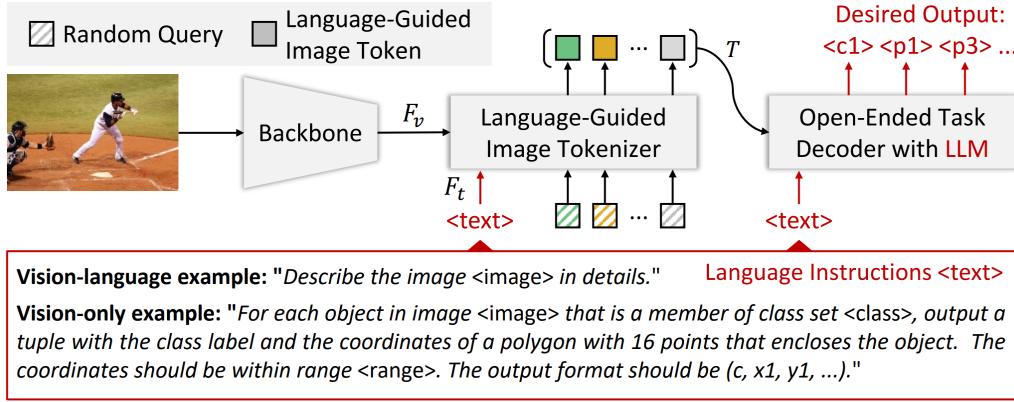


Figure 4.9: VisionLLM (Wang et al., 2023h) is proposed to bridge vision systems with LLMs in a sequential decoding manner. Image credit: Wang et al. (2023h).

et al., 2015), grounded captioning (Plummer et al., 2015), visual grounding, object localization and visual question answering (Antol et al., 2015). The model is trained in three stages: pre-training, multi-task finetuning, and task-specific finetuning.

Pix2SeqV2 (Chen et al., 2022d) slightly differs from UniTab in that it unifies two different vision tasks: referring segmentation and keypoint detection. Following Pix2Seq (Chen et al., 2022c), Pix2SeqV2 represents objects in an image as $[y_{min}, x_{min}, y_{max}, x_{max}, text]$. Then, it introduces a unique task prompt for each task, which contains task type information or a combination of task types and specific locations. For mask decoding, a mask contour is converted into a polygon and then its coordinates extracted from the polygon (Castrejon et al., 2017). A similar strategy is also used for referring segmentation, as in Polyformer (Liu et al., 2023e).

LLM-augmented. Recent works have also explored building a generic decoding interface based on LLMs, which are pre-trained on large amounts of text data and human instructions. Kosmos-2 (Peng et al., 2023b) exploits the pretrained LLMs of Kosmos-1 (Huang et al., 2023b) and augments the grounded multi-modal data by collecting a web-scale grounded image-text pair dataset (GRIT) consisting of 91M images. VisionLLM (Wang et al., 2023h) appends an even larger LLM (e.g., LLaMa (Touvron et al., 2023)) on top of an image tokenizer, as shown in Figure 4.9. The resultant model exhibits a very strong vision-language reasoning capacity and decent localization ability for object detection, segmentation, etc. Some other works that combine LLMs with grounding are DetGPT (Pi et al., 2023) and GPT4ROI (Zhang et al., 2023k). To further equip the model with the segmentation capability, both BubaGPT (Zhao et al., 2023c) and LISA (Lai et al., 2023) use an extra referring segmentation model to segment images by taking texts or embeddings as input, respectively. PaLI-X (Chen et al., 2023g) is by far the largest unified model that can cope with multilingual vision and vision-language tasks.

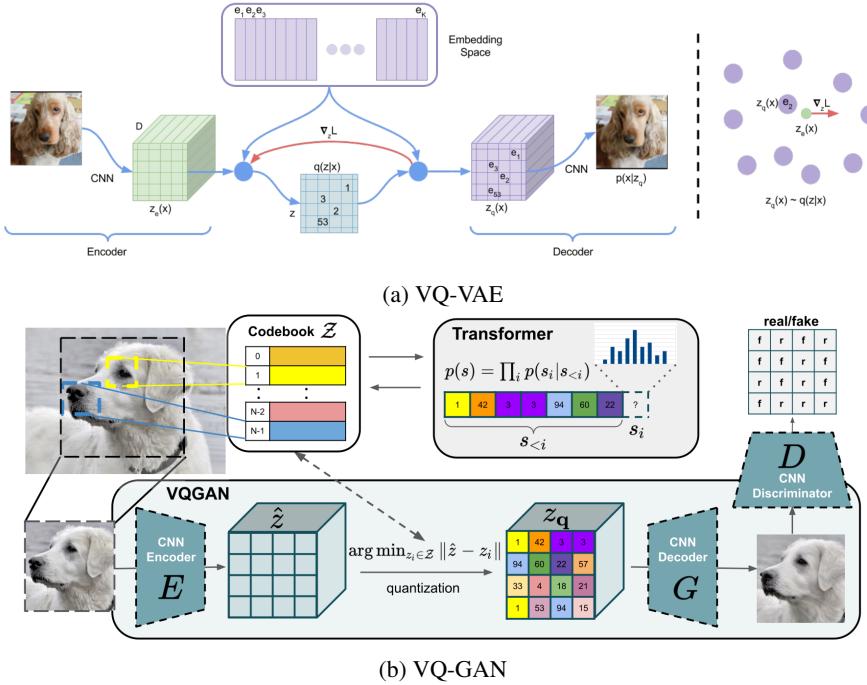


Figure 4.10: Illustration of VQ-VAE (Oord et al., 2017) and VQ-GAN (Esser et al., 2021).

Dense and continuous outputs

There are also some tasks that require dense and continuous outputs, such as image segmentation (He et al., 2017), depth estimation (Mertan et al., 2022), image inpainting and editing (Elharrouss et al., 2020; Brooks et al., 2023). Except for segmentation masks which can be approximated by polygons (Liu et al., 2023e; Chen et al., 2022d), most dense and continuous outputs cannot be easily converted into discrete tokens due to the high-dimensional space. Thus, we have to resort to an image-oriented tokenizer. Akin to the language tokenizer, an image tokenizer encodes raw images and extracts discrete tokens spanning the visual feature space. The most representative work is VQ-VAE (Oord et al., 2017; Razavi et al., 2019). As shown in Figure 4.10 (a), VQ-VAE learns an encoder z_e , a decoder z_q and a discrete codebook $e = \{e_1, \dots, e_K\}$ consisting of K embeddings. Given the input x , the posterior categorical probability $q(z|x)$ is defined as:

$$q(z = k|x) = \begin{cases} 1, & \text{for } k = \operatorname{argmin}_i \|z_e(x) - e_i\|. \\ 0, & \text{otherwise.} \end{cases} \quad (4.4)$$

where the decoder z_q takes x (or its representation e_k) as input to predict class label. As a variant of VQ-VAE, VQ-GAN uses a discriminator and the perceptual loss (Larsen et al., 2016; Lamb et al., 2016) to maintain a good balance between output quality and model efficiency (via high compression rate). In Figure 4.10 (b), we see that the discriminator is applied at the patch level to regularize the decoding of images at high resolution. Below, we discuss some most recent works that attempt to unify different vision and multi-modal tasks that involve dense outputs.

UVIM (Kolesnikov et al., 2022) is one of the first works that employ a dense decoding process to unify various core vision tasks, including panoptic segmentation, depth estimation and colorization. The learning process consists of two stages: (i) Base encoder-decoder f and restricted oracle Ω are learned to predict outputs given input images, where f takes raw image as input and Ω takes the desired output as input to decode the oracle code; (ii) Instead of using the desired output as input to the oracle Ω , the model learns a language model to produce the oracle code for the input raw image. Notably, the encoder-decoder model used here is trained with VQ-VAE objectives. As the first step to unify vision tasks with a single model, UVIM shows promising results on three vision tasks.

Unified-IO (Lu et al., 2022a) is another representative work. Compared to UVIM, it scales to many more vision tasks and datasets. Unlike the training procedure of UVIM, Unified-IO first trains differ-

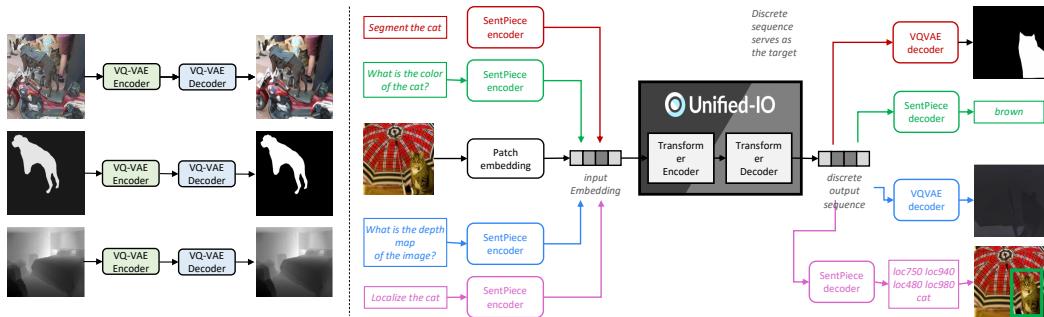


Figure 4.11: Unified-IO (Lu et al., 2022a) unifies different vision tasks by first pre-training VQ-VAE for each task and then an encoder-decoder for tasks jointly. Image credit: Lu et al. (2022a).

ent VQ-VAE models for different tasks, as depicted in Figure 4.11 left. After obtaining all VQ-VAE encoder-decoders, 90 datasets are combined to train another transformer encoder-decoder end-to-end, as shown on the right side. Similar to previous works, it also uses a language decoder to obtain the organic and numeric texts to generate coordinate outputs. After the second-stage pre-training, the model achieves state of the art on the GRIT benchmark (Gupta et al., 2022c) and exhibits compelling compositionality, although the performance still lags behind the strongest models on common tasks. As a follow-up, a soft-token strategy is proposed in Ning et al. (2023) to improve the accuracy for next token decoding. In addition, a masked modeling strategy is proposed to learn robust representations. Evaluated on instance segmentation and depth estimation, the model achieves state-of-the-art performance on NYUv2 (Silberman et al., 2012) and competitive performance on segmentation. A recent work uses image inpainting as the general task to unify different pixel-level vision tasks (Bar et al., 2022). Given the target discrete tokens produced by VQ-GAN, the method exploits a masked autoencoder to decode the missed image regions, using the task input-output examples as prompts. Painter (Wang et al., 2023i) extends this pipeline to facilitate more vision tasks and obtains competitive performance on various standard benchmarks.

Diffusion-augmented. Unlike the above works that learn their own decoding models, some recent works utilize the off-the-shelf stable diffusion model to build generalist vision models. For example, Prompt Diffusion (Wang et al., 2023m) initializes a model using Stable Diffusion and ControlNet (Zhang and Agrawala, 2023), and trains the in-context image-to-image model jointly on six different vision-language tasks, including segmentation, depth estimation, *etc*. InstructDiffusion Geng et al. (2023) also uses the diffusion model but explicitly introduces task-specific instructions to the diffusion process. Moreover, it uses task-specific training and human alignment training to enable a generalist interface for vision tasks.

4.3.2 Functionality Unification

Unlike I/O unification, functionality unification attempts to unify different tasks based on the task characteristics, with the awareness that they are neither fully isolated nor fully aligned. At a high level, vision tasks produce three types of outputs: (*i*) location outputs, (*ii*) semantic outputs, and (*iii*) pixel-level outputs. For example, both object detection and phrase grounding need to localize objects in the image, while both generic segmentation and referring segmentation produce masks. On the other hand, many tasks require semantic (or text) outputs to represent either concept names or textual descriptions.

Multi-task learning

Some early works explore multi-task learning methods for unifying different vision or vision-language tasks.

Vision models. A few works explore using CNNs for learning with different vision tasks at different levels. For example, Cross-stitch Networks (Misra et al., 2016) develops a strategy to split

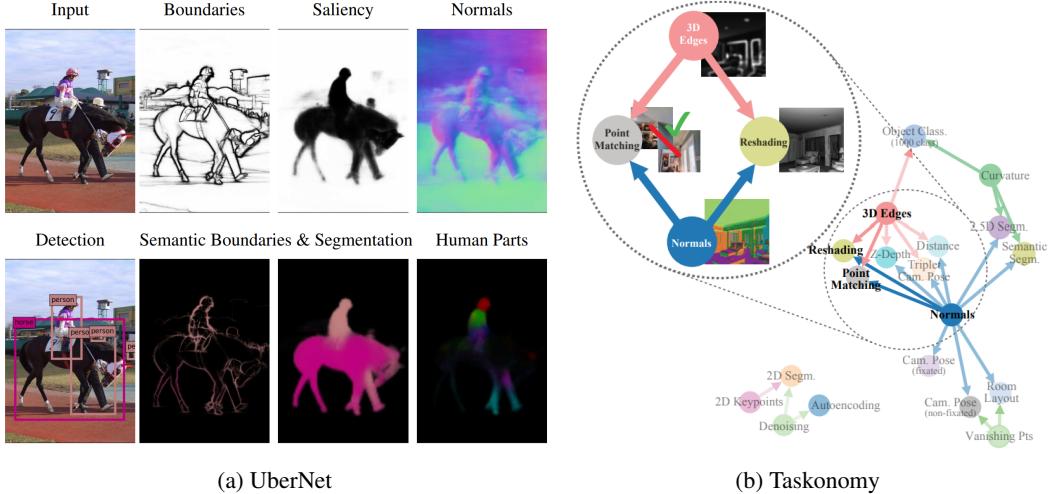


Figure 4.12: (a) UberNet (Kokkinos, 2017) can be applied to 7 vision tasks using a unified and budget-controllable CNN architecture. (b) Taskonomy (Zamir et al., 2018) further studies the relationship across vision tasks by exploiting a multi-task transfer modeling. Image credit: Kokkinos (2017) and Zamir et al. (2018).

different numbers of layers from the top in CNNs so as to adapt to different vision tasks. Results show that the best-performing multi-task architecture depends on the tasks of interest and can hardly generalize to new tasks. UberNet (Kokkinos, 2017) takes one step further to use a single universal CNN architecture and sophisticatedly design a routing mechanism to save the memory and computing cost, as shown in Figure 4.12 (a). Both works require some tweaking to the CNN architecture so that they can adapt to different levels of tasks and loss types. But they unfortunately fail to build the synergy across tasks to improve model performance. Taskonomy (Zamir et al., 2018) specifically studies the relationship among vision tasks. It first trains task-specific models for each individual task and then performs transfer modeling across tasks in the latent space. The task affinity is then calculated in the latent space, providing us with the taskonomy. The result shows that vision tasks have different affinities for different groups, as shown in Figure 4.12 (b). For example, surface normal estimation is heavily related to reshaping and point matching. Curvature extraction is related to image segmentation tasks. This study provides deep insights for multi-task vision modeling (Xu et al., 2018; Crawshaw, 2020).

Multi-modal models. The emergence of Transformers significantly facilitates the advancement of multi-task multi-modal learning. Among them, 12in1 (Lu et al., 2020) is one of the pioneering works that combine 12 vision-language tasks in a single BERT-based architecture. It uses task-specific heads for individual tasks and a commonly shared trunk ViLBERT (Lu et al., 2019). Results show that multi-task learning can achieve substantial improvements over single-task learning while reducing the model parameters significantly. Later on, UniT (Hu and Singh, 2021b) exploits an encoder-decoder architecture and expands to vision-only tasks like object detection. Additionally, it allows end-to-end training on the task pool without relying on pre-trained detectors. Similar to 12in1, it also uses a task-specific head for each task, motivated by the empirical result that sharing the same head usually hurts performance. Likewise, E2E-VLP (Xu et al., 2021) proposes an end-to-end pipeline for both localization tasks and text generation. Both UniT and E2E-VLP demonstrate the versatility of the encoder-decoder architecture of DETR (Carion et al., 2020). Following the same spirit, GPV (Gupta et al., 2022b) proposes an end-to-end task-agnostic architecture for different vision and vision-language tasks. It uses DETR to extract boxes and region features and then exploits a cross-attention module for fusion, followed by a vision decoder and a language decoder for decoding different outputs.

The above vision and multi-modal models unify different tasks by incorporating different modules or heads designed to cope with different tasks, and can hardly achieve synergy across tasks. In the following, we discuss recent model unification research that aims to make the best use of synergy among various vision and multi-modal tasks.

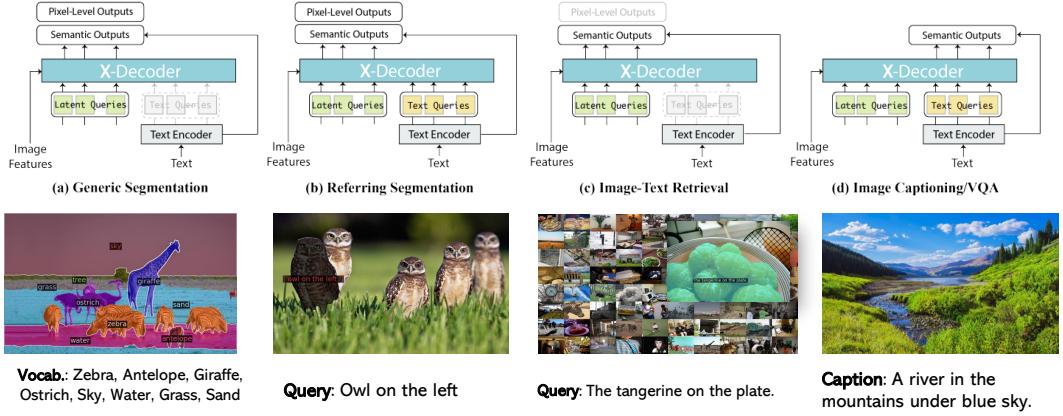


Figure 4.13: A generalist model X-Decoder (Zou et al., 2023a) that unifies different vision and vision-language tasks in a functional manner. It uses a single decoder with the same suite of parameters, but different routing mechanisms to tackle different tasks. Image credit: Zou et al. (2023a).

Unified learning

The barrier across tasks is gradually blurred thanks to the use of Transformers (Vaswani et al., 2017) and the development of open-set models as we discussed earlier. It is now possible to bind inputs from different modalities to learn a shared semantic space. A number of works (Zhang et al., 2022b; Zou et al., 2023a; Li et al., 2023g) have recently been proposed to unify vision and vision-language tasks by *using one model for all*. After pre-training, the single model can be applied to tackle all tasks in a zero-shot manner and the performance can be further improved via task-specific finetuning. Note that unified learning in this context differs from previous works of large-scale pre-training. Like GPT which serves as a universal language interface after pre-training, a unified vision model is not only a representation learning engine but also an interface that supports as many tasks as possible in a zero-shot manner. Below, we review a few representative works.

GLIPv2 (Zhang et al., 2022b) is proposed by extending GLIP (Li et al., 2022f) to support a wide range of vision and vision-language tasks, including grounded captioning, visual question answering, etc. GLIPv2 seamlessly integrates localization pre-training and Vision-Language Pre-training (VLP) through three distinct pre-training tasks: (i) phrase grounding, which serves as a vision-language adaptation of detection tasks; (ii) region-word contrastive learning, introducing a novel contrastive task at the region-word level; and (iii) masked language modeling. In a zero-shot manner, this pre-trained model can be applied to different tasks and attain plausible performance across the board. Unlike previous works (*e.g.*, GPV (Gupta et al., 2022b)), it merges the localization module and vision-language matching module in a coherent manner, which makes model training from fused data much more efficient and effective.

X-Decoder (Zou et al., 2023a) follows the generic design of encoder-decoder architecture. Given an input image, it first uses an image encoder to extract features at multiple scales. Afterward, a text encoder is used to encode a textual query into a sequence of embeddings. The visual features, textual queries and the non-semantic or latent queries are fed to a decoder to predict the outputs. Three critical designs are proposed to empower the generalization ability of X-Decoder to a variety of vision and vision-language tasks: (i) It defines two types of queries and outputs. Specifically, the queries for the decoder are categorized into latent queries and text queries, which undertake generic vision and vision-language tasks, respectively. Likewise, the output is categorized into pixel-level masks and semantic embeddings; (ii) A single text encoder is exploited to encode the textual corpus from all tasks. The common text encoder is used to encode referring phrases, text descriptions, and image captions in the task of referring segmentation, image-text retrieval and image captioning, respectively; (iii) It fully decouples the image and text encoder, and use all the outputs as queries. As such, it can learn from both intra-image supervisions and inter-image ones, which is essential to learn stronger pixel-level representations and support different granularity of tasks. As shown in Figure 4.13, the pre-trained model can support different tasks by taking different routing while sharing the same suite of parameters.

Uni-Perceiver-v2 (Li et al., 2023g) is another generalist model that unifies vision and vision-language tasks. Similar to X-Decoder, the model exploits a vision encoder, a text encoder and a general decoder. Differently, it introduces a region proposal network on top of the vision backbone to explicitly predict the boxes and masks, which are then encoded as “queries” for the general decoder. To jointly train on datasets at different levels, it introduces a unified max-likelihood estimation strategy for tasks with localization and without localization.

4.4 From Static to Promptable Models

The success of Large Language Models (LLMs) such as ChatGPT (OpenAI, 2023b) have shown the importance of modern AI models in interacting with humans, and have provided a glimpse of AGI (Bubeck et al., 2023). The ability to interact with humans requires a user-friendly interface that can take as many types of human inputs as possible and generate responses that humans can easily understand. In NLP, such a universal interaction interface has emerged and evolved for a while from early models like GPT (Brown et al., 2020) and T5 (Raffel et al., 2020), to more advanced techniques like prompting (Shin et al., 2020; Zhao et al., 2021; Li and Liang, 2021) and chain-of-thought (Wei et al., 2022a; Kojima et al., 2022; Schick et al., 2023). However, most vision models are still static in that they are less flexible than LLMs to various prompts. Most recently, a number of works have proposed to enhance the static vision models with the capabilities to support: (i) multi-modal prompting; (ii) in-context prompting.

4.4.1 Multi-modal Prompting

Vision is different from language by nature. To enable a smooth interaction between humans and AI, a model requires not only language prompts but also other types of prompts to complement the missing information or resolve the ambiguity in language. Recently, a number of works have explored how to combine or augment language prompts with other types of prompts, such as spatial prompts (Kirillov et al., 2023), visual prompts (Zou et al., 2023b) and other modalities (Girdhar et al., 2023; Liu et al., 2023f). In the following, we review some representative works.

Spatial prompting. Vision is rooted in the physical world, and as such it is not only semantic but also spatial by nature. Spatial prompting can be considered as a way to modulate the vision models through the inputs of location information, which could be a point, a box, or an arbitrary stroke, *etc*. Such clues have been heavily used in UI designs of computers (*e.g.*, mouse) and mobile devices (*e.g.*, touch screen). In computer vision, interactive segmentation (Mortensen and Barrett, 1998; McGuinness and O’Connor, 2010; Chen et al., 2021c, 2022i) naturally requires such capability so that the model can take multiple clicks from users and gradually refine the segmentation mask. However, most of these works are still designed task-specifically and lack enough flexibility to support different types of spatial prompts.

SAM (Kirillov et al., 2023) is one of the pioneering works that propose a convenient spatial prompting interface and learn a foundation model for image segmentation. As shown in Figure 4.14, the model can take points or boxes as the prompts, and segment images in arbitrary granularity. The ability to segment images following the user instructions from humans makes it readily a foundation to build many more models and applications (Zhang et al., 2023c). To name a few, a number of works (Ma and Wang, 2023; Roy et al., 2023) start with SAM and train a promptable segmentation model for the medical domain. Spatial prompting is particularly beneficial in that the textual annotations for medical images are usually limited and hard to interpret. Similar cases also happen in other industry domains (Tang et al., 2023a). To further improve point prompting, SAMAug (Dai et al., 2023a) proposes to refine the points using the max entropy criterion and saliency map, which can help to determine the most informative locations the model should look at.

Visual prompting. In many cases, textual descriptions of objects are not necessarily clear to convey the information. For example, given an unrecognizable or indescribable object, people may fail to express themselves clearly about the object. In this case, showing one or a few examples would be more informative and straightforward. With this idea, a lineup of works have studied exemplar-based visual modeling, such as image-to-image retrieval (Yoon et al., 2021; Datta et al., 2008; Zhang et al., 2018), image co-segmentation (Joulin et al., 2010; Jerripothula et al., 2016) and visual object tracking (Yilmaz et al., 2006; Luo et al., 2021; Wu et al., 2013). Most recently, this strategy has

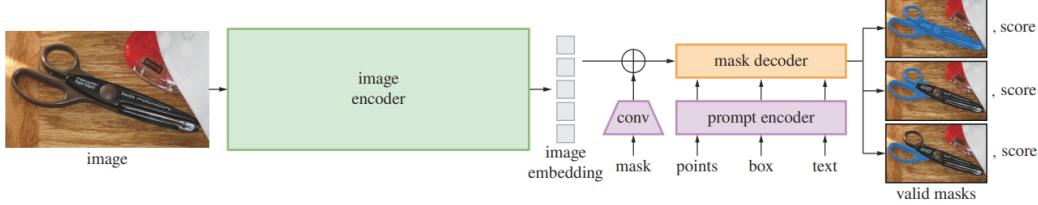


Figure 4.14: SAM (Kirillov et al., 2023) proposes a promptable segmentation model which can take different spatial prompts in addition to text prompts. It further develops a data annotation engine to scale up the mask-annotated data. Image credit: Kirillov et al. (2023).

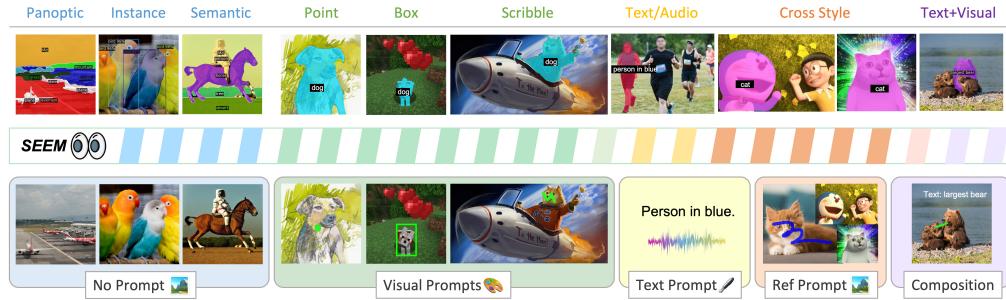


Figure 4.15: SEEM (Zou et al., 2023b) can take different types of prompts as inputs for various image segmentation tasks. Image credit: Zou et al. (2023b).

been formulated as visual prompting in that different types of visual inputs are usually encoded to some unified format and then fed into a Transformer architecture, as shown in LLMs.

SEEM (Zou et al., 2023b) is one of the representative works that enable visual prompting to a vision model for image segmentation. As shown in Figure 4.15, SEEM differs from the aforementioned SAM and can take visual prompts by drawing points, boxes, and strokes on an image that can be the target image or another reference image. It develops a new module called a visual sampler that can extract visual features from an image according to the locations specified by users. Based on the visual sampler, the model can even take another reference image as input without any training like that. As a result, it shows impressive performance not only for various image segmentation tasks but also for video object segmentation in a zero-shot manner.

PerSAM (Zhang et al., 2023h) develops a personalized segmentation model on top of SAM and takes one shot as the input. It learns a specific model that takes a source image plus a mask as input and then predicts the mask for a target image. To extract the visual prompts, mask pooling is taken and used as the input tokens to the decoder of PerSAM. It also proposes a way to extract the positive and negative priors based on feature matching to facilitate pre-trained SAM models with comprehensive clues. Like most prompt learning methods in LLMs, a plausible feature for PerSAM is that it can be easily attained by some off-the-shelf models like SAM. SAM-PT (Rajić et al., 2023) further applies this strategy to video object segmentation. Inspired by the spatial prompting in SAM, it exploits a point-tracking system (Harley et al., 2022) to track different points (both positive and negative ones) and then ask SAM to segment the image given the points. It exhibits strong point tracking performance as well as segmentation.

Others. Some other works combine a wide range of visual prompting types. For example, Painter (Wang et al., 2023i) reformulates different vision tasks (*e.g.*, depth estimation, image segmentation) all as prompting and learns a decoder in an in-context learning manner. The prompts are combinations of raw images and the corresponding dense annotations (*e.g.*, depth or segmentation maps). In contrast, Prism (Liu et al., 2023f) makes use of many off-the-shelf vision models to extract different information from the raw images and then feed the information to a vision-language model. To facilitate the interplay across multiple modalities, ImageBind (Girdhar et al., 2023) learns a universal alignment among image/video, language, audio and depth. Once the embedding space is learned, it can be used to compose different types of prompts by simply doing the summations.

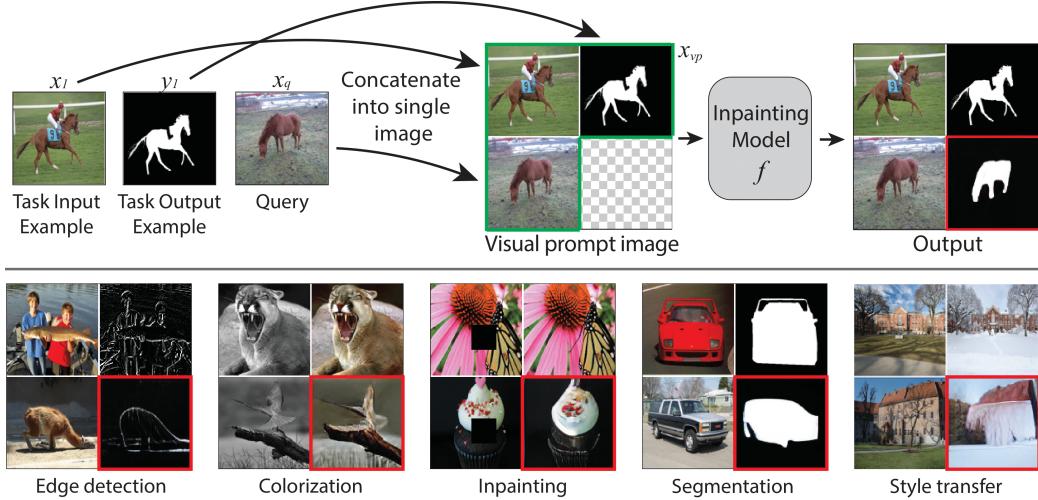


Figure 4.16: Visual prompting via visual inpainting proposed in Bar et al. (2022). Image credit: Bar et al. (2022).

4.4.2 In-context Prompting

The in-context learning capability has been observed in many LLMs such as GPT-3 (Radford et al., 2019), which makes the model more configurable via prompting without any model parameter updates. In contrast, till now, the in-context learning capability for vision models is still less studied. Flamingo (Alayrac et al., 2022) is one of the pioneering works that demonstrate in-context language generation for multi-modal inputs, which is acquired by learning from interleaved image-text pair data. Likewise, Kosmos-1 (Huang et al., 2023b) is another work that takes visual inputs as a foreign language so that the in-context learning ability in LLMs can be naturally translated to multi-modal inputs. However, both methods take multi-modal data as inputs but merely generate texts as outputs. As we discussed earlier, vision tasks require outputs of different types beyond texts. How to endow the in-context learning ability for vision systems is still an open question. Below, we review some recent attempts toward the goal.

Visual prompting via inpainting is proposed in Bar et al. (2022) to teach the model to predict dense outputs, such as edges, masks, depths, etc. as shown in Figure 4.16. Given an input image $\mathbf{x} \in \mathcal{R}^{H \times W \times 3}$ and a binary mask $\mathbf{m} \in \{0, 1\}^{H \times W}$, an inpainting model is to predict the missing region $\mathbf{y} = f(\mathbf{x}, \mathbf{m})$. The authors exploit a pre-trained VQ-GAN to encode the original image into discrete tokens and ask another ViT encoder to predict the masked regions. To make sure the model understands the visual “context” in the images, the authors collected a new dataset called *Computer Vision Figures* dataset which consists of 88k images from Arxiv papers. After pre-training, the model is used to predict the content at the bottom-right corner.

Concurrently, Painter (Wang et al., 2023i) extends a similar idea of visual in-context learning to more diverse datasets and benchmarks. Unlike Bar et al. (2022), it predicts the output in the continuous pixel space instead of discrete tokens. For different tasks, the authors define rules to convert the output spaces into image spaces. For example, it uses different colors to represent different individual instances in the image for the segmentation task. After unifying the input and output format, the authors use vanilla ViT as the encoder and masked image modeling (He et al., 2022a). A follow-up work called SegGPT (Wang et al., 2023j) is built on top of Painter and designed specifically for image segmentation tasks. The pre-trained model can be easily extended for exemplar-based image segmentation tasks.

Hummingbird (Balažević et al., 2023) resorts to a different method for in-context visual learning. Instead of using masked modeling, the authors propose to leverage attention across target and source images to aggregate the information. As shown in Figure 4.18, the models take multiple input images (first row) and corresponding semantic label maps (second row). Given a query image, it first finds the nearest neighbor feature locations in the prompt images for the query points and then projects the same matches to the semantic label maps so as to aggregate the label for the target query. This

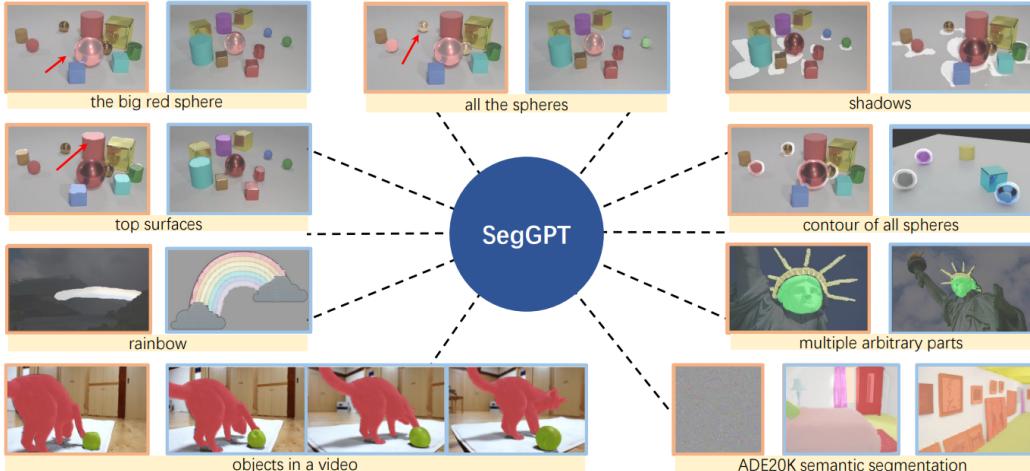


Figure 4.17: SegGPT (Wang et al., 2023j) proposes to perform in-context learning for image segmentation. Image credit: Wang et al. (2023j).



Figure 4.18: Hummingbird (Balažević et al., 2023) is proposed for in-context visual scene understanding through nearest-neighbor retrieval. Image credit: Balažević et al. (2023).

strategy is akin to earlier works that build classification models based on K-nearest-neighbor but differently applied to dense prediction tasks.

Discussion. In-context learning is arguably an appealing feature. On one hand, there are a number of works that attempt to bridge vision with LLM so as to inherit the in-context learning capability such as Flamingo (Alayrac et al., 2022) and Kosmos-1 (Huang et al., 2023b). On the other hand, researchers resort to pure vision-based in-context learning to address vision-specific tasks such as image segmentation, depth estimation, *etc*. Thus far, there is no single model that can take multi-modal inputs and predict different types of outputs as well in an in-context learning manner, which may render a promising future direction along this line.

4.5 Summary and Discussion

To the end, an illustrative summary of the works that have been covered in this chapter is shown in Figure 4.19. There is a clear trend in the vision community to build *open-world, unified and interactive* vision models. Nevertheless, there are still some intrinsic differences between vision and language. First, vision differs from language in that it captures the physical world with raw signals. We need to develop some sophisticated tokenization methods to compress the raw data into compact “tokens”. In the language domain, this can be easily done by using some well-established heuristic tokenizers (Sennrich et al., 2016). Second, unlike language, vision data itself is not labeled and thus difficult to convey information or knowledge. It always requires human labors to annotate the visual contents in either a semantic or spatial manner. Third, language data is homogeneous while vision data and tasks are heterogeneous. Last but not least, storing vision data is much more costly

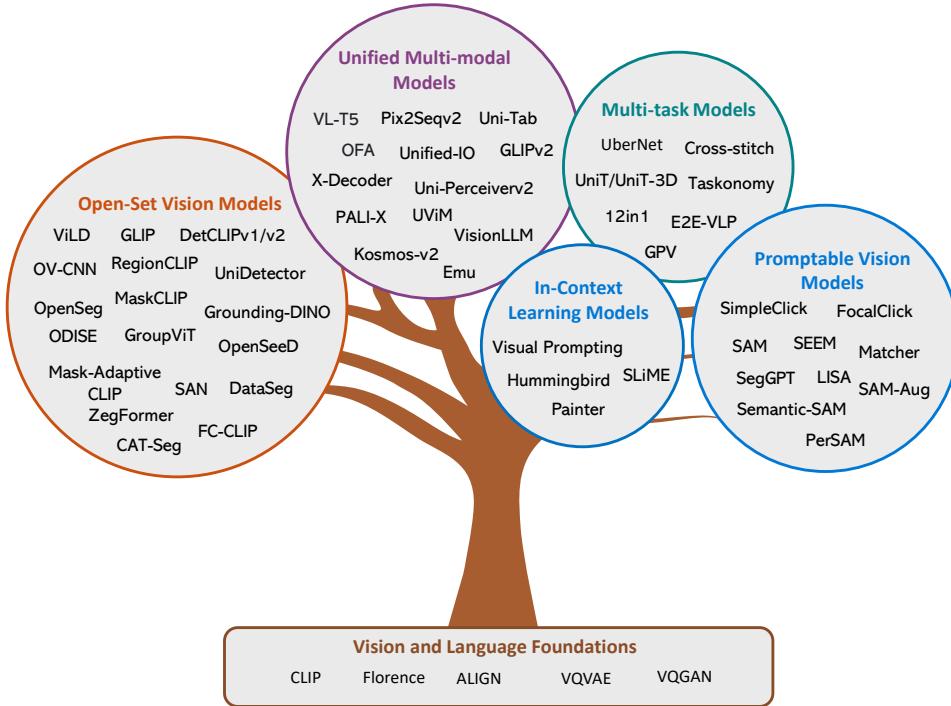


Figure 4.19: A summary of topics covered in this chapter. A lot of effort has been spent to unify vision models from different aspects to enable more intelligent vision systems.

than language data. For example, GPT-3 consumes 45 TB of training data, while the ImageNet dataset which contains 1.3M images costs more than hundreds of gigabytes. When it comes to video data like Howto100M (Miech et al., 2019), the storage cost already exceeds that of training corpus for GPT-3. All these differences cast some open questions that need to be addressed in the vision community, detailed below.

- **Computer vision in the wild.** Due to the heterogeneous nature, the current vision data we use for training models can hardly cover the full picture of the physical world. Despite the effort in building open-set vision models, we are still facing significant challenges in coping with novel or long-tail scenarios.
- **Scaling law in vision.** As discussed in Kaplan et al. (2020); Hoffmann et al. (2022), the performance of large language models improves smoothly with the increase of model size, data scale, and amount of computes. As the scale increases, some intriguing emerging properties are further observed in LLMs. In contrast, it is still not clear what is the right path to scale vision models, not to mention the emerging properties in such models.
- **Vision-centric or language-centric models.** Currently, the boundary between vision and language is gradually dismissed. However, due to intrinsic differences between vision and language, it is still not clear whether we should further scale up the vision models and integrate language models or the combination of moderate vision models and LLMs is sufficient to address most (if not all) of the problems.

With that being said, we are close yet still far away from an intelligent vision system that can perceive the world like humans. We hope the literature review in this chapter could provide an overall picture of the existing efforts, and inspire the pursuit of next-generation vision models.

Chapter 5

Large Multimodal Models: Training with LLM



In this chapter, we comprehensively explore large multimodal models (Alayrac et al., 2022; OpenAI, 2023a). We begin with Section 5.1 to delve into the background of such models, with the focus on the basics of image-to-text generative models and their representative model instances in various case studies. We also discuss the state-of-the-art OpenAI Multimodal GPT-4 (OpenAI, 2023a) and identify the existing research gaps in the field. To better understand the process of instruction tuning in large language models, Section 5.2 examines its importance and its role in self-instruct and open-source LLMs. Moving forward, we explore instruction-tuned large multimodal models in Section 5.3, shedding light on their basics, significance and applications. Additionally, Section 5.4 touches upon advanced topics in the realm of multimodal models to provide a deeper understanding of the subject. Finally, we assess the current progress in the field by evaluating how close we are to achieving the OpenAI Multimodal GPT-4 in Section 5.5, a major milestone in AI research.

5.1 Background

5.1.1 Image-to-Text Generative Models

LMMs in their current form is primarily an image-to-text generative model, which takes images as input, and outputs a text sequence. One example is illustrated in Figure 5.1 (a) Left. All of the model variants share a very similar model architecture and training objective.

- *Model Architecture.* As illustrated in Figure 5.1 (a) Right, the model typically consists of an image encoder to extract visual features, and a language model to decode the text sequence. The vision and language modalities can be optionally connected by trainable connection module. The image encoder and language model can be either trained from scratch or initialized from pre-trained models.
- *Training Objective.* As illustrated in Figure 5.1 (b), it typically employs an auto-regressive loss on the output text tokens. For the attention map in the Transformers (Vaswani et al., 2017), image tokens can attend to each other, and the current text token attends to all image tokens and the previous text tokens.

5.1.2 Case Studies

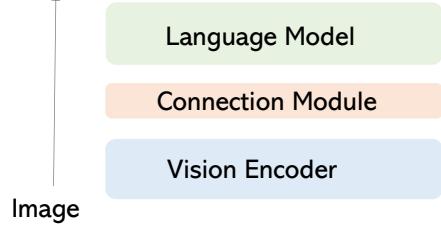
We use some prominent LMMs as examples to illustrate how the aforementioned network architecture can be instantiated in different models, while maintaining the same auto-regressive training objective.

Case study I: LMM trained with image-text pairwise instances. Most LMMs are trained on a large number of image-text pairs, where each training sample is a pair. GIT (Wang et al., 2022a) and

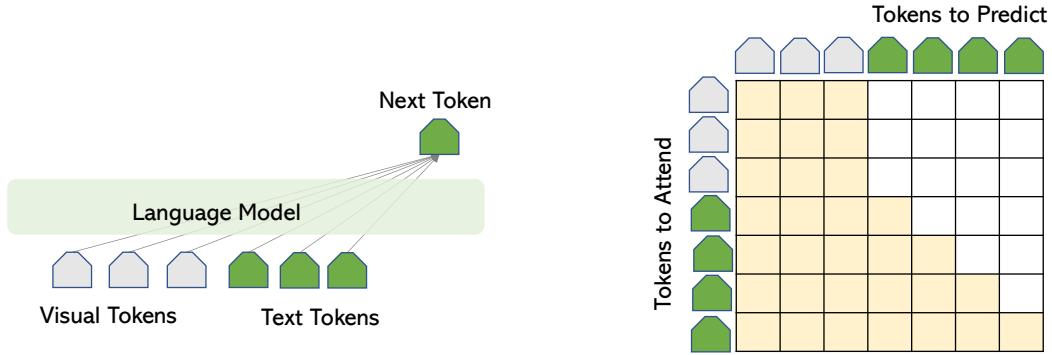
A dog lying on the grass next to a frisbee



Language



(a) Left: An example of image-to-text generation task; Right: Model architecture.



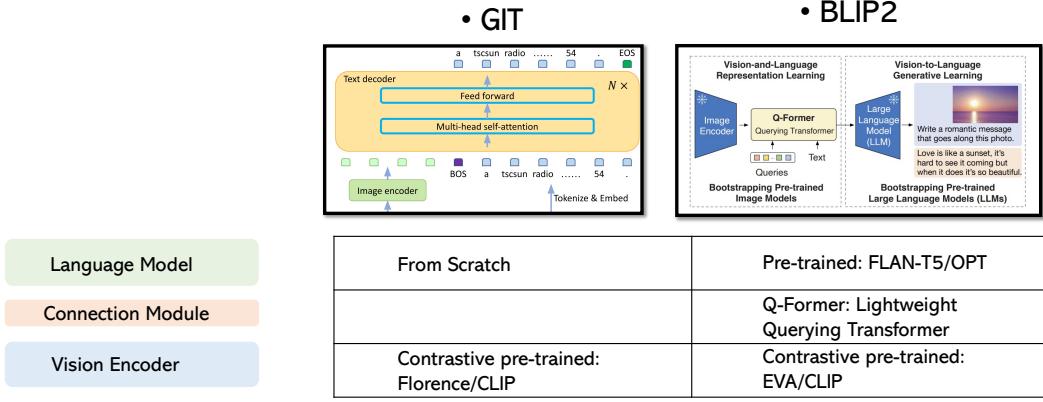
(b) Training objective and attention mask. For each row, the yellow elements indicate that the prediction token attends the tokens on the left.

Figure 5.1: Illustration of image-to-text generation task, architecture, and training objective.

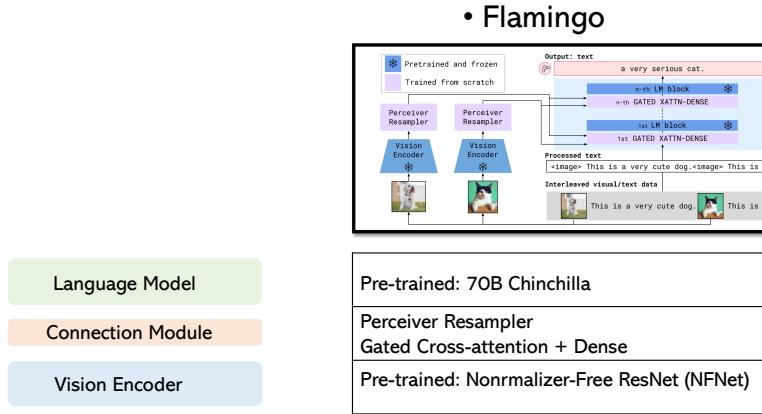
BLIP2 (Li et al., 2023h) are two large models that achieve state-of-the-art (SoTA) performance on many datasets. The comparisons are shown in Figure 5.2(a). GIT initializes the image encoder with contrastively pre-trained Florence model (Yuan et al., 2021), and trains the language model from scratch. On the other hand, BLIP2 freezes the weights of a pre-trained image encoder and a pre-trained language model, while trains a lightweight Q-former module to connect the image encoder and the language model.

Case study II: LMM trained with interleaved image-text sequence instances. We use Flamingo (Alayrac et al., 2022) as an example, shown in Figure 5.2(b). It connects the frozen pre-trained image encoder and language model – by adding novel architectural components in between. Specifically, Perceiver Sampler module helps reduce computational complexity, and Gated Transformer module helps to stabilize training during the initial stage. Flamingo is trained on a mixture of complementary large-scale multimodal data coming only from the web, without using any data annotated for machine learning purposes. After this training is done, Flamingo can be directly adapted to vision tasks via simple few-shot learning without any additional task-specific tuning.

Multimodal in-context-learning. Beside the SoTA performance on dozens of academic benchmarks, probably the most appealing aspect of Flamingo is the emerging property: Multimodal In-Context-Learning. Specifically, given a couple of image-text pairs as examples, Flamingo can zero-shot task transfer to unseen problems, such as solving visual math problems. This means Flamingo can tackle a number of difficult problems with just a handful of task-specific examples, without any additional training required. For example in Figure 5.3, two new tasks are presented to Flamingo. The top row provides two image-text pairs as the context in the prompt, where the text describes the name of the animal in the image, followed by the geographical information of the animal. Flamingo is able to understand the patterns presented in the examples, and output the corresponding information for a new image. In the bottom row, the text first shows the optical character recognition (OCR) result of the image, followed by the answer to the math problem. Flamingo follows the task instruction illustrated in the multimodal context, and outputs the correct answer for a new math problem in the third image. This intriguing in-context learning capability makes Flamingo the GPT-3 moment (Brown et al., 2020) in the multimodal domain.



(a) Example 1: LMM trained with image-text pairs.



(b) Example 2: LMM trained with image-text pairs and interleaved image-text data.

Figure 5.2: Examples of image-to-text generation models. Image credit: Wang et al. (2022a); Li et al. (2023h); Alayrac et al. (2022).

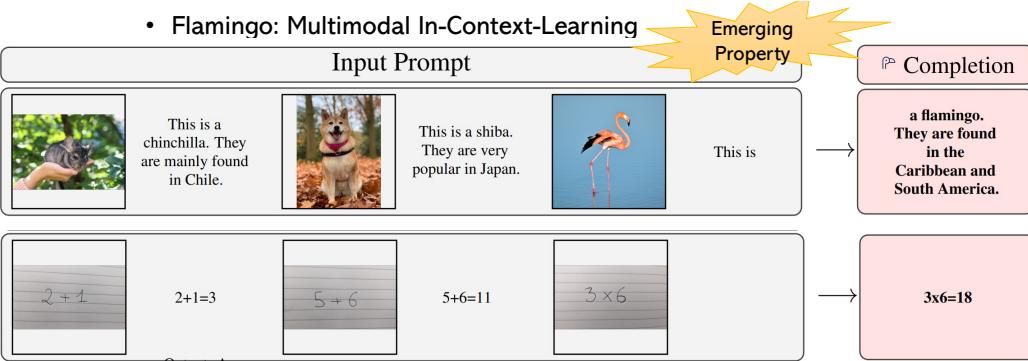


Figure 5.3: The emerging property of pre-training on web-scale interleaved image-text data: multi-modal in-context-learning. Examples are adopted from Alayrac et al. (2022).

5.1.3 OpenAI Multimodal GPT-4 and Research Gaps

In March 2023, OpenAI released GPT-4 (OpenAI, 2023a), with impressive capability in visual understanding and reasoning. Though the model details are not revealed, there is no doubt that GPT-4 enables many new scenarios, based on the examples highlighted in the technique report. For instance, two popular visual examples are illustrated in Figure 5.4. The first one identifies the uncommon visual region and exhibits strong complex reasoning performance. The second one recognizes text in the image and captures the mere across image-text. For a while, the research community had

OpenAI MultiModal GPT-4

- *Model Details: Unknown*
- *Capability: Strong zero-shot visual understanding & reasoning on many user-oriented tasks in the wild*

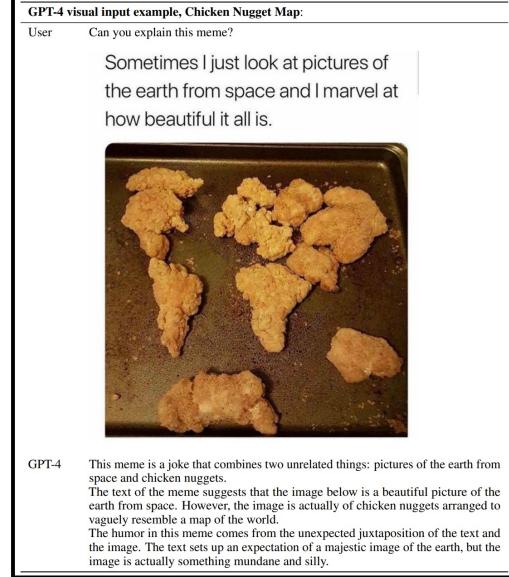


Figure 5.4: OpenAI Multimodal GPT-4. Visual examples are from [OpenAI \(2023a\)](#).

no clue how this new ability is achieved (probably because they are not tightened to any established academic tasks/datasets), but all are determined that these are exciting results. It naturally raises a question: how can we build Multimodal GPT-4 like models?

To answer it, we start to review the big models from OpenAI, by highlighting the most appealing properties for each model in Figure 5.5. There are several key observations: (i) GPT-2 ([Radford et al., 2019](#)) is the auto-regressive counterpart in the BERT era ([Devlin et al., 2019](#)) for the pre-train-then-finetune paradigm. Compared with GPT-2, GPT-3 ([Brown et al., 2020](#)) is a 175B model trained on web-scale text corpus, which exhibits two emerging properties with a frozen model: in-context-learning ([Brown et al., 2020](#)) and chain-of-thoughts (CoT) reasoning ([Wei et al., 2022a](#)). This means, without any additional training, the model can tackle a wide range of new problems with just a few task-specific examples and by properly prompting it step-by-step, respectively. It further leads to the modeling paradigm from task-specific finetuning to prompting frozen models, where the latter shows higher generalizability and lower adaptation cost in task transfer. (ii) ChatGPT and InstructGPT ([Ouyang et al., 2022](#)) show the importance of instruction-following and alignment with human intents for LLMs, by finetuning the base language model GPT-3/GPT-3.5 on high quality instruction-following data, and improving them with a reward model via reinforcement learning with human feedback. (iii) GPT-4 not only improves the language ability of previous models, but also allows visual signals as additional input for visual understanding and reasoning. We see that the newer generation model maintains/improves the existing properties of the previous ones, and enable new properties.

In other words, from GPT-3 to GPT-4, we see two new properties: instruction-following and multimodal input. This reveals the gap between existing LMMs (*e.g.*, Flamingo) and multimodal GPT-4: how to perform instruction-following and alignment research in the multimodal space, which is the focus of this chapter.

5.2 Pre-requisite: Instruction Tuning in Large Language Models

Note that instruction-following is a notion originated in NLP. To study the intuition behind it and have a full picture of its history, we first revisit instruction tuning with LLMs.

Traditional language data. As a typical data instance in NLP, sequence-to-sequence (seq2seq) representation is widely adopted for many language tasks: each data instance consists of two parts: one sequence as the input and another sequence as the output. We provide two examples in Fig-

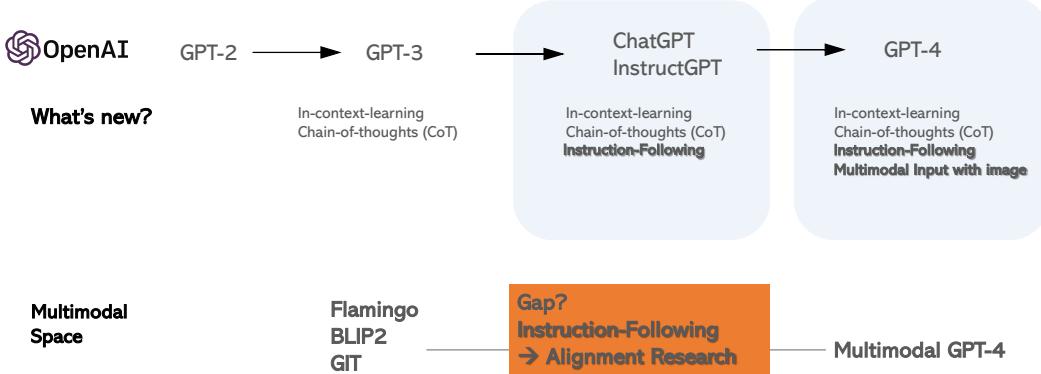


Figure 5.5: Recap on Language Modeling: OpenAI LLM development history. The unique properties for each generation model are highlighted, from which the research gap is revealed for LMM.

ure 5.6 (a). Without any task instruction specified, we know they are translation and summarization tasks, respectively.

This seq2seq representation is also the conventional data format in NLP research, where task instructions are implicit. Based on each data domain, individual models are trained. Or sometimes one model is trained with multi-task objectives over multiple data domain without specifying the task instructions. For both cases, the models are hard to generalize to new tasks in a zero-shot fashion, as they are not trained to understand task instructions, thus cannot distinguish and generalize what task to perform during testing time.

Instructional language data. Recently, researchers have started to explicitly add task instructions into the model training, as shown in Figure 5.6 (b). Interestingly, the task instruction of most NLP tasks can be expressed in natural language as well. It leads a new data format: instruction-input-output triplets. Based on the new format, one single model can be trained to perform multiple tasks, each with its specific instructions. Since models have observed many task instructions and many instances for each task during training, it is more natural and easier for them to generalize to new tasks by task composition in the inference stage.

For example, in the evaluation stage, a new task that requires both summarization and translation is provided in Figure 5.6 (c). Though the model has never seen this new task during training, it observes individual task basis, and learns to perform on new tasks. Note that we humans are always creating new tasks in our daily life, and presumably these new tasks would never been observed by models. It is thus appealing if a model is able to solve thousands of new tasks in the wild without training. This is partially why ChatGPT is becoming popular and prevalent so quickly.

5.2.1 Instruction Tuning

How can we collect a diverse set of high-quality instruction-following data? There are two general schemes. One is through human-human interaction, where humans (task providers) provide the annotation statement and requirements, based on which another group of humans complete the annotation tasks. Such a scheme is typically costly and time consuming. The other scheme is via human-machine interaction, where similarly humans provide the annotation statement and requirements, but it is now the machines/models that complete the annotation tasks.

To enable LLMs to follow natural language instructions and complete real-world tasks, researchers have been exploring methods to instruction-tune LLMs. This is implemented by either finetuning the model on a wide range of tasks using human-annotated prompts and feedback (Ouyang et al., 2022), or supervised finetuning using public benchmarks and datasets augmented with manually or automatically generated instructions (Wang et al., 2022f). Among these methods, Self-instruct tuning (Wang et al., 2022e) is a simple and effective method of aligning LLMs to human intent, by learning from instruction-following data generated by SoTA LLMs. It turns out that the line of instruction-tuning research has produced effective means to improve zero-shot and few-shot generalization abilities of LLMs. Self-instruct leverages the in-context-learning ability of LLM. The pipeline is illustrated in Figure 5.7. Humans create a few examples (*i.e.* seed examples) as the con-

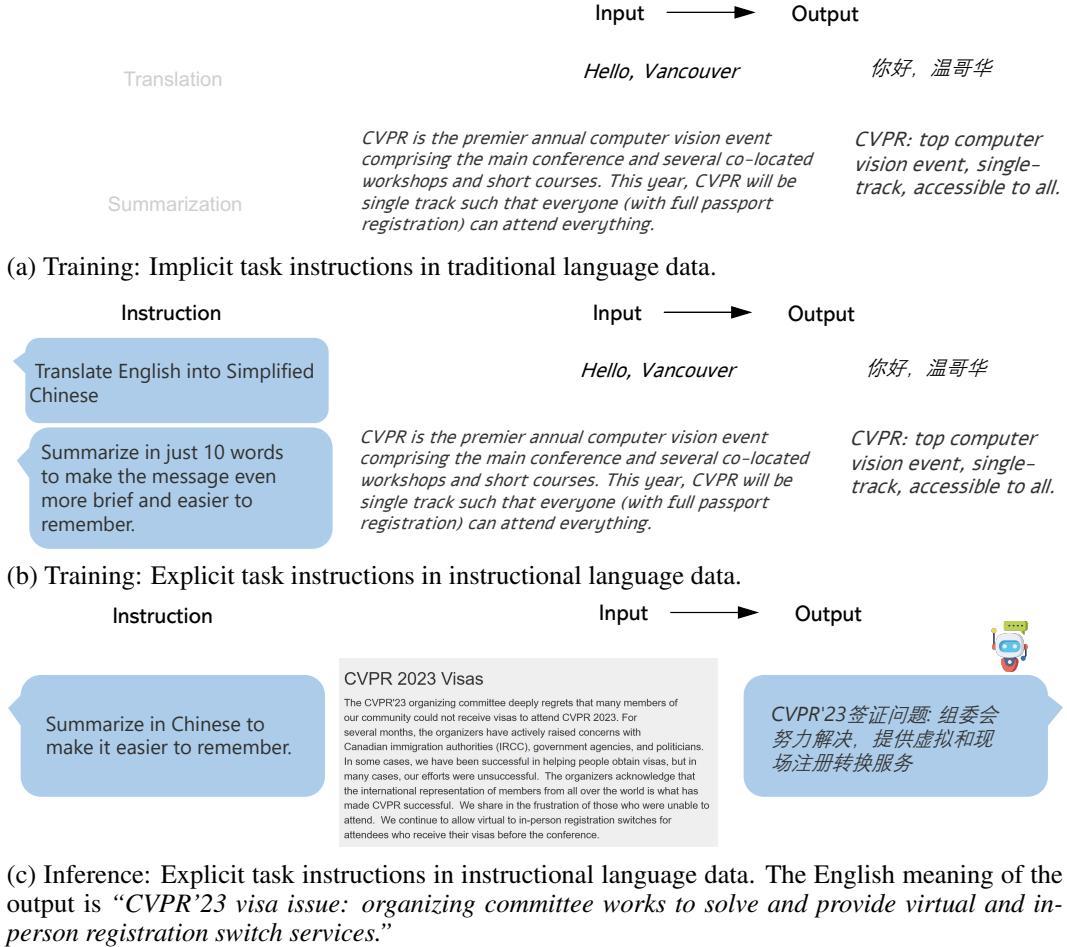


Figure 5.6: Examples of task instructions in traditional and instructional language data, respectively.

text, and ask LLM such as GPT-3 or GPT-4 to create more instructions and responses that follow the requirements stated in the prompt. The machine-generated instruction-following data can be further selected to construct with the prompt for in-context-learning in the next data generation iteration. The procedure iterates until a given number of samples are collected. Due to the relatively lower cost and higher response speed of API calls (compared with human annotations), self-instruct is becoming more favorable in the research community.

5.2.2 Self-Instruct and Open-Source LLMs

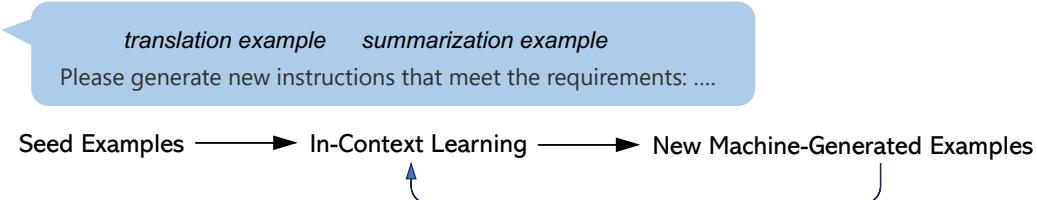


Figure 5.7: Illustration of the self-instruct pipeline (Wang et al., 2022e).

The open-source community has witnessed a surge of open LLMs. The success of ChatGPT ([OpenAI, 2022](#)) and GPT-4 ([OpenAI, 2023a](#)) offers tremendous opportunities to improve open-source LLMs using instruction-tuning. Figure 5.8 compares several open-source instruction-tuned LLMs. LLaMA ([Touvron et al., 2023](#)) is a series of open-sourced LLMs, which match the performance of

	LLaMA 	Alpaca 	Vicuna 	GPT4-Alpaca 	...	Tulu 
Data Source		GPT-3.5	ShareGPT (Human & GPT)	GPT-4 (text-only)	...	Mixed Data
Instruction-following Data (#Turns)	None	52K	500K (~150K conversions)	52K	...	

Figure 5.8: Model examples of the LLaMA family.

proprietary LLMs such as GPT-3. To teach LLaMA to follow instructions, Self-instruct tuning has been quickly adopted given its superior performance and low cost. For example, to name a few early attempts in this line of research, Stanford Alpaca (Taori et al., 2023) uses 52K instruction-following samples generated by GPT-3.5, while Vicuna (Vicuna, 2023) uses around 500K high-quality instruction-following samples (150K conversions) between user and GPT (ShareGPT, 2023). To advance the SoTA of instruction-tuning for LLMs, Peng et al. (2023a) uses GPT-4 as the teacher to generate the responses to the Alpaca instructions. Many follow-up works (Zhang et al., 2023i) improve the instruction-following data to enable the open LLMs with better alignment quality in chat. For a comprehensive review, we refer the readers to a recent paper (Wang et al., 2023k), where a LLM Tulu is trained on a mix of several high-quality instruction data, and comprehensive comparisons are conducted across multiple benchmarks.

Quick assessment of LLM chatbots. To study the quality of LLM Chatbots, we consider *Vicuna-Instructions-80*¹ (Vicuna, 2023), a dataset with 80 questions that baseline models (Touvron et al., 2023) find challenging. Besides generic instructions, the instructions fall into 8 categories, including knowledge, math, Fermi, counterfactual, roleplay, generic, coding, writing and common-sense. To quantitatively compare the performance, GPT-4 is used to rate the response from score 1 to 10 for any two given chatbots, then compute the relative score. Surprisingly, it turns out this evaluation metric is quite consistent across different settings. The open-source LLaMA family seems to perform closely to SoTA proprietary chatbots.

Further discussions. There are several important topics on LLMs that we have not covered in this chapter, but are worthwhile future exploring.

- *Data-centric AI.* We emphasize that the development of these open-source LLMs is data-centric (Mazumder et al., 2022), rather than model-centric, so that we hope the readers could align with this perspective when discussing the topic. As the training objectives and network architectures are becoming similar or even identical to GPT-like models, the key differential factor is data. For example, behaviors of the aforementioned LLMs are determined by the instruction tuning data.
- *False promise?* There is a debate on that the open LLMs could catch up with the proprietary LLMs is a false promise (Gudibande et al., 2023). To align the discussions, we argue that there are two distinctive abilities for LLMs: the instruction-following ability to know which task to perform, and massive knowledge storage to complete the task with high quality. Imitation models are good at the former, by mimicking ChatGPT’s style but perform poorly in terms of factuality in their responses. In Gudibande et al. (2023), the authors conclude that there exists a substantial capabilities gap between open and closed LLMs that, with current methods, can only be bridged using an unwieldy amount of imitation data or by using more capable base LLMs. They also advocate that the highest leverage action for improving open-source models is to tackle the difficult challenge of developing better base LLMs. However, unfortunately, the resources to train such base LLMs are only available in a few industry labs. It seems more

¹<https://github.com/lm-sys/FastChat/blob/main/fastchat/eval/table/question.jsonl>

promising for most academic research labs to explore the opportunities in alignment research with affordable resources, or explore the techniques to reduce the compute barriers.

- *Base LLMs.* Developing more capable or commercial usable LLMs is of great value. Besides LLaMA, the open-source community has developed variants of base LLMs such as LLaMA-2, OpenLLAMA (Geng and Liu, 2023), MPT (Team, 2023) and Falcon (Penedo et al., 2023), or released the training recipe (Computer, 2023).

5.3 Instruction-Tuned Large Multimodal Models

In this section, we illustrate how to build the minimum prototype of multimodal GPT-4 with open-source resources. Specially, we use LLaVA (Liu et al., 2023c) as the running example, a similar idea is also proposed in its con-current work MiniGPT-4 (Zhu et al., 2023a).

The research in the multimodal space has often been inspired by the latest advances in NLP in recent years. One successful recipe is to explore what would happen if the most intriguing and successful NLP ideas are borrowed for the vision-and-language community, for example, self-instruct. However, the unique challenge with self-instruct in multimodal research is that there is no strong multimodal teacher publicly available. Therefore, the research question becomes: how can we use language models such as language-only GPT-4 to create multimodal instruction following data.

Data Creation

Instead of directly feeding images into OpenAI GPT-4, we use their symbolic sequence representations shown in Figure 5.9 (a). In LLaVA, both captions and bounding boxes are considered, due to the following reasons: (i) it is empirically found that GPT-4 can understand both well, in contrast to the poor performance of ChatGPT in understanding bounding box coordinates. (ii) They are often complementary to each other and hence can represent the image as informative as possible.

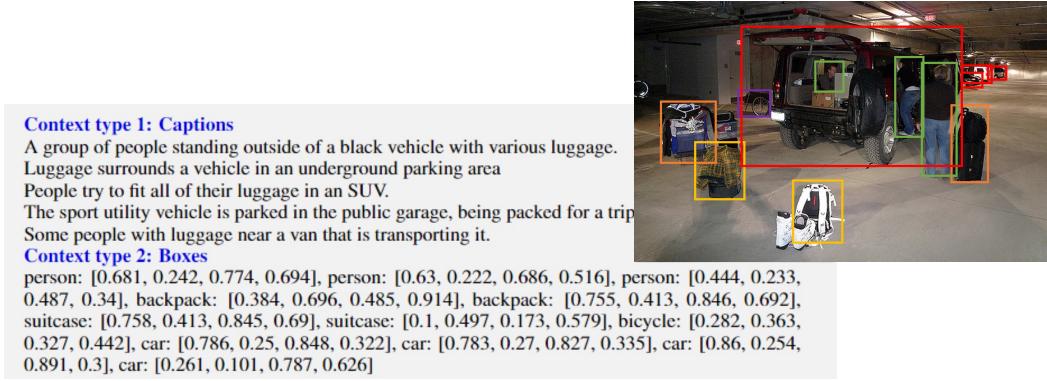
As shown in Figure 5.9 (b), three types of instruction-following data are considered: (i) multi-turn conversations so that users can chat with the model; (ii) detailed description so that long-form responses can be generated from the model; and (iii) complex reasoning, which is more about the implication of the image, rather than the image content. For example, “*what challenge do these people face?*”, which requires to first recognize that the image is about a SUV in the parking area, and there are quite a few luggage placed on the ground, and then to infer that the challenge is how the luggage can be packed into the SUV due to the tight space of the trunk. In total, 158K samples are collected over three types. To summarize, the spirit is that whatever tasks one wants the model to perform in the serving stage, it is important to create the corresponding instruction-following data for training.

Network Architecture and Training

As illustrated in Figure 5.10, the network architecture of LLaVA is an instantiation of the general image-to-text generative model framework introduced in Figure 5.1 of Section 5.1. Specifically, LLaVA connects the pre-trained CLIP ViT-L/14 visual encoder (Radford et al., 2021) and large language model Vicuna (Vicuna, 2023), via a simple projection matrix (*i.e.*, the linear projection layer). A two-stage instruction-tuning procedure is adopted to train the model. (i) *Stage 1: pre-training for feature alignment.* Only the projection matrix is updated, based on a subset of CC3M (Changpinyo et al., 2021). (ii) *Stage 2: finetuning end-to-end.* Both the projection matrix and LLM are updated on the proposed multimodal instruction-following data for daily user-oriented applications.

Performance

Visual chat: towards building multimodal GPT-4 level chatbot. LLaVA is finetuned on the generated multimodal instruction-following data, which contains a diverse set of task instructions and responses for daily user-oriented applications. It is empirically observed that finetuning the linear projection layer only is sufficient for the chat demo/scenarios, though it requires longer training time. To evaluate the model performance, an evaluation dataset named LLaVA-Bench is constructed, with two subsets: (i) LLaVA-Bench (COCO): 30 unseen COCO images with 90 new language-image instructions, (ii) LLaVA-Bench (In-the-Wild): 24 images with 60 questions. Each image can be associated with three types of instructions: conversation, detailed description and complex reasoning.



(a) The sequence representation of the image data.

Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV).
Question: Where is the vehicle parked?
Answer: The vehicle is parked in an underground parking area, likely in a public garage.
Question: What are the people in the image doing?
Answer: The people are trying to fit all of their luggage into the SUV, likely preparing for a trip.

Response type 2: detailed description
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip.
In the parking area, various luggage items are scattered around the vehicle. There are two backpacks, one located near the left rear wheel and the other closer to the right side of the vehicle. Additionally, there are two suitcases, one on the right side of the car and another further away near the center of the parking area. A bicycle can also be seen on the left side of the vehicle.
Other cars are parked around the main SUV, with one car positioned behind it and slightly to the left, another behind and slightly to the right, and the third car further behind on the right side.

Response type 3: complex reasoning
Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings to accommodate. They might have to strategize and arrange the luggage efficiently to ensure that everything fits properly into the vehicle. Additionally, they need to consider the comfort of the passengers and visibility while driving, so the placement of the luggage must not obstruct the driver's view or make the passengers uncomfortable during the trip.

(b) Three types of instruction-following data for the given image.

Figure 5.9: Examples of multimodal instruction-following data. Image credit: Liu et al. (2023c).

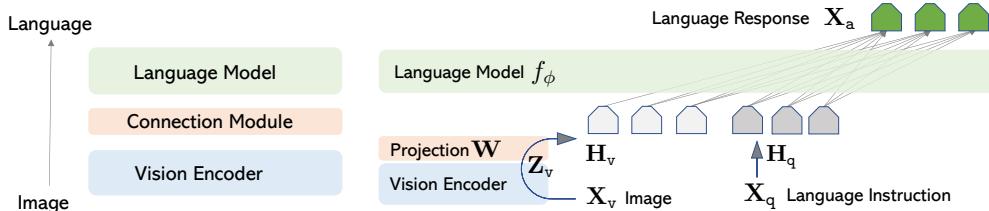


Figure 5.10: Network architecture: Left: General LMM; Right: LLaVA. Image credit: Liu et al. (2023c).

The ground-truth answers are collected by manually re-writing GPT-4 output. We test LLaVA and use language-only GPT-4 to rate their responses from score 1 to 10. Overall, LLaVA achieves 85.1% relative score compared with ground-truth on LLaVA-Bench (COCO), and 73.5% on LLaVA-Bench (In-the-Wild). On the latter dataset, Google Bard (July 19, 2023) and Microsoft BingChat (June 29, 2023) achieves 77.8% and 71.5%, respectively. It indicates the effectiveness of the proposed self-instruct method in multimodal settings. One examples is shown in Table 5.1.

Science QA: New SoTA with the synergy of LLaVA with GPT-4. LLaVA is finetuned on a multimodal reasoning dataset in the science domain (Lu et al., 2022b). LLaVA alone achieves 90.92% in accuracy. We further explores with language-only GPT-4 as the judge, to predict the final answer based on its own previous answers and the LLaVA answers. This “GPT-4 as judge” scheme yields a new SoTA of 92.53%.

OCR in the wild: An emerging property. LLaVA has never been explicitly trained on OCR data, *i.e.* images that contains scene text that is described in the corresponding caption. Surprisingly, the model shows strong zero-shot OCR task transfer ability in the wild.

5.4 Advanced Topics

The history of recent instruction-tuned LMMs are illustrated in Figure 5.11 (a). Due to the popularity of ChatGPT and GPT-4, instruction-tuned LMM appears as an emerging line of research in the past three months after GPT-4 was proposed. Alpaca (Taori et al., 2023) and Vicuna (Vicuna, 2023) were proposed to make LLaMA more instruction-following in the language domain in March. In two weeks, MiniGPT-4 (Zhu et al., 2023a) and LLaVA (Liu et al., 2023c) were proposed to make Vicuna to see and chat about the visual world. In ten days, LLaMA-Adapter v2 (Gao et al., 2023b) and mPlug-OWL (Ye et al., 2023b) started to compare performance with MiniGPT-4/LLaVA, indicating the beginning of model evolution. The data points in April are relatively sparse. In May, a large number of LMM papers appeared on arXiv, which improve this line of research from many different aspects. The momentum is till going in June.

It is easy to lose track of all the recent papers for the readers, so as well in our literature review. To better organize the literature, we group them based on specific research topics, shown in Figure 5.11 (b). The early LMMs with billions of parameters include GPT-4 (OpenAI, 2023a), Flamingo (Alayrac et al., 2022), PaLM-E (Driess et al., 2023) and KOSMOS-1 (Huang et al., 2023b). In contrast to these proprietary LMMs, LLaVA and MiniGPT-4 open the opportunities to build LMMs with open-source resource. We will discuss several topics as below, in addition to the extensions of RLHF (Gunjal et al., 2023), dense prediction (Wang et al., 2023h; Zang et al., 2023; Chen et al., 2023d), video (Zhang et al., 2023f; Luo et al., 2023c; Li et al., 2023i), image generation (Koh et al., 2023) and embodied agent (Mu et al., 2023).

More Modalities (Beyond VL)

While LMM extends LLM by adding the vision modality, it is natural to further extend the framework to include more modalities beyond vision and language. Following this spirit, several attempts have been made, including ChatBridge (Zhao et al., 2023e), PandaGPT (Su et al., 2023), SpeechGPT (Zhang et al., 2023d) and X-LLM (Chen et al., 2023c). PandaGPT leverages ImageBind to add more modalities into LMMs. The ImageBind model (Girdhar et al., 2023) learns a single, shared representation space for text, image/video, audio and sensors that record depth (3D), thermal (infrared radiation), or inertial measurement units (IMU), which calculate motion and position. ImageBind provides a holistic understanding of the visual world that connects objects in a photo with how they will sound, their 3D shape, how warm or cold they are, and how they move. By training a projection layer for one modality in LMM, the model can zero-shot transfer to infer over other modalities, thanks to the shared multimodal embedding space. Another representative model is SpeechGPT, where language and speech modalities are enabled for both inputs and outputs. Despite of rich model variations, the idea to connect diverse modalities is similar to LMM that adds images into LLMs. NExT-GPT (Wu et al., 2023c) connects an LLM with multimodal adaptors and different diffusion decoders, enabling NExT-GPT to perceive inputs and generate outputs in arbitrary combinations of text, images, videos, and audio. The LMM framework has also been successfully extended to speech (Zhao et al., 2023c), 3D (Wang et al., 2023l; Hong et al., 2023), and point cloud (Xu et al., 2023c).

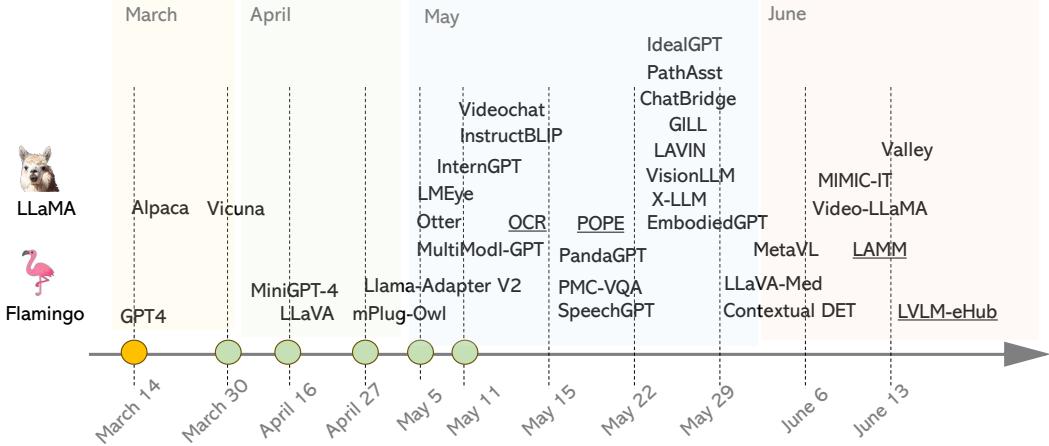
Improving Visual Instruction Data Quantity and Quality

Given the convergence of model architectures to GPT-like network, the performance of LMM is primarily determined by its training data. Therefore, it is crucial to improve the quantity and quality of visual instruction tuning data. SVIT (Zhao et al., 2023a) follows the same data generation pipeline

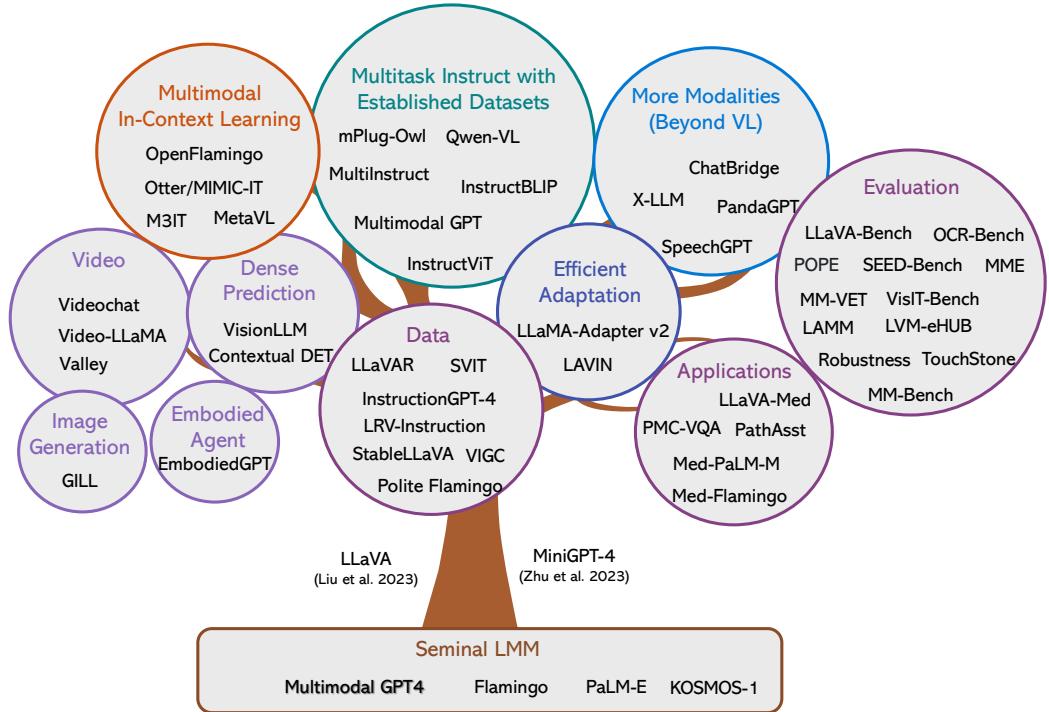
Visual input example, Chicken Nugget Map:

	<p>Sometimes I just look at pictures of the earth from space and I marvel at how beautiful it all is.</p> 
User LLaVA	<p>Can you explain this meme in detail?</p> <p>The meme in the image is a creative and humorous take on food, with a focus on chicken nuggets, as the center of the universe. The meme begins with the phrase "Sometimes I just look at pictures of the Earth from space and I marvel at how beautiful it all is..." The main image features a tray of cooked chicken nuggets, with the chicken pieces positioned to resemble the continents and islands on a world map. The meme playfully suggests that the chicken nuggets represent the Earth, and the various locations depicted in the photo are actually chicken nugget versions of different places. The punchline of the meme appears in smaller text at the bottom, reading "I mean, it's not the real Earth, but how beautiful it is all is." This lighthearted and entertaining meme combines food and humor, using the image of chicken nuggets to create a funny and imaginative representation of the world.</p>
User GPT-4	<p>Can you explain this meme?</p> <p>This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets. The text of the meme suggests that the image below is a beautiful picture of the earth from space. However, the image is actually of chicken nuggets arranged to vaguely resemble a map of the world. The humor in this meme comes from the unexpected juxtaposition of the text and the image. The text sets up an expectation of a majestic image of the earth, but the image is actually something mundane and silly.</p>
User Bing Chat	<p>Can you explain this meme in detail?</p> <p>Sure, I can try to explain this meme in detail. This is a **chicken nugget meme**, which is a popular online meme that typically features a picture of a chicken nugget with a caption that is funny or relatable. The meme has been popular since the early 2000s, and continues to be one of the most popular memes today. This particular meme is based on a contrast between the text and the image. The text above the image reads "Sometimes I just look at pictures of the earth from space and I marvel at how beautiful it all is." This is a sentiment that many people can relate to, as looking at the earth from space can inspire awe and wonder. However, the image below the text shows a baking tray with chicken nuggets arranged in the shape of the world map. This is a humorous twist, as chicken nuggets are not usually associated with beauty or marvel, but rather with fast food or junk food. The image also implies that the person who made this meme loves chicken nuggets so much that they see them everywhere, even when looking at pictures of the earth from space. The meme is meant to be funny and absurd, and to poke fun at the person's obsession with chicken nuggets. I hope this explanation helps you understand this meme better.</p>
User Bard	<p>Can you explain this meme in detail?</p> <p>Sure. The meme you sent me is a humorous take on the idea of looking at pictures of the earth from space and marveling at its beauty. The text in the meme says "Sometimes I just look at pictures of the earth from space and I marvel at how beautiful it all is." However, the image shows a map of the world made out of chicken nuggets. This juxtaposition of the text and the image is what creates the humor in the meme. The text suggests that the speaker is looking at something beautiful and awe-inspiring, but the image shows something that is more mundane and humorous. The meme also plays on the fact that chicken nuggets are often seen as a junk food. By depicting a map of the world made out of chicken nuggets, the meme suggests that even something as mundane as chicken nuggets can be beautiful, or at least, more beautiful than we might think. Overall, the meme is a humorous way to poke fun at the idea of finding beauty in the mundane. It also shows how our perception of beauty can be influenced by our expectations. Here are some other interpretations of the meme: (1) The meme could be seen as a commentary on the way we consume images of the earth from space. We often see these images as representations of beauty and perfection, but the meme reminds us that the earth is also a messy and imperfect place. (2) The meme could also be seen as a critique of the fast food industry. The use of chicken nuggets to represent the world suggests that we are increasingly seeing the world through the lens of consumerism. Ultimately, the meaning of the meme is up to the interpretation of the viewer. However, it is clear that the meme is a humorous and thought-provoking way to explore the idea of beauty.</p>
User MiniGPT-4	<p>Can you explain this meme in detail?</p> <p>This meme is a picture of a pan full of fried chicken. The chicken is cooked and appears to be ready to be served. The caption reads, "Sometimes it's just like looking at pictures from space and marveling at how beautiful it all is." This meme is likely meant to be humorous and playful, as it compares the beauty of space to the simple pleasure of eating fried chicken. It is likely meant to be a lighthearted and funny way to express appreciation for both the wonders of space and the simple pleasures of food.</p>
User BLIP-2	<p>Can you explain this meme in detail?</p> <p>sometimes i just look at pictures of the earth from space and marvel how beautiful it is</p>
User OpenFlamingo	<p>Can you explain this meme in detail?</p> <p>It's a picture of a chicken nugget on the International Space Station.</p>

Table 5.1: Example prompt comparing LLaVA, GPT-4, BingChat, Bard, MiniGPT-4, BLIP-2, and OpenFlamingo's visual reasoning capabilities in understanding the humor. LLaVA and GPT-4 both explain the meme and its humor, while GPT-4 produces a more concise answer. Table credit: Liu et al. (2023c).



(a) The surge of papers on LMMs from March 14, 2023 to June 19, 2023. Those with an underline indicate benchmarks, otherwise indicate models.



(b) Summary and categorization of papers on LMMs.

Figure 5.11: Review and summary for the emerged LMM literature. Due to space constraints, some methods are not displayed visually, but we aim to address them in the accompanying text.

as in LLaVA, but further includes region description to prompt GPT-4, in addition to the caption and box data as shown in Figure 5.9 (a). The data is scaled up to 3.2 million, which is 20 times larger than the data used in LLaVA.

Unlike existing studies that primarily focus on positive instruction samples, LRV-Instruction (Liu et al., 2023a) includes both positive and negative instructions for more robust instruction-tuning. Other examples along this line include LLavar (Zhang et al., 2023o) that adds OCR-related instruction-tuning data for text-rich image understanding, and StableLLaVA (Li et al., 2023o) that considers model-synthesized images for image-dialogue data. Polite Flamingo (Chen et al., 2023b) trains LLM to re-write the instruct data. Instead of leveraging GPT-4 for data generation, VIGC (Wang et al., 2023a) considers to utilize LMM to generate instruction-tuning data and progres-

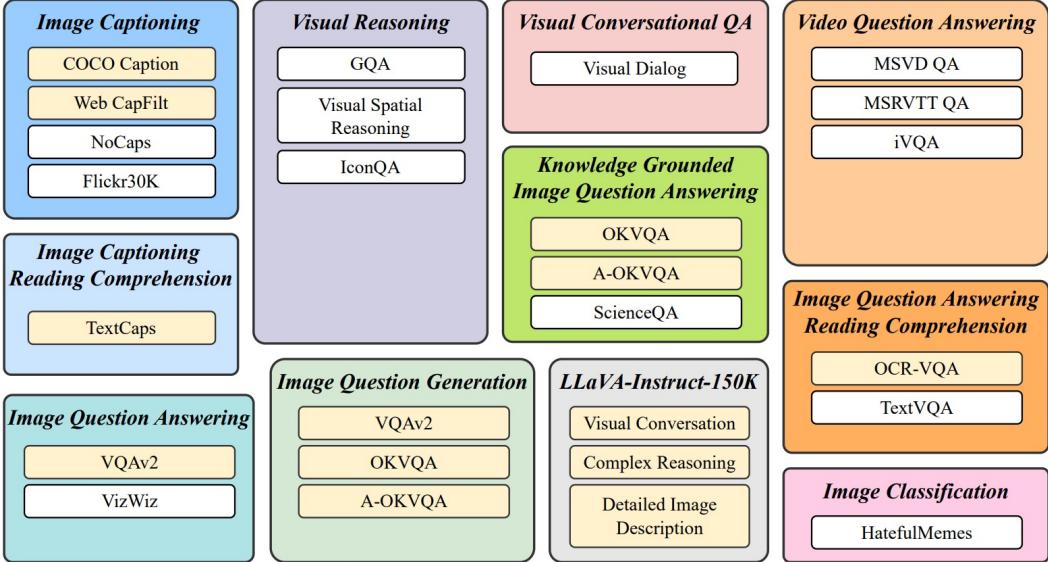


Figure 5.12: The vision-language tasks covered in InstructBLIP. Image credit: [Dai et al. \(2023b\)](#).

sively enhance its quality on-the-fly. Similar to the “less is more” observation in LIMA ([Zhou et al., 2023a](#)) from the NLP domain, InstructionGPT-4 shows that the quality of the instruction-tuning data is more important than its quantity, where they finetune a better version of MiniGPT-4 with 200 high-quality samples (6%), selected from the 3500 samples used in the original MiniGPT-4.

Multitask Instruct with Established Academic Datasets/Tasks

As discussed earlier in Section 5.2, instruction tuning in the language domains is implemented in two different ways: finetuning the model on a wide range of tasks using human-annotated prompts and feedback ([Ouyang et al., 2022](#)), or supervised finetuning using public benchmarks and datasets augmented with manually or automatically generated instructions ([Wang et al., 2022f](#)). The former is good at user-oriented daily life tasks, and the latter is good at achieving decent performance on established benchmarks. LLaVA and MiniGPT-4 fall into the former class. Several other works either target for the latter class or combine both classes, including MultiInstruct ([Xu et al., 2022b](#)), mPlug-OWL ([Ye et al., 2023b](#)), InstructBLIP ([Dai et al., 2023b](#)), Multimodal-GPT ([Gong et al., 2023](#)), Instruction-ViT ([Xiao et al., 2023](#)) and Qwen-VL ([Bai et al., 2023a](#)).

For example, MultiInstruct is an early attempt before open-source LLaMA for instruction tuning with multimodal datasets. InstructBLIP is a recent work that combines chat and benchmark instruction-following data. As shown in Figure 5.12, InstructBLIP transforms 26 publicly available datasets, covering a wide variety of tasks and capabilities, into instruction tuning format. Trained on 13 held-in datasets, InstructBLIP attains SoTA zero-shot performance across all 13 held-out datasets, substantially outperforming BLIP-2 and larger Flamingo models. Qwen-VL scales up both image-text pair data for pre-traning and academic datasets for multi-task pre-traning, and achieve excellent performance on many tasks.

Multimodal In-Context-Learning

Similar to the behavior of LLMs, which can address a language task by processing examples of the task in their text prompt, multimodal in-context-learning refers to a visual and text interface that can steer the model towards solving a multimodal task. Given a few example pairs of visual inputs and expected text responses composed in the multimodal prompt, the model can be queried with a question about a new image or video, and then generate an answer. The direction to extend in-context-learning from language to multi-modalities has been explored, including OpenFlamingo ([Awadalla et al., 2023](#)), Otter ([Li et al., 2023d](#)), M³IT ([Li et al., 2023j](#)), MetaVL ([Monajatipoor et al., 2023](#)) and Sparkles ([Huang et al., 2023d](#)).

OpenFlamingo (Awadalla et al., 2023) is an open source version of DeepMind’s Flamingo model, trained on Multimodal C4 dataset (Zhu et al., 2023b), which is a billions-scale corpus of interleaved image-text data. To explicitly enhance the multimodal in-context-learning ability of LMMs, MIMIC-IT (Li et al., 2023c) dataset is constructed, which is 2.4M multimodal instruction instances with in-context examples. By tuning OpenFlamingo on MIMIC-IT, a new model Otter is obtained with a stronger instruction-following ability. Using two image-text pairs as the context, Otter learns the concise answer style demonstrated by the examples, otherwise a tedious response is generated.

Parameter-Efficient Training

While finetuning very large models often leads to high performance, it is prohibitively expensive; For example, regular 16-bit finetuning of a LLaMA-65B model (Touvron et al., 2023) requires more than 780 GB of GPU memory (Dettmers et al., 2023). Therefore, it is critical to reduce the memory footprint of LLMs/LMMs, especially when it comes to improve the accessibility of large models to a wider community.

Parameter-efficient training is an effective approach for LMM adaptation. It freezes most of the model parameters, and only allows a fraction of trainable parameters to update with domain-specific data. For example, LLaMA Adapter v2 (Gao et al., 2023b) and LAVIN (Luo et al., 2023a) only have 14M and 3.8M trainable parameters, compared with 7B/13B LLM parameters, respectively. Another efficient training method is quantization. The recent QLoRA (Dettmers et al., 2023) finetunes 65B LLaMA for 24 hours on a single GPU, achieving 99.3% of the performance level of ChatGPT. Since instruction tuning typically involves a small amount of data, it makes parameter-efficient training or model quantization the practical approach, especially when with limited GPU resources. Both LoRA (Hu et al., 2021) and QLoRA are supported in LLava codebase to allow LMM training with less GPUs. It is empirically shown in Lu et al. (2023d) that LoRA/QLoRA can achieve similar performance with full-modal tuning when scaling LLava to 33B and 65B, when training with around 150K instruct data and evaluating with LLava-Bench.

Benchmarks

While LMMs have shown excellent visual recognition and reasoning in an open-set manner with free-form text across many scenarios, the evaluation of LMMs is becoming an urgent and challenging problem. Several related benchmarks have been developed to evaluate various aspects of LMMs, ranging from their specific abilities including OCR (Liu et al., 2023k), hallucination (POPE (Li et al., 2023l) and HaELM (Wang et al., 2023d)) and adversarial robustness (Zhao et al., 2023d), to comprehensive evaluation such as LAMM (Yin et al., 2023), LVLM-eHub (Xu et al., 2023b). We summarize the LMM evaluation benchmarks in Table 5.2. Among them, LLava-Bench is the first attempt to designed open-world visual chat benchmark specifically for LMM. Recently, early multimodal experiments have been conducted to compare open-source LMM with commercial ones such as BingChat and Bard and LLava-Bench (Liu et al., 2023c) and LVLM-eHub (Shao et al., 2023).

It is surprising that LMMs shows strong zero-shot OCR performance in the wild, without explicitly training on text recognition data. To shed light on the hidden mystery of OCR in LMMs, a comprehensive empirical study is conducted in Liu et al. (2023k) to compare open-source LMMs on 24 academic text recognition datasets, shown in Figure 5.13. Three observations are highlighted: (i) LLava consistently outperforms MiniGPT-4 on 21 out of 24 datasets, despite that the training data in LLava is an order of magnitude smaller. (ii) Training with significantly more training data leads to higher OCR performance, as demonstrated by BLIP2 (Li et al., 2023h) and mPLUG-Owl. (iii) In most cases, supervised SoTA results significantly outperform zero-shot LMM. However, it is worth noting that in the WordArt dataset (Xie et al., 2022a), which primarily features challenging artistic text, BLIP2 surpasses supervised SoTA. This reveals the potential of LMM in recognizing more complex text types.

Applications

The success of ChatGPT/GPT-4 in the general domain has inspired the interests in building assistants in the vertical domains such as medicine, gaming and education. Such domain-specific assistants can have the several advantages over the general domain counterpart: (i) training with high-quality domain-specific data makes the assistants more helpful; (ii) the model size can be smaller, with

Benchmark	Capability to Evaluate	Statistics	Metric
LLaVA-Bench (Liu et al., 2023c)	Multi-turn QA, detailed description, reasoning	Two subsets: 90 samples on COCO and 60 samples on In-the-Wild	Relative score via GPT-4 evaluation
OCR-Bench (Liu et al., 2023k)	Zero-shot OCR	A suite of 23 OCR-related academic tasks	Accuracy
MMBench (Liu et al., 2023j)	Perception (coarse, fine-grained single-instance and cross-instance) and Reasoning (attribute, relation, logic)	2974 multiple-choice samples in 20 ability dimensions	Circular evaluation via ChatGPT answer extraction
M3Exam (Zhang et al., 2023i)	Multilingual, multimodal, and multi-level assessment	12,317 questions in 9 languages, with 2,816 questions involving images	Accuracy on multiple-choice questions
MME (Fu et al., 2023)	Perception and Cognition	14 tasks	Accuracy on “yes” or “no”
LAMM (Yin et al., 2023)	Various 2D/3D vision tasks	9 image task with 62K samples, and 3 point cloud tasks with 12K samples	Traditional CV task metrics
LVLM-eHub (Xu et al., 2023b)	six multimodal capabilities such as VQA and embodied AI	47 standard text-related visual benchmarks	CIDEr and accuracy; Arena with human judgment
SEED-Bench (Li et al., 2023b)	Comprehension of both the image and video modality	19K multiple choice questions in 12 dimensions	Accuracy on multiple-choice questions
VisIT-Bench (Bitton et al., 2023)	Real-life vision-language instructions	592 samples in 70 tasks	Elo, matches
MM-VET (Yu et al., 2023d)	Integrated capabilities in recognition, OCR, spatial, knowledge, math, language	200 samples	GPT-4 evaluation
TouchStone (Bai et al., 2023b)	Five abilities: basic description, visual recognition, visual comprehension, visual storytelling, and multi-image analysis	908 dialogues in 27 tasks	GPT-4 evaluation
SciGraphQA (Li and Tajbakhsh, 2023)	Scientific graph question-answering	3K test samples	CIDEr, BLEU-4, and ROUGE

Table 5.2: Comparisons of recently proposed LMM evaluation benchmarks.

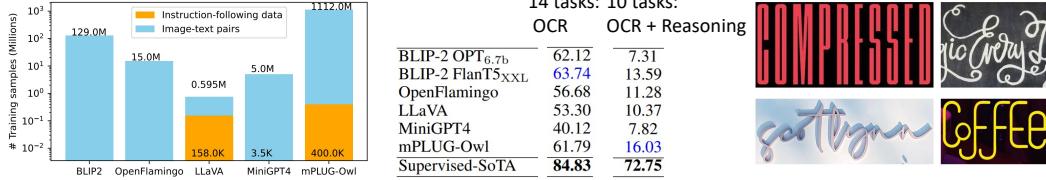


Figure 5.13: Zero-shot OCR performance of LMMs on 24 datasets. Image credit: Liu et al. (2023k).

lower severing cost; and (iii) the sensitive user prompt data can be maintained internally by serving the model locally, to avoid privacy issue.

To improve text recognition ability of LMM, OCR-specific models have been developed, including BLIVA (Hu et al., 2023), LLaVAR (Zhang et al., 2023o), mPlug-DocWL (Ye et al., 2023a). LMMs have been recently explored in the biomedical domain (Sun et al., 2023c; Zhang et al., 2023m; Li et al., 2023e), where conversational generative AI has demonstrated remarkable promise for empowering biomedical practitioners. LLaVA-Med (Li et al., 2023e) is a cost-efficient approach for training a vision-language conversational assistant that can answer open-ended research questions about biomedical images. The key idea is to leverage a large-scale, broad-coverage biomedical figure-caption dataset extracted from PubMed Central, use GPT-4 to self-instruct open-ended instruction-following data from the captions, and then finetune a large general-domain vision-language model LLaVA using a novel curriculum learning method. Specifically, the model first learns to align biomedical vocabulary using the image-caption pairs as is, then learns open-ended conversational semantics using GPT-4 generated instruction-following data, broadly mimicking how a layperson gradually acquires biomedical knowledge. In Figure 5.14, we provide examples of the biomed visual conversations with different chatbots. LLaVA-Med precisely answers the questions requiring biomedical knowledge, while LLaVA behaves like a layperson, that hallucinates based on common-sense. LLaVA-Med has inspired several generalist biomedical AI models, including Google Med-

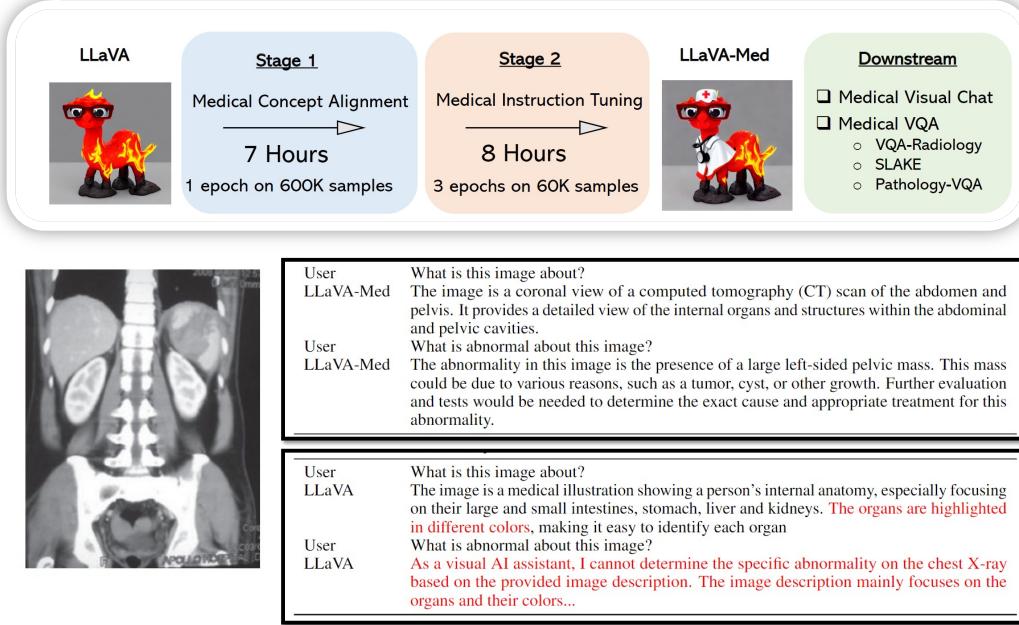


Figure 5.14: Application of LMMs to biomedical images. Top: Domain adaptation from LLaVA to LLaVA-Med. Bottom: The chat behaviors of two chatbots. Image credit: [Li et al. \(2023e\)](#).

PaLM-M ([Tu et al., 2023](#)), Stanford Med-Flamingo ([Moor et al., 2023](#)) and radiology generalist ([Wu et al., 2023b](#)).

5.5 How Close We Are To OpenAI Multimodal GPT-4?

With all the works mentioned above, are we close to (or, even surpassing) OpenAI Multimodal GPT-4? It is encouraging to see that the open-source community has quickly developed a variety of models and prototypes for various new capabilities. For example, LLaVA/Mini-GPT4 paves the way towards building multimodal chatbots, with some examples that reproduce the results in OpenAI GPT-4 technique report; CM3leon ([Yu and et al, 2023](#)), Emu ([Sun et al., 2023a](#)), GILL ([Koh et al., 2023](#)) extends LMMs for end-to-end image generation, to the best of our knowledge, this is a capability that the current GPT-4 does not exhibit. From the perspective of enabling new capabilities with the minimum prototypes, the open-source community seems close to OpenAI Multimodal GPT-4, by exploring the baby steps towards building the general-purpose multimodal assistant.

However, there is still a clear large gap in terms of scaling a given capability, *e.g.*, for the visual reasoning capability that we have observed in LLaVA. There are two more visual examples from OpenAI technical report, to correctly answer the questions, it requires models to understand multiple high-resolution images and long sequence text depicted in the image, as well as responding with domain knowledge. It requires much more compute and more powerful language models, which are not available to most people.

In summary, we have presented the background and strong capabilities of LMM, reviewed instruction tuning in LLMs, and showed how to build a prototype such as LLaVA and MiniGPT-4 using open-source resources. We also summarized the most recent papers emerged on this line of research to help those who are interested to gain the momentum to start the journey of LMM research. To discuss the next steps to work on as a community, one sustainable suggestion can be that those with resources can continue focusing on the scaling success and study new emerging properties, while others focus on prototypes for new functionalities and evaluation, as well as developing techniques to reduce the computational barriers and thus allow easier accessibility to large models.

Chapter 6

Multimodal Agents: Chaining Tools with LLM



Large Language Models (LLMs) (Chowdhery et al., 2022; OpenAI, 2023a) have shown intriguing properties generalizing to user prompts in various domains, and rapidly adapting to new scenarios, using in-context learning with a few examples. Inspired by such strong capabilities, researchers are now exploring a new modeling paradigm that shifts from standalone models for solving finite, pre-defined problems, into synergistically chaining multiple tools or experts with LLMs to solve complicated, open problems. Unlike what has been introduced in Chapter 5, such a system can be built without any training involved, just by using a few demonstration examples to teach the LLM to generate proper calling to existing tools.

In this chapter, we review the fast-evolving literature on chaining different multimodal experts with LLMs to solve complicated multimodal understanding problems, referred to as *multimodal agents*. We start with an overview on the evolution of this modeling paradigm in Section 6.1, highlighting the differences between traditional approaches and the new modeling paradigm of chaining tools with LLM. Section 6.2 gives a general overview of multimodal agents. Pivoting on an exemplary multimodal agent MM-REACT (Yang* et al., 2023), Section 6.3 comprehensively reviews how to build a multimodal agent, its emerging capabilities in multimodal understanding, and how it can be easily extended to incorporate the latest and strongest LLM and potentially millions of tools. Finally, in Section 6.4, we end the chapter with discussions on advanced topics, such as how to improve/evaluate multimodal agents, the diverse applications powered by multimodal agents.

6.1 Overview

We first revisit the evolution of modeling paradigms, from task-specific models to the most recent large multimodal models, which all require data curation and model training. We then introduce the new modeling paradigm of chaining tools with LLM, which may not require any training, but instead directly takes advantage of a pre-trained LLM and existing tools that are widely available through open-source platforms or APIs.

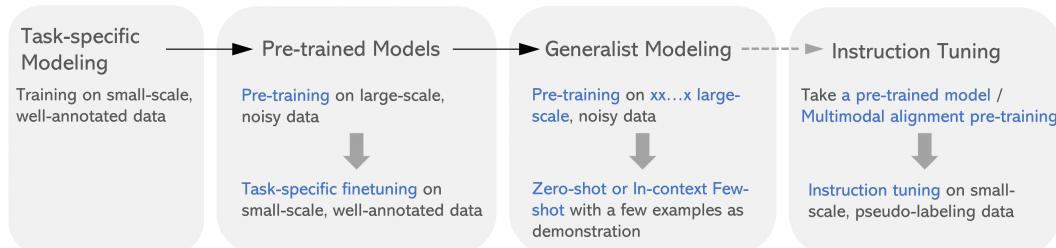


Figure 6.1: Evolution of modeling paradigm.

Evolution of modeling paradigm. As summarized in Figure 6.1, we are witnessing the transition from task-specific models towards general-purpose assistants across language, vision, and multi-modal research.

We started with **task-specific models** that are trained on small-scale well-annotated data. This results in dedicated models (Anderson et al., 2018; Li et al., 2019a; Yu et al., 2019) for each task or even each dataset.

We then transitioned to the phase of **pre-trained models**, with the pretrain-then-finetune paradigm widely adopted across both NLP and vision-language (VL) research. During pre-training, the model can take advantages of large-scale, web-crawled noisy data, for example, millions to billions of image-text pairs (Chen et al., 2020d; Wang et al., 2022a), or billions of text tokens (Devlin et al., 2019; Liu et al., 2019). However, it is still mostly task-specific finetuned, requiring similarly small-scale, well-annotated data as the ones used in training task-specific models. This paradigm has led to many well-known models, such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) in NLP, and UNITER (Chen et al., 2020d), OSCAR (Li et al., 2020b) in VL. These early VL foundation models were considered to be large-scale (trained with 10M image-text pairs), but may be of intermediate or even small size in today’s view (billions of pairs).

Nowadays, we are entering a new era of **generalist modeling**, where the pre-training has been further scaled up to trillions of text tokens (Gao et al., 2023b). For downstream adaptation, these generalist models have shown strong performance with in-context few-shot learning on a few demonstration examples, or even zero-shot evaluation. These models are what we now refer as large language/multimodal models, including the GPT family (OpenAI, 2022, 2023a), PaLM family (Chowdhery et al., 2022; Driess et al., 2023), LLaMa (Touvron et al., 2023), Flamingo (Alayrac et al., 2022).

Based on the generalist models, the pipeline of building **instruction-following models** covered in Chapter 5, similarly follows the pretrain-then-finetune paradigm. For example, Alpaca (Taori et al., 2023), is built on top of the pre-trained LLaMa (Touvron et al., 2023), then finetuned on a smaller-scale instruction tuning dataset. Similarly, for instruction-following VL models (*e.g.* LLaVA (Li et al., 2023e)), an additional stage of image-text alignment pre-training is introduced to align the visual representations to the frozen LLM first, followed by visual instruction tuning.

New modeling paradigm: chaining tools with LLM. LLMs (Brown et al., 2020; Chowdhery et al., 2022; OpenAI, 2023a) have demonstrated exceptional abilities to tackle new tasks with only a few examples or textual instructions, showing the promise of serving as general-purpose foundations for many applications. Despite being versatile and impressive, they encounter challenges with the basic functionalities, such as mathematical reasoning and information retrieval. Furthermore, a fundamental limitation of not only LLMs but also other large-scale models nowadays, is that they only represent the world described by their training data, which will inevitably become outdated over time. Regularly re-training the model with the latest information is simply not feasible.

Meanwhile, many tasks with real-world impact cannot be readily tackled by LLMs alone. For example, accessing up-to-date information and performing computations, can be done via existing tools (*e.g.*, search engine or calculator). Hence, recent research in language modeling has explored a new modeling paradigm by supplementing LLMs with external NLP tools (Nakano et al., 2021; Huang et al., 2022b; Ahn et al., 2022), including calculators, search engines, translation systems, calendars, or even API calls on other models.

The above studies mainly focus on a single modality, *i.e.*, language, in which the output of the tools are in text format, thereby can naturally be fed into LLMs as additional knowledge. However, we live in a multimodal world and a truly intelligent agent should be able to perform advanced multimodal reasoning and actions. How to enable LLMs with perception of multimodal signals via tool using, is the focus of the remaining part of this chapter.

6.2 Multimodal Agent

There are several representative works on building multimodal agent with tool use of vision experts, including VISPROG (Gupta and Kembhavi, 2022b), Visual ChatGPT (Wu et al., 2023a) and MM-ReAct (Yang* et al., 2023). VISPROG is the very first work on using programming language to chain different vision tools with a LLM. Visual ChatGPT enables dialogue-based image editing by

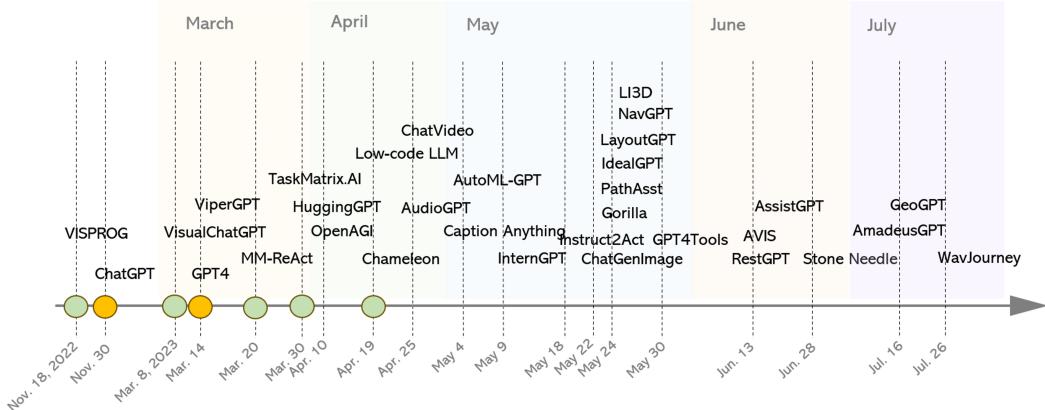


Figure 6.2: The surge of multimodal agents from November 18, 2022 to July 26th, 2023.

Model	LLM	Tools	Tool Size	Tool Execution	Multimodal
ART (Paranjape et al., 2023)	GPT-3	🌐 API 🐍	3	Program	✗
Toolformer (Schick et al., 2023)	GPT-J	🌐 API 🐍	5	Natural language	✗
VISPROG (Gupta and Kembhavi, 2022b)	GPT-3	🌐 API 🐍	> 10	Program	✓
Visual ChatGPT (Wu et al., 2023a)	ChatGPT	🌐 API 🐍	> 10	Natural Language	✓
ViperGPT (Surís et al., 2023)	GPT-3 Codex	🌐 API 🐍	> 10	Program	✓
MM-ReAct (Yang* et al., 2023)	ChatGPT/GPT-4	🌐 API 🐍	> 10	RegExp Match	✓
HuggingGPT (Shen et al., 2023b)	ChatGPT	🌐 API 🐍	> 10	Natural Language	✓
Chameleon (Lu et al., 2023b)	GPT-4	🌐 API 🐍	> 10	Natural Language	✓

Table 6.1: Glossary of representative works on chaining tools with LLMs. The LLMs used in these works include GPT-3 (Brown et al., 2020), GPT-J (Wang and Komatsuzaki, 2021), ChatGPT (OpenAI, 2022), GPT-3 Codex (Chen et al., 2021a) and GPT-4 OpenAI (2023a). : search engine. : code. : APIs other than search engine. : open-source models.

complementing ChatGPT (OpenAI, 2022) with various image generation tools. MM-ReAct shows that when collaborating various advanced vision experts, ChatGPT can perform complex multimodal actions and reasoning. Figure 6.2 presents the fast-evolving literature in multimodal agents from November 18, 2022 to July 26th, 2023. Among them, we include a few more exemplary multimodal agents in Table 6.1, along with two representative works in the NLP domain.

An overview of a typical multimodal agent framework is illustrated in Figure 6.3. The user directly interacts with the **Tool Allocator**, which functions as the brain of the agent. In current literature, the tool allocator is usually a LLM. To achieve the user’s goal, the LLM will outline all the steps necessary with either a single tool or collaborating multiple tools together. Subsequently, it will retrieve from all the candidate tools for the needed tools, and execute possibly multiple rounds of tools to fulfill the human requirement. Finally, the execution results from the tools are gathered as inputs of the LLM to generate a response to the user. Next, we cover the three key components of multimodal agents.

Tools. Tools are external modules that are callable by the LLM to obtain extra information that is missing from the model weights, including open-source models, public/private APIs, or code interpreters. As LLMs only accept language inputs, one must include tools that can process multimodal inputs to build a multimodal agent.

Planning. During planning, the LLM decomposes the user requests into smaller, manageable sub-problems, and outlines a step-by-step solution, each of which involves calling an external tool. There are two ways to teach LLMs for planning. One is to prompt the LLM with in-context few-shot examples of all candidate tools. This approach can extend the general model directly but is limited by the context length. The other approach relies on large amounts of annotated data to fine-tune the LLM, which most likely will damage the robustness and generalizability of the model.

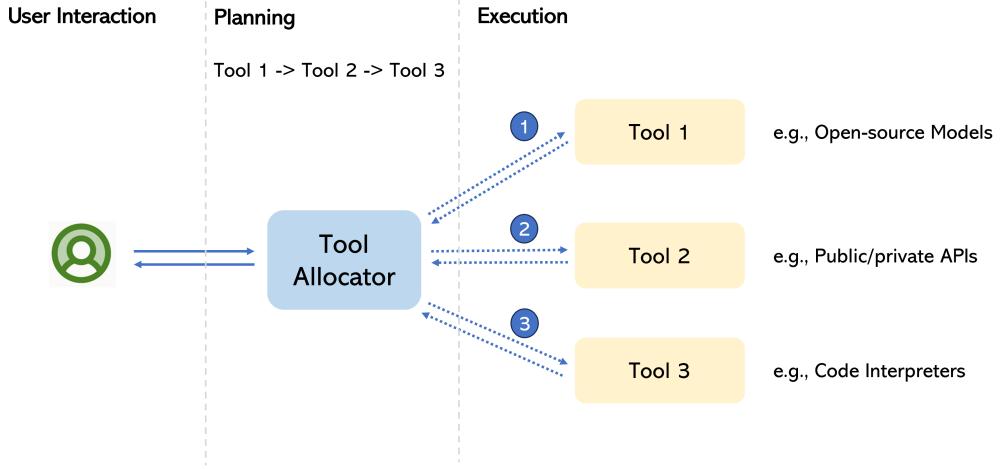


Figure 6.3: An overview of multimodal agent framework.

Execution. The generated plan is further translated into executable calls to the required tools, which can be done via regular expression matching (Yang* et al., 2023); directly prompting LLMs to generate executable programs (Surís et al., 2023); or leveraging in-context few-shot learning capability of LLMs by providing natural language instructions that describe the roles of each module together with a few calling examples (Lu et al., 2023b). The execution results are fed back to the LLM to generate a response to the user.



Figure 6.4: Input/Output modalities of MM-REACT (Yang* et al., 2023).

6.3 Case Study: MM-REACT

We use MM-REACT (Yang* et al., 2023) as a case study to show how to build a multimodal agent, its emerging capabilities in multimodal understanding, and how it can be easily extended to incorporate the latest and strongest LLM and potentially millions of tools.

6.3.1 System Design

MM-ReAct designs the system paradigm that composes numerous multimodal tools¹ with ChatGPT (OpenAI, 2022) for multimodal reasoning and action. By augmenting the language-only ChatGPT with various multimodal tools, MM-REACT supports both inputs and outputs in multimodalities, including text, image and video, as shown in Figure 6.4.

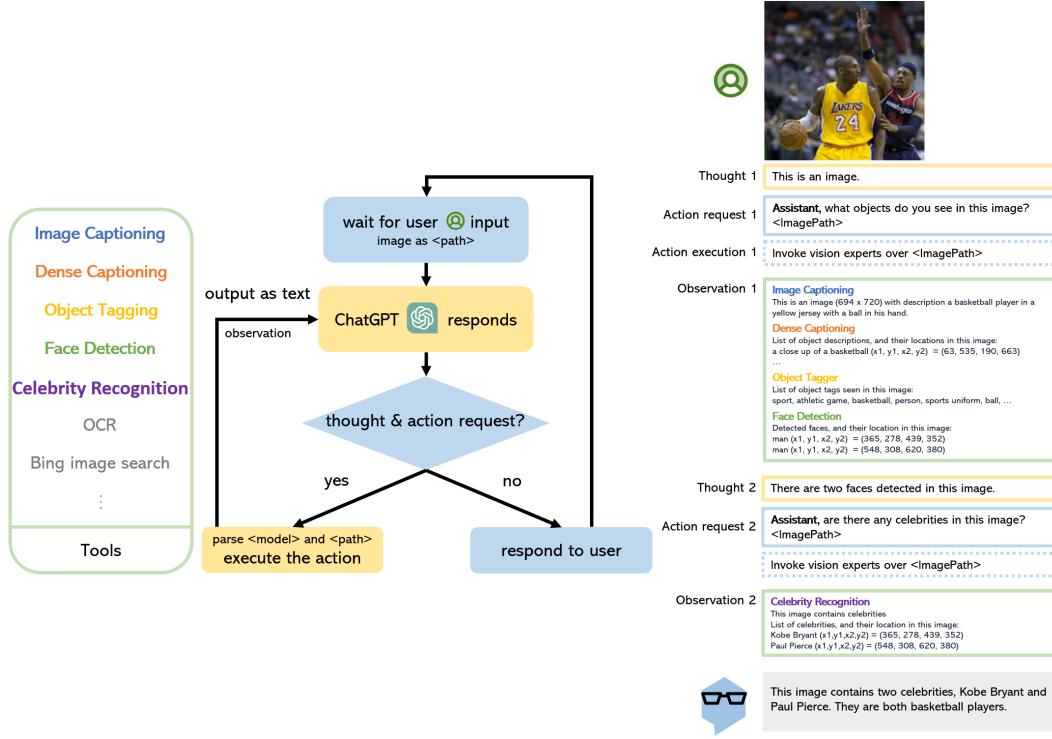


Figure 6.5: System design of MM-REACT (Yang* et al., 2023).

Figure 6.5 shows the system design of MM-REACT. The **multimodal tools** explored in MM-REACT are mainly computer vision models that take an image as input and interpret the image content from different perspectives. For instance, the image captioning model generates a natural description, the OCR model extracts the scene text in the image, the celebrity recognition model identifies the celebrity names, and the object detection model extracts the salient object with bounding box locations. LLMs such as ChatGPT serves as the brain of the agent, which plans on which tools to use, and in what order, based on the input image and the user intent. Next, with the example in Figure 6.5, we unfold the planning and execution of MM-REACT behind the scene.

User prompt. As ChatGPT only accepts language inputs, to enable image as inputs, we simply use the file path as the input to ChatGPT. The file path functions as a placeholder, allowing ChatGPT to treat it as a black box and later seek help from different tools during the planning stage. Besides the input image, the user can also provide the intent in text format (*e.g.*, a question about the input image). When there is no text input from the user, the goal is to get a general understanding about the image.

Planning. Upon receiving the input image and user prompt, ChatGPT plans for what tools to use. Inspired by REACT (Yao et al., 2022c), MM-REACT instructs ChatGPT to respond with certain watchwords, such as “Assistant, what objects are there in the image? <file path>”, if a specific tool is required (*i.e.*, *action request* in Figure 6.5). In practice, one can tell whether a multimodal tool is needed by simply string-matching the keyword “Assistant,” in ChatGPT’s response.

¹In Yang* et al. (2023), these tools are referred as experts. We unify the terminology as tools throughout this chapter.



Figure 6.6: Emerging capabilities of MM-REACT for multimodal reasoning and action. Image credit: Yang* et al. (2023).

MM-ReAct encourages ChatGPT to show the *thought* (reasoning) process to highlight why an external tool is needed, which has been proven beneficial in NLP studies (Yao et al., 2022c). In addition, for generating proper calling to each tool, both instructions and in-context examples are added as the prefix when prompting ChatGPT. Each tool is described with the model name, a general description of its capability, the input data format, and the output information. After describing each tool, a few in-context dialogue examples are also included for enhanced performance.

Execution. Given the action request from ChatGPT, the tool name and the file path can be parsed via regular expression matching, which are used to invoke the tool (*action execution*).

Take the example shown in Figure 6.5, upon receiving the input image, ChatGPT first invokes a series of tools for a general understanding about the image. The invoked tools include image captioning for an overall description of the image; dense captioning to get the region-level, more detailed description about the objects in the image; object tagging to get the tags of the objects in the image; face detection to get the box coordinates of the two faces mentioned in the object tags. The outputs from the tools (*i.e. observations*) are serialized as text, and fed back to ChatGPT.

Combining observations with the chat history, ChatGPT can further invoke additional experts or return the final answer to the user. In this specific example, ChatGPT invokes a second round of *thought-action-observation* over the two faces detected in the image and calls celebrity recognition to get the names of these two persons.

Response generation. When ChatGPT decides no external tools are needed, it takes consideration of all observations gathered and summarize them as the response to the user, which is “This image contains two celebrities, Kobe Bryant and Paul Pierce. They are both basketball players.” for the example shown in Figure 6.5.

If the user continues to interact with MM-REACT, it repeats the process described above, but with all observations and chat history available when planning for the tools needed. For instance, if the user then asks “how many championship rings did the player on the left win in his career”, it is not available in the existing observations nor chat history, but ChatGPT has the bounding boxes to decide who is on the left, and also the names of the players. It plans to invoke Bing Search to find the right answer, which should be 5.

6.3.2 Capabilities

Figure 6.6 shows the representative capabilities and application scenarios that MM-REACT demonstrates, including visual math and text reasoning, understanding visual-conditioned jokes/memes, spatial/coordinate understanding, visual planning and prediction, multi-image reasoning, multi-hop document understanding, open-world concept understanding, video analysis and summarization.

In addition, we show an example of the full response from MM-REACT on multi-image reasoning in Figure 6.7, which may not be easily achievable by visual instruction tuning in Chapter 5. For more comprehensive examples of all emerging capabilities of MM-REACT, we refer the reader to the original paper.

Receipt 1:

Receipt 2:

Receipt 3:

Receipt 4:

How much did I spend on grocery?

You spent \$25.97 on groceries at The Shop Store #100 in Chicago, IL.

How much did I spend on dining out?

You spent \$23.45 on dining out at the Windmill Family Restaurant.

How much in total did I spend on travel?

You spent \$1299.90 on airfare and \$43.83 on an Uber ride for a total of \$1343.73 on travel.

How much in total did I pay for taxes?

You paid \$323.23 in total for taxes. This includes \$321.90 for taxes and fees for the flight purchase and \$1.33 for sales tax for the Windmill Family Restaurant expenses.

Figure 6.7: Examples of MM-REACT’s capabilities and application scenarios: **multi-image reasoning**. Image credit: Yang* et al. (2023).

I. Principe de la détection de rayonnement avec un bolomètre

Comme illustré sur la figure 1 un bolomètre est constitué d'un absorbeur qui reçoit le rayonnement que l'on désire détecter. Sa température T , supposée uniforme, est mesurée à l'aide d'un thermomètre incorporé, constitué d'un matériau conducteur dont la résistance $R(T)$ varie avec la température T ; cette variation est caractérisée par le coefficient $\alpha = \frac{1}{R} \frac{dR}{dT}$. L'ensemble possède la capacité thermique C_R .

Un barreau, conducteur thermique, homogène, de longueur L , de section S et de conductivité thermique λ et sans échanges thermiques latéraux, relie le bolomètre à un thermostat de température T_b fixe.

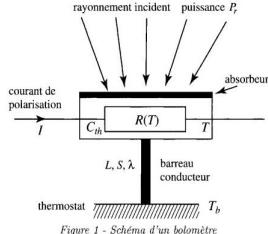


Figure 1 - Schéma d'un bolomètre

I.1. Dans cette première partie, on étudie le comportement du bolomètre en régime stationnaire, à la température de fonctionnement $T = T_b$.

I.1.a) Exprimer la température $T(x)$ en tout point du barreau conducteur repéré par son abscisse x ; on prendra $T(0) = T_0$ et $T(L) = T_b$.

GPT-4 (Multimodal)

User Answer question I.1.a. Think step-by-step.

GPT-4 To answer question I.1.a, we need to express the temperature $T(x)$ in every point of the conductive bar, which is characterized by the abscissa x . We are given that $T(0) = T_0$ and $T(L) = T_b$.

For a stationary regime, the heat equation for the one-dimensional heat conduction in the bar is:

$$\frac{d^2T}{dx^2} = 0$$

Integrating once, we get:

$$\frac{dT}{dx} = C_1$$

Integrating a second time:

$$T(x) = C_1x + C_2$$

To determine the constants C_1 and C_2 , we use the boundary conditions:

$$T(0) = T_0, \text{ which gives } C_2 = T_0$$

$$T(L) = T_b, \text{ which gives } C_1 = (T_b - T_0) / L$$

Finally, the expression for $T(x)$ is:

$$T(x) = (T_b - T_0) * (x / L) + T_0$$

This represents a linear temperature profile along the conductive bar, with a slope of $(T_b - T_0) / L$.



Answer question I.1.a. Think step-by-step.

MM-ReAct w/ ChatGPT

To answer question I.1.a, one would need to use the principles of thermodynamics and heat transfer to determine the temperature at any point in the conductor bar. This would likely involve solving a differential equation to find the temperature distribution along the bar. The solution would depend on the specific properties of the bar, such as its length, cross-sectional area, and thermal conductivity, as well as the temperature of the bolometer and the temperature of the thermostat. The solution would also depend on the power of the incident radiation and the current of polarization. Once the temperature distribution is found, the temperature at any point in the bar can be determined by evaluating the solution at that point.



MM-ReAct w/ GPT-4 (Language-only)

To answer question I.1.a, we'll consider the heat conduction equation along the thermal conductor bar. We're working under the assumptions of stationary mode and one-dimensional heat conduction with no lateral heat exchanges.



Start with Fourier's Law of heat conduction:

$$q = -k * (dT/dx)$$

where q is the heat flow (W/m^2), k is the thermal conductivity ($\text{W/m}\cdot\text{K}$), and dT/dx is the temperature gradient along the x -axis.

In a stationary regime, the heat flow q is constant throughout the bar. Thus, we have:

$$-k * (dT/dx) = \text{constant}$$

Integrate the equation with respect to x :

$$\int (dT/dx) dx = \int (\text{constant}/k) dx$$

After integration, we get:

$$T(x) = (\text{constant}/k) * x + C$$

where C is the integration constant.

Apply the boundary conditions to find the constants:

$$T(0) = T_0 \Rightarrow C = T_0$$

$$T(L) = T_b \Rightarrow (\text{constant}/k) * L + T_0 = T_b$$

Solve for the constant:

$$\text{constant} = k * (T_b - T_0) / L$$

Substitute the constant back into the $T(x)$ equation:

$$T(x) = (k * (T_b - T_0) / L) * x / k + T_0$$

$$T(x) = ((T_b - T_0) / L) * x + T_0$$

So, the temperature $T(x)$ at any point x along the thermal conductor bar is given by:

$$T(x) = ((T_b - T_0) / L) * x + T_0$$

Figure 6.8: Extensibility of multimodal agents: upgrading LLMs. Image credit: Yang* et al. (2023).

6.3.3 Extensibility

One favorable property of tool chaining to build multimodal agents is that the system can be easily extended and enhanced, from two perspectives. One is to upgrade the core part of the system, the LLM, and the other is to expand the number of external tools.

Upgrading LLM. The system design of MM-REACT allows for upgrading the core part of the system, the LLM, to newer and more powerful models as they come out, without the need of re-training. We show an example in Figure 6.8 on upgrading ChatGPT to language-only GPT-4, which improves MM-REACT to potentially match the performance of multimodal GPT-4.

Plug-and-play (adding more tools). Existing multimodal agents incorporates tools via a plug-and-play mechanism, allowing adding more tools without training. One prominent work along this direction is HuggingGPT (Shen et al., 2023b), which proposes to leverage all open-source models hosted on huggingface. Chameleon (Lu et al., 2023b), incorporates not only huggingface models, but also open-source models from GitHub, Bing search API, and python compiler. RestGPT (Song et al., 2023) proposes a multi-level online planning framework that effectively handles the practical challenges associated with integrating LLMs with more than 100 RESTful APIs. However, it remains challenging in scaling this framework to thousands to millions of tools, which is the potential future demonstrated in TaskMatrix.AI (Liang et al., 2023b).

Moreover, one can leverage SAM (Kirillov et al., 2023) as a tool to allow for more types of human interaction with the multimodal agent other than text. Recall in MM-REACT, the user intent is all captured by the natural language query from the user. In InternGPT (Liu et al., 2023l), by connecting the tool SAM with GPT, it allows for more ways to interact with the system, for example, via clicks, scribbles, and drawing bounding boxes. These additional interactions, to some extent, are mimicking the action of finger-pointing when we humans are having a conversation.

6.4 Advanced Topics

In this section, we discuss more advanced topics and shed light on potential future directions.

6.4.1 Comparison to Training with LLM in Chapter 5

We have covered two directions on building advanced multimodal systems based on LLMs. As the key distinction, the multimodal agents in this chapter leverages LLMs’ high-level planning abilities to allocate various multimodal tools, while training multimodal models with LLMs in Chapter 5 solely leverages LLMs for text generation conditioned on multimodal inputs.

Nonetheless, both of these methods exhibit their respective advantages and disadvantages. On one hand, instruction tuning enables an end-to-end model that directly interprets rich semantics in multimodal inputs, but requires data curation and training, hence more computationally expensive. However, limited instruction tuning data may limit its capabilities in certain scenarios, such as OCR. On the other hand, one can build a multimodal agent without any training by chaining LLMs with abundant off-the-shelf models/APIs/code interpreters as tools, and leveraging in-context few-shot examples to teach LLMs on planning. However, as there is no training, the system may fail to invoke the right tool. Moreover, weak domain experts may produce noisy outputs, that can confuse LLM on planning or reasoning, leading to weak performance.

Though these two approaches exhibit distinct variations,, we envision the possibility of an intermediate domain that amalgamates the strengths of both paradigms, and raise the following questions. Now that we have open-source LMM such as LLaVA (Liu et al., 2023c), can we replace the LLM with LLaVA as the tool allocator? If so, what capabilities would require a tool to be enabled? And what problems can be solved by instruction tuning. These are interesting directions that may worth exploring in the near future.

6.4.2 Improving Multimodal Agents

Existing multimodal agents mainly rely on in-context few-shot examples to teach LLM on planning, which can be unreliable, leading to inaccurate tool using. To improve the accuracy in planning, several works have been proposed and we group them into three categories below.

Composing tools via code generation. Most existing multimodal agents uses natural language to prompt LLM for planning which tool to use. Researchers (Gupta and Kembhavi, 2023; Surís et al., 2023) have also been exploring using programming language for more accurate execution. Visual programming (Gupta and Kembhavi, 2023) is a prominent work along this direction, which

uses the in-context learning ability of GPT-3 (Brown et al., 2020) to generate python-like modular programs from natural language instructions for compositional visual tasks ViperGPT Surís et al. (2023) instructs GPT-3 Codex (Chen et al., 2021a) to generate Python code to compose multimodal tools for a one-round query answering. However, as the codes are still generated by a LLM, the problem of inaccurate tool using still remains.

Improving accuracy in tool using: self-assessment. A recent work AssistGPT (Gao et al., 2023a) tries to improve the accuracy in tool using via self-assessment. It adds a stage of inspection and learning loop into the system. When the round of plan and execution is finished, the system inspects the outcome, and determines whether the reasoning path of calling the tool is a success or not, if so, save it as an in-context example, to teach LLM for a more accurate tool calling in the future rounds.

Improving accuracy in tool using: instruction tuning. Another thread on improving accuracy in tool using is to combine the system with instruction tuning (Patil et al., 2023; Yang et al., 2023c). One can generate a dataset of instruction-API pairs via self-instruct to tune a smaller LLM (e.g. , Vicuna-7B (Vicuna, 2023)).

LMM as the tool allocator? In addition, as LMMs evolve, we envision that the LLM can be replaced by a LMM as the tool allocator in the system, to enable even more advanced application scenarios. If the tool allocator can take multimodal inputs, there is no need to unify the outputs of tools into text sequence, allowing more natural interactions between the tool allocator and multimodal tools, particularly those producing multimodal outputs. For instance, one can imagine using multimodal GPT-4 (OpenAI, 2023a) to coordinate various image or video generation tools to make a short movie by providing it with a sketch of the storyline and visual examples of the main characters.

6.4.3 Diverse Applications of Multimodal Agents

By composing tools from a specific domain, this new system paradigm can also support diverse domain-specific applications.

Yu et al. (2023b) composes LLMs with image synthesis tools and object-level/pixel-level image understanding tools to build a data synthesis pipeline to provide diverse annotations on synthesized image. Instruct2Act (Huang et al., 2023c) complements the LLM with robotic executors, to enable robotic actions based on multi-modal instructions. When chaining a pool of audio models with LLM, AudioGPT (Huang et al., 2023a) can understand and generate speech, music, sound and talking head. Similarly, WavJourney (Liu et al., 2023i) further supports compositional audio creation with storylines encompassing speech, music, and sound effects. With tracking, captioning, audio understanding models, ChatVideo (Wang et al., 2023c) enables ChatGPT to understand multi-channel videos. Other application scenarios include 3D scene generation (Lin et al., 2023; Feng et al., 2023), medical image understanding (Liu and Zuo, 2023; Sun et al., 2023c) and vision-language navigation (Zhou et al., 2023b).

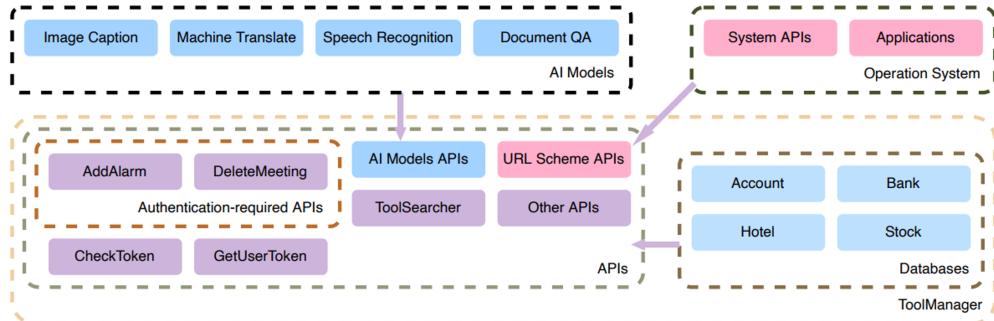


Figure 6.9: Example of evaluation benchmark focusing on toll using accuracy.

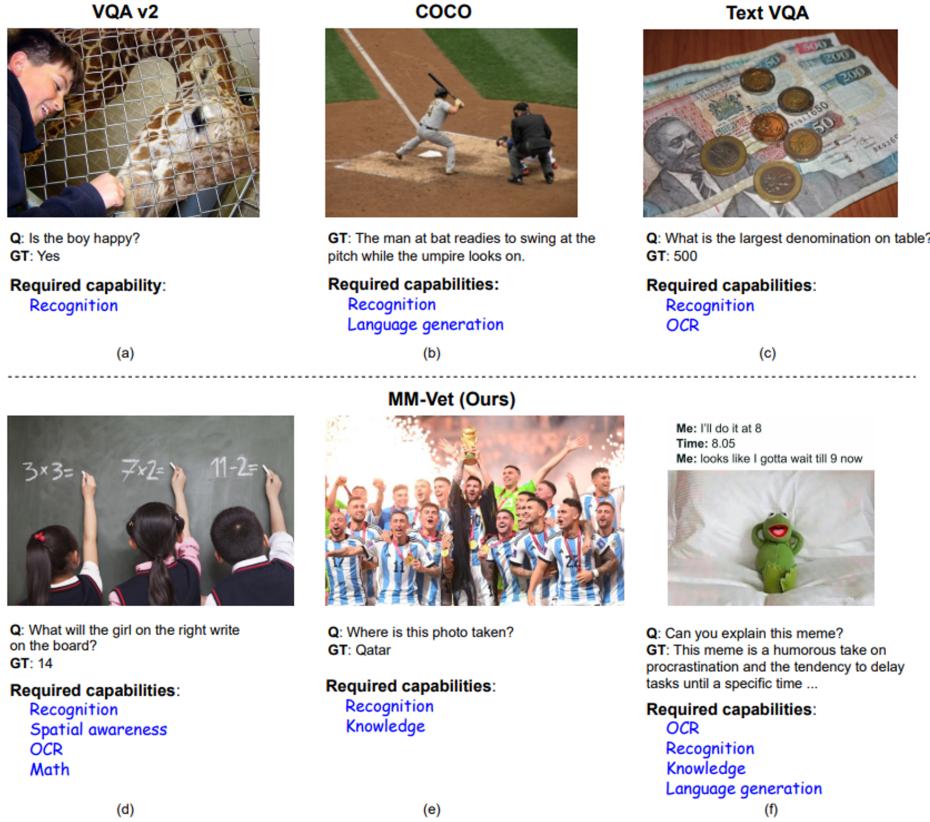


Figure 6.10: Example of evaluation benchmark focusing on emergent capabilities. MM-Vet focuses on the integration of different core VL capabilities, including recognition, OCR, knowledge, language generation, spatial awareness, and math. Image credit: [Yu et al. \(2023d\)](#).

6.4.4 Evaluation of Multimodal Agents

Multimodal tool using. Although we have seen qualitative examples of new scenarios enabled by multimodal agents, it remains unclear how these agents perform in terms of the accuracy in tool using. API-Bank ([Li et al., 2023k](#)) is a starting point on building pipeline in systematically evaluating tool-augmented LLMs.

Emergent capabilities. Existing VL benchmarks focus on specific capabilities of interest, such as visual recognition ([Antol et al., 2015](#)), image description ([Chen et al., 2015; Agrawal et al., 2019](#)), as well as other benchmarks for specialized capabilities such as scene text understanding ([Sidorov et al., 2020; Gurari et al., 2018](#)), commonsense reasoning ([Zellers et al., 2019](#)), outside knowledge ([Schwenk et al., 2022](#)). The intriguing abilities shown in large multimodal models and multimodal agents are not examined by existing benchmarks, such as solving math problems written on the blackboard, reasoning about events and celebrities in news images, or explaining visual jokes. Furthermore, the long, chatty outputs from these systems poses challenges to today’s evaluation metrics. Researchers ([Fu et al., 2023; Liu et al., 2023j](#)) have started to design comprehensive evaluation samples to facilitate the LMM evaluation. As an attempt to test multimodal systems on integrated capabilities, MM-Vet ([Yu et al., 2023d](#)) defines 6 core VL capabilities and examines the 16 integrations of interest derived from the capability combination (Figure 6.10). In addition, to accommodate for the open-ended free-form text outputs, MM-Vet proposes an LLM-based evaluator to enable evaluation across different question types and answer styles.

6.4.5 Tool Creation

Imagine if we have a completely new scenario without a robust tool to use. Can we create a tool based on the user need on-the-fly? In NLP, CREATOR (Qian et al., 2023) proposes to create tools by writing python code for math reasoning, as opposed to calling math solver API such as Wolfram Alpha. Cai et al. (2023) further explores the capabilities of LLMs to make tools, and experiment with two LLMs, one as the tool maker and the other as the tool user to collaboratively solve complicated tasks, such as scheduling a meeting. In terms of multimodal agents, the challenge is how to create a tool that can process multimodal inputs. One may follow ViperGPT (Surís et al., 2023) to instruct LLMs to generate python programs leveraging pre-existent python packages such as OpenCV. AutoML GPT (Zhang et al., 2023j) envisions that one can utilize LLMs to automate the model training pipeline. There may be potential to develop novel multimodal deep learning tools tailored to more effectively address the requirements of users.

6.4.6 Retrieval-Augmented Multimodal Agents

In real-life applications, a substantial portion of information resides within databases, and user needs may require accurate retrieval of such information. Meanwhile, it is infeasible to encode all the world knowledge into the weights of pre-trained models, particularly when it comes to the long-tail concepts and fast-evolving data.

In NLP, several works augment LLMs with external data encoded with structured language and relation representations (Peters et al., 2019; Guu et al., 2020; Lewis et al., 2020). Given input texts, such retrieved-augmented models utilize a retriever that retrieves relevant documents from an external memory, and uses a generator to generate predictions given the retrieved documents.

Motivated by retrieval-augmented models in NLP, several recent works leverage visual and / or textual knowledge to improve vision tasks, such as image classification (Long et al., 2022), captioning (Yang et al., 2023a), question answering (Wu et al., 2021; Marino et al., 2021; Yang et al., 2022d; Chen et al., 2022e), image generation (Blattmann et al., 2022; Sheynin et al., 2022; Chen et al., 2022f; Zhou et al., 2022c), and multi-modal tasks simultaneously (Yasunaga et al., 2022). RAC (Long et al., 2022) improves long-tail classification by retrieving from a non-parametric memory consisting of pre-encoded images and text. K-LITE (Shen et al., 2022a) enhances the text prompts with the retrieved external knowledge that is encoded in natural language. REACT (Liu et al., 2023d) retrieve from billions of the paired knowledge of image-text and aims to improve task transfer performance for core vision problems. Among them, RA-CM3 (Yasunaga et al., 2022) builds the first retrieval-augmented LMM with a multimodal retriever to retrieve multimodal documents, and a retrieval-augmented generator that can generate both text and image. Chaining tools with LLM shares a strong connection with the retrieval-augmented methods in that both leverage external knowledge to provide additional information for the core model to utilize. In the multimodal regime, the image itself can be used as the query to gain external knowledge, either retrieved from a knowledge base, or extracted from another pre-trained vision expert models.

Chapter 7

Conclusions and Research Trends

Multimodal foundation models have garnered significant interest among scholars in the fields of computer vision and multimodal vision-language research. Although prevailing research topics, approaches and methodologies have been evolving – encompassing image self-supervised learning, language-image contrastive learning, text-to-image generation, unified vision modeling, and large language-and-vision assistants – they converge on a common overarching objective: the creation of general-purpose models and systems capable of following human intents and effortlessly executing a diverse array of vision and vision-language tasks in the wild. In this chapter, we provide a concise summary of what has been reviewed, and delve into the prevailing research tendencies in the field.

7.1 Summary and Conclusions

This paper surveys the most recent advances at the frontier of multimodal foundation model research, categorized into two classes discussed below.

- **Specific-purpose multimodal foundation models.** There is a diverse set of problems to tackle in the computer vision community. To lay a comprehensive foundation for the introduction of general-purpose visual assistants, we have discussed many seminar papers in the era of pre-training. The major paradigm during this period is pre-training on a large amount of problem-related data, and then transferring to a number of real-world scenarios of the same problem type in a zero- or few-shot fashion. More specifically, we have presented two general topics: (*i*) *Visual Understanding* in Chapter 2: individual multimodal foundation models have developed to analyze the content of visual data in the image, region, pixel levels, prospectively. The language-augmented vision models are a popular family, contributing to the recent success of visual understanding tasks in the wild. (*ii*) *Visual Generation* in Chapter 3: text-to-image generation models have served the foundation for image synthesis, which has been successfully extended to allow user controllability and customization at more fine-grained manners. The availability and creation of large amount of problem-related data has played a key role in making these multimodal foundation models possible.
- **General-purpose assistants.** We have reviewed recently emerged literature on building general-purpose assistants, which often possess an unified network architecture, an unified input-output data format, and a general interface that facilitates easy interaction with humans. Inspired by the success in NLP that LLM such as ChatGPT/GPT-4 is a general assistant for a wide range of language tasks, researchers in computer vision have explored various solutions to their counterpart for vision tasks. Based on how LLM is leveraged in the methodology, existing works can be categorized into three topics: (*i*) *Unified Vision Models* in Chapter 4: The spirit of unifying modeling in LLM is borrowed to build unified vision models at different levels and across different tasks. (*ii*) *Training with LLM* in Chapter 5: Starting with a pre-trained LLM, visual data is connected to LLM for end-to-end training. (*iii*) *Chaining with LLM* in Chapter 6: By freezing LLM, existing vision experts can be triggered by prompt engineering LLM to complete specific vision tasks.

The comparisons among these models are summarized in Table 7.1.

Models	Advantages	Disadvantages
Visual Understanding <i>Specific-Purpose Models</i>	Well studied and scalable solutions on image-level understanding; Emerging interests and success on region-level and pixel-level visual understanding	High training cost; No successful scalable recipe beyond the billion-image level
	Well studied and scalable solutions on image-level generation; Emerging interests and success in controllable/customized image generation	High training and inference cost; Debate between diffusion and auto-regressive solutions for the best recipe; More studies are needed for video generation
Unified Vision Models <i>General-Purpose Assistants</i>	Promises to unlock new emerging capabilities and scenarios	High risks in modeling and high training cost
	Some new emerging capabilities and scenarios are enabled with light model training	The performance is bounded by LLM
Chaining with LLM	Fast system development cycles with low cost as no training is involved	Low flexibility in improving system performance; No new emerging capabilities

Table 7.1: Comparisons of different multimodal foundation model families covered in this paper.

7.2 Towards Building General-Purpose AI Agents

At the end of each chapter, we have discussed future trends for individual topics. The paper itself is organized to demonstrate the transition from specialist multimodal foundation models to general-purpose visual assistants. Though powerful, existing visual assistants such as Flamingo (Alayrac et al., 2022) and multimodal GPT-4 (OpenAI, 2023b) are in the preliminary form, compared with grand vision on building a general-purpose multimodal AI agent via foundation models. In what follows, we highlight a number of research trends towards this goal.

Generalist agents with multi-modality. This aligns with the grand goal of building a single generalist agent that interacts with world like humans through fusing multiple channels such as language, vision, speech and actions. From this perspective, the notion of multimodal foundation models becomes somewhat indistinct on its own. Instead, it serves as a crucial component of the agent for perceiving and synthesizing visual signals. For example, Gato (Reed et al., 2022) and PaLM-E (Driess et al., 2023) perform a wide range of language, multimodal and control tasks with a single set of model weights, where visual perception is a crucial component in understanding the environment. It also raises challenges in the effective and scalable pre-training objectives for unified vision and multimodal modeling.

Alignment with human intents. AI alignment research focuses on steering AI systems towards humans’ intended goals, values, or ethical guidelines. An AI system is deemed aligned when it effectively promotes the desired goals. Though language has exhibited its generality in expressing human intents, it is not always the best option. As demonstrated in SAM (Kirillov et al., 2023) and ControlNet/GLIGEN (Zhang and Agrawala, 2023; Li et al., 2023n), human intents can be more precisely and conveniently represented in visual prompts such as key points, bounding boxes and sketch drawing, for visual understanding and generation tasks, respectively. Building foundation models that are well equipped with such a multimodal human-machine interaction interface is a key step to unlock new use scenarios, where human intents are best represented visually. For example, the spatial arrangement of elements within a scene, as well as the artistic style and visual appeal of a piece of visual art.

Planning, memory, and tool use. It is highlighted in Weng (2023) that a LLM-powered autonomous agent system can be built, where LLM functions as the agent’s brain, complemented by several key components: planning, memory and tool use. Following the framework, we could foresee the role of multimodal foundation models in this AI agent system. (*i*) Planning. To complete complex tasks in real-world scenarios, the agent should be able to decompose large tasks into smaller, manageable subgoals, enabling efficient handling of complex tasks. In the ideal case, the

AI agent possesses the self-improvement ability, engaging in self-assessment and introspection regarding previous actions, enabling it to learn from errors and enhance its approach for subsequent endeavors, ultimately leading to better outcomes. Visual modality is a common channel to represent state of the environment. To facilitate planning, it raises challenges in improving the capability of the current visual understanding models in perceiving more fine-grained visual details and longer sequence videos. *(ii) Memory.* For short-term memory, in-context learning (or prompt engineering) is utilized as short-term memory for the model to learn. Interleaved multimodal prompts can enable new scenarios to clarify the human intents. For long-term memory, it provides the agent with the capability to recall external knowledge over extended sessions, which can be implemented by fast retrieving from a multi-modal vector space (Liu et al., 2023d). In term of modeling, foundation models are required to learn the new skills to effectively leverage both types of memory. *(iii) Tool use.* The agent learns to utilize external APIs for knowledge that is missing from the foundation model weights. New capabilities are required to deal with the vision modality in several scenarios. For example, based on both the input visual signal and instructions, the model decides and plans whether certain external APIs are needed to complete the goal, such as code execution of detection/segmentation/OCR/generator experts.

The field of multimodal foundation models is evolving at a rapid speed, with new directions/methods emerging frequently. There are many important research topics that are not discussed in this paper, mostly due to the daily-updated research innovation. We are optimistic about the future of multimodal foundation models, not only because we are convinced that foreseeable exciting research innovations/ideas in individual areas are becoming reality by following the path of LLM in the near future, but also because connecting computer vision with the broader AI community, and building general-purpose AI agents is going to significantly advance the daily life of human being.

Acknowledgments

This book is largely based on our CVPR 2023 tutorial on vision foundation models. Many people have supported us and provided valuable feedback to the writing of this book. We thank all the authors who have contributed to the related papers, which makes the tutorial and book possible. We are also grateful to Mark de Jongh, the editor from the journal of *Foundations and Trends® in Computer Graphics and Vision*, for inspiring and encouraging us to write the book on multimodal foundation models.

Bibliography

- (2022). Stable diffusion. <https://github.com/CompVis/stable-diffusion>.
- Agarwal, A., Karanam, S., Joseph, K., Saxena, A., Goswami, K., and Srinivasan, B. V. (2023). A-star: Test-time attention segregation and retention for text-to-image synthesis. *arXiv preprint arXiv:2306.14544*.
- Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., and Anderson, P. (2019). nocaps: novel object captioning at scale. In *ICCV*.
- Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., et al. (2022). Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*.
- Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. (2022). Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*.
- Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., and Kochut, K. (2017). Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268*.
- Amrani, E., Karlinsky, L., and Bronstein, A. (2022). Self-supervised classification network. In *ECCV*.
- Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.
- Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. (2015). Vqa: Visual question answering. In *ICCV*.
- Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., and Saunshi, N. (2019). A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*.
- Assran, M., Caron, M., Misra, I., Bojanowski, P., Bordes, F., Vincent, P., Joulin, A., Rabbat, M., and Ballas, N. (2022). Masked siamese networks for label-efficient learning. In *ECCV*.
- Avrahami, O., Aberman, K., Fried, O., Cohen-Or, D., and Lischinski, D. (2023a). Break-a-scene: Extracting multiple concepts from a single image. *arXiv preprint arXiv:2305.16311*.
- Avrahami, O., Fried, O., and Lischinski, D. (2022a). Blended latent diffusion. *arXiv preprint arXiv:2206.02779*.
- Avrahami, O., Hayes, T., Gafni, O., Gupta, S., Taigman, Y., Parikh, D., Lischinski, D., Fried, O., and Yin, X. (2023b). Spatext: Spatio-textual representation for controllable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18370–18380.
- Avrahami, O., Lischinski, D., and Fried, O. (2022b). Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218.
- Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Jitsev, J., Kornblith, S., Koh, P. W., Ilharco, G., Wortsman, M., and Schmidt, L. (2023). Openflamingo.
- Awais, M., Naseer, M., Khan, S., Anwer, R. M., Cholakkal, H., Shah, M., Yang, M.-H., and Khan, F. S. (2023). Foundational models defining a new era in vision: A survey and outlook. *arXiv preprint arXiv:2307.13721*.

- Bachman, P., Hjelm, R. D., and Buchwalter, W. (2019). Learning representations by maximizing mutual information across views. *NeurIPS*.
- Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J., and Auli, M. (2022). Data2vec: A general framework for self-supervised learning in speech, vision and language. In *ICML*.
- Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *ICLR*.
- Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. (2023a). Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.
- Bai, S., Yang, S., Bai, J., Wang, P., Zhang, X., Lin, J., Wang, X., Zhou, C., and Zhou, J. (2023b). Touchstone: Evaluating vision-language models by language models.
- Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al. (2022). ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*.
- Balažević, I., Steiner, D., Parthasarathy, N., Arandjelović, R., and Hénaff, O. J. (2023). Towards in-context scene understanding. *arXiv preprint arXiv:2306.01667*.
- Bansal, A., Chu, H.-M., Schwarzschild, A., Sengupta, S., Goldblum, M., Geiping, J., and Goldstein, T. (2023). Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 843–852.
- Bansal, A., Sikka, K., Sharma, G., Chellappa, R., and Divakaran, A. (2018). Zero-shot object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 384–400.
- Bao, H., Dong, L., and Wei, F. (2022). BEiT: Bert pre-training of image transformers. In *ICLR*.
- Bar, A., Gandelsman, Y., Darrell, T., Globerson, A., and Efros, A. (2022). Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, 35:25005–25017.
- Bardes, A., Ponce, J., and LeCun, Y. (2021). Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*.
- Berant, J., Chou, A., Frostig, R., and Liang, P. (2013). Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1533–1544.
- Bitton, Y., Bansal, H., Hessel, J., Shao, R., Zhu, W., Awadalla, A., Gardner, J., Taori, R., and Schimdt, L. (2023). Visit-bench: A benchmark for vision-language instruction following inspired by real-world use.
- Black, K., Janner, M., Du, Y., Kostrikov, I., and Levine, S. (2023). Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*.
- Blattmann, A., Rombach, R., Oktay, K., and Ommer, B. (2022). Retrieval-augmented diffusion models. *arXiv preprint arXiv:2204.11824*.
- Bolya, D., Zhou, C., Xiao, F., and Lee, Y. J. (2019). Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9157–9166.
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
- Brooks, T., Holynski, A., and Efros, A. A. (2023). Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. In *NeurIPS*.
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.
- Byeon, M., Park, B., Kim, H., Lee, S., Baek, W., and Kim, S. (2022). Coyo-700m: Image-text pair dataset. <https://github.com/kakaobrain/coyo-dataset>.

- Cai, T., Wang, X., Ma, T., Chen, X., and Zhou, D. (2023). Large language models as tool makers. *arXiv preprint arXiv:2305.17126*.
- Cai, Z., Kwon, G., Ravichandran, A., Bas, E., Tu, Z., Bhotika, R., and Soatto, S. (2022). X-detr: A versatile architecture for instance-wise vision-language tasks. In *ECCV*.
- Cao, L., Zhang, B., Chen, C., Yang, Y., Du, X., Zhang, W., Lu, Z., and Zheng, Y. (2023). Less is more: Removing text-regions improves clip training efficiency and robustness. *arXiv preprint arXiv:2305.05095*.
- Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In *ECCV*.
- Caron, M., Bojanowski, P., Joulin, A., and Douze, M. (2018). Deep clustering for unsupervised learning of visual features. In *ECCV*.
- Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. *NeurIPS*.
- Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *ICCV*.
- Castrejon, L., Kundu, K., Urtasun, R., and Fidler, S. (2017). Annotating object instances with a polygon-rnn. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5230–5238.
- Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.-H., Murphy, K., Freeman, W. T., Rubinstein, M., et al. (2023). Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*.
- Chang, H., Zhang, H., Jiang, L., Liu, C., and Freeman, W. T. (2022). Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325.
- Changpinyo, S., Sharma, P., Ding, N., and Soricut, R. (2021). Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*.
- Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., and Cohen-Or, D. (2023). Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *arXiv preprint arXiv:2301.13826*.
- Chen, C., Zhang, B., Cao, L., Shen, J., Gunter, T., Jose, A. M., Toshev, A., Shlens, J., Pang, R., and Yang, Y. (2023a). Stair: Learning sparse text and image representation in grounded tokens. *arXiv preprint arXiv:2301.13081*.
- Chen, D., Liu, J., Dai, W., and Wang, B. (2023b). Visual instruction tuning with polite flamingo. *arXiv preprint arXiv:2307.01003*.
- Chen, F., Han, M., Zhao, H., Zhang, Q., Shi, J., Xu, S., and Xu, B. (2023c). X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. *arXiv preprint arXiv:2305.04160*.
- Chen, F., Zhang, D., Han, M., Chen, X., Shi, J., Xu, S., and Xu, B. (2022a). Vlp: A survey on vision-language pre-training. *arXiv preprint arXiv:2202.09061*.
- Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., and Zhao, R. (2023d). Shikra: Unleashing multimodal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*.
- Chen, L., Zhai, M., He, J., and Mori, G. (2019). Object grounding via iterative context reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0.
- Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.
- Chen, M., Laina, I., and Vedaldi, A. (2023e). Training-free layout control with cross-attention guidance. *arXiv preprint arXiv:2304.03373*.
- Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. (2021a). Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.
- Chen, Q., Chen, X., Zeng, G., and Wang, J. (2022b). Group detr: Fast training convergence with decoupled one-to-many label assignment. *arXiv preprint arXiv:2207.13085*.

- Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020a). A simple framework for contrastive learning of visual representations. In *ICML*.
- Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. E. (2020b). Big self-supervised models are strong semi-supervised learners. *NeurIPS*.
- Chen, T., Saxena, S., Li, L., Fleet, D. J., and Hinton, G. (2022c). Pix2seq: A language modeling framework for object detection. In *ICLR*.
- Chen, T., Saxena, S., Li, L., Lin, T.-Y., Fleet, D. J., and Hinton, G. (2022d). A unified sequence interface for vision tasks. *arXiv preprint arXiv:2206.07669*.
- Chen, W., Hu, H., Chen, X., Verga, P., and Cohen, W. W. (2022e). Murag: Multimodal retrieval-augmented generator for open question answering over images and text. *arXiv preprint arXiv:2210.02928*.
- Chen, W., Hu, H., Li, Y., Rui, N., Jia, X., Chang, M.-W., and Cohen, W. W. (2023f). Subject-driven text-to-image generation via apprenticeship learning. *arXiv preprint arXiv:2304.00186*.
- Chen, W., Hu, H., Saharia, C., and Cohen, W. W. (2022f). Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*.
- Chen, X., Ding, M., Wang, X., Xin, Y., Mo, S., Wang, Y., Han, S., Luo, P., Zeng, G., and Wang, J. (2022g). Context autoencoder for self-supervised representation learning. *arXiv preprint arXiv:2202.03026*.
- Chen, X., Djolonga, J., Padlewski, P., Mustafa, B., Changpinyo, S., Wu, J., Ruiz, C. R., Goodman, S., Wang, X., Tay, Y., et al. (2023g). Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*.
- Chen, X., Fan, H., Girshick, R., and He, K. (2020c). Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.
- Chen, X., Fang, H., Lin, T., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. (2015). Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.
- Chen, X. and He, K. (2021). Exploring simple siamese representation learning. In *CVPR*.
- Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., et al. (2022h). Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.
- Chen, X., Xie, S., and He, K. (2021b). An empirical study of training self-supervised vision transformers. In *ICCV*.
- Chen, X., Zhao, Z., Yu, F., Zhang, Y., and Duan, M. (2021c). Conditional diffusion for interactive segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7345–7354.
- Chen, X., Zhao, Z., Zhang, Y., Duan, M., Qi, D., and Zhao, H. (2022i). Focalclick: Towards practical interactive image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1300–1309.
- Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. (2020d). UNITER: Universal image-text representation learning. In *ECCV*.
- Chen, Z., Duan, Y., Wang, W., He, J., Lu, T., Dai, J., and Qiao, Y. (2022j). Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*.
- Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., and Girdhar, R. (2022). Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299.
- Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., and Jitsev, J. (2023). Reproducible scaling laws for contrastive language-image learning. In *CVPR*.
- Cho, J., Lei, J., Tan, H., and Bansal, M. (2021). Unifying vision-and-language tasks via text generation. In *ICML*.
- Cho, J., Li, L., Yang, Z., Gan, Z., Wang, L., and Bansal, M. (2023). Diagnostic benchmark and iterative inpainting for layout-guided image generation. *arXiv preprint arXiv:2304.06671*.

- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. (2022). Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.
- Computer, T. (2023). Redpajama-data: An open source recipe to reproduce llama training dataset.
- Crawshaw, M. (2020). Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*.
- Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65.
- Dai, H., Ma, C., Liu, Z., Li, Y., Shu, P., Wei, X., Zhao, L., Wu, Z., Zhu, D., Liu, W., et al. (2023a). Samaug: Point prompt augmentation for segment anything model. *arXiv preprint arXiv:2307.01187*.
- Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., and Hoi, S. (2023b). Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*.
- Dai, X., Chen, Y., Xiao, B., Chen, D., Liu, M., Yuan, L., and Zhang, L. (2021). Dynamic head: Unifying object detection heads with attentions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7373–7382.
- Datta, R., Joshi, D., Li, J., and Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (Csur)*, 40(2):1–60.
- Deng, C., Wu, Q., Wu, Q., Hu, F., Lyu, F., and Tan, M. (2018). Visual grounding via accumulated attention. In *CVPR*.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *CVPR*.
- Desai, K. and Johnson, J. (2021). Virtex: Learning visual representations from textual annotations. In *CVPR*.
- Desai, K., Kaul, G., Aysola, Z., and Johnson, J. (2021). Redcaps: Web-curated image-text data created by the people, for the people. In *NeurIPS, Track on Datasets and Benchmarks*.
- Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.
- Dhariwal, P. and Nichol, A. (2021). Diffusion models beat gans on image synthesis. In *NeurIPS*.
- Ding, J., Xue, N., Xia, G.-S., and Dai, D. (2022a). Decoupling zero-shot semantic segmentation.
- Ding, Z., Wang, J., and Tu, Z. (2022b). Open-vocabulary panoptic segmentation with maskclip. *arXiv preprint arXiv:2208.08984*.
- Dong, B., Zeng, F., Wang, T., Zhang, X., and Wei, Y. (2021). Solq: Segmenting objects by learning queries. *Advances in Neural Information Processing Systems*, 34:21898–21909.
- Dong, X., Bao, J., Zhang, T., Chen, D., Zhang, W., Yuan, L., Chen, D., Wen, F., and Yu, N. (2022). Bootstrapped masked autoencoders for vision bert pretraining. In *ECCV*.
- Dong, X., Bao, J., Zhang, T., Chen, D., Zhang, W., Yuan, L., Chen, D., Wen, F., Yu, N., and Guo, B. (2023). Peco: Perceptual codebook for bert pre-training of vision transformers. In *AAAI*.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.
- Dou, Z.-Y., Kamath, A., Gan, Z., Zhang, P., Wang, J., Li, L., Liu, Z., Liu, C., LeCun, Y., Peng, N., Gao, J., and Wang, L. (2022a). Coarse-to-fine vision-language pre-training with fusion in the backbone. In *NeurIPS*.
- Dou, Z.-Y., Xu, Y., Gan, Z., Wang, J., Wang, S., Wang, L., Zhu, C., Liu, Z., Zeng, M., et al. (2022b). An empirical study of training end-to-end vision-and-language transformers. In *CVPR*.
- Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al. (2023). PaLM-E: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

- Du, Y., Liu, Z., Li, J., and Zhao, W. X. (2022). A survey of vision-language pre-trained models. In *IJCAI survey track*.
- El-Nouby, A., Izacard, G., Touvron, H., Laptev, I., Jegou, H., and Grave, E. (2021). Are large-scale datasets necessary for self-supervised pre-training? *arXiv preprint arXiv:2112.10740*.
- Elharrouss, O., Almaadeed, N., Al-Maadeed, S., and Akbari, Y. (2020). Image inpainting: A review. *Neural Processing Letters*, 51:2007–2028.
- Ermolov, A., Siarohin, A., Sangineto, E., and Sebe, N. (2021). Whitening for self-supervised representation learning. In *ICML*.
- Esser, P., Rombach, R., and Ommer, B. (2021). Taming transformers for high-resolution image synthesis. In *CVPR*.
- Everingham, M. and Winn, J. (2011). The pascal visual object classes challenge 2012 (voc2012) development kit. *Pattern Analysis, Statistical Modelling and Computational Learning, Tech. Rep*, 8(5).
- Fan, L., Krishnan, D., Isola, P., Katabi, D., and Tian, Y. (2023a). Improving clip training with language rewrites. *arXiv preprint arXiv:2305.20088*.
- Fan, Y., Watkins, O., Du, Y., Liu, H., Ryu, M., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Lee, K., and Lee, K. (2023b). Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *arXiv preprint arXiv:2305.16381*.
- Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., and Cao, Y. (2023). Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*.
- Fang, Z., Wang, J., Hu, X., Liang, L., Gan, Z., Wang, L., Yang, Y., and Liu, Z. (2022). Injecting semantic concepts into end-to-end image captioning. In *CVPR*.
- Feichtenhofer, C., Li, Y., He, K., et al. (2022). Masked autoencoders as spatiotemporal learners. *NeurIPS*.
- Feng, C., Zhong, Y., Jie, Z., Chu, X., Ren, H., Wei, X., Xie, W., and Ma, L. (2022a). Promptdet: Towards open-vocabulary detection using uncurated images. In *European Conference on Computer Vision*, pages 701–717. Springer.
- Feng, W., He, X., Fu, T.-J., Jampani, V., Akula, A. R., Narayana, P., Basu, S., Wang, X. E., and Wang, W. Y. (2022b). Training-free structured diffusion guidance for compositional text-to-image synthesis. In *The Eleventh International Conference on Learning Representations*.
- Feng, W., Zhu, W., Fu, T.-j., Jampani, V., Akula, A., He, X., Basu, S., Wang, X. E., and Wang, W. Y. (2023). Layoutgpt: Compositional visual planning and generation with large language models. *arXiv preprint arXiv:2305.15393*.
- Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., and Mikolov, T. (2013). Devise: A deep visual-semantic embedding model. In *NeurIPS*.
- Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Qiu, Z., Lin, W., Yang, J., Zheng, X., et al. (2023). Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.
- Gadre, S. Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al. (2023). Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*.
- Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., and Taigman, Y. (2022). Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131*.
- Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D. (2022). An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.
- Gan, Z., Chen, Y.-C., Li, L., Zhu, C., Cheng, Y., and Liu, J. (2020). Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*.
- Gan, Z., Li, L., Li, C., Wang, L., Liu, Z., Gao, J., et al. (2022). Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*.
- Gao, D., Ji, L., Zhou, L., Lin, K. Q., Chen, J., Fan, Z., and Shou, M. Z. (2023a). Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn. *arXiv preprint arXiv:2306.08640*.

- Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., et al. (2023b). Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.
- Gao, P., Ma, T., Li, H., Lin, Z., Dai, J., and Qiao, Y. (2022). Convmae: Masked convolution meets masked autoencoders. *arXiv preprint arXiv:2205.03892*.
- Ge, Y., Ge, Y., Zeng, Z., Wang, X., and Shan, Y. (2023). Planting a seed of vision in large language model. *arXiv preprint arXiv:2307.08041*.
- Geng, X. and Liu, H. (2023). Openllama: An open reproduction of llama.
- Geng, Z., Yang, B., Hang, T., Li, C., Gu, S., Zhang, T., Bao, J., Zhang, Z., Hu, H., Chen, D., et al. (2023). Instructdiffusion: A generalist modeling interface for vision tasks. *arXiv preprint arXiv:2309.03895*.
- Ghiasi, G., Gu, X., Cui, Y., and Lin, T.-Y. (2022a). Open-vocabulary image segmentation. In *ECCV*.
- Ghiasi, G., Gu, X., Cui, Y., and Lin, T.-Y. (2022b). Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, pages 540–557. Springer.
- Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., and Misra, I. (2023). Imagebind: One embedding space to bind them all. In *CVPR*.
- Girshick, R. (2015). Fast r-cnn. In *ICCV*.
- Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2015). Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):142–158.
- Gong, T., Lyu, C., Zhang, S., Wang, Y., Zheng, M., Zhao, Q., Liu, K., Zhang, W., Luo, P., and Chen, K. (2023). Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*.
- Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. (2020). Bootstrap your own latent-a new approach to self-supervised learning. *NeurIPS*.
- Gu, X., Cui, Y., Huang, J., Rashwan, A., Yang, X., Zhou, X., Ghiasi, G., Kuo, W., Chen, H., Chen, L.-C., et al. (2023). Dataseg: Taming a universal multi-dataset multi-task segmentation model. *arXiv preprint arXiv:2306.01736*.
- Gu, X., Lin, T.-Y., Kuo, W., and Cui, Y. (2021). Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*.
- Gu, X., Lin, T.-Y., Kuo, W., and Cui, Y. (2022). Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*.
- Gudibande, A., Wallace, E., Snell, C., Geng, X., Liu, H., Abbeel, P., Levine, S., and Song, D. (2023). The false promise of imitating proprietary llms. *arXiv preprint arXiv:2305.15717*.
- Gunjal, A., Yin, J., and Bas, E. (2023). Detecting and preventing hallucinations in large vision language models. *arXiv preprint arXiv:2308.06394*.
- Gupta, T., Kamath, A., Kembhavi, A., and Hoiem, D. (2022a). Towards general purpose vision systems: An end-to-end task-agnostic vision-language architecture. In *CVPR*.
- Gupta, T., Kamath, A., Kembhavi, A., and Hoiem, D. (2022b). Towards general purpose vision systems: An end-to-end task-agnostic vision-language architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16399–16409.
- Gupta, T. and Kembhavi, A. (2022a). Visual programming: Compositional visual reasoning without training. *arXiv preprint arXiv:2211.11559*.
- Gupta, T. and Kembhavi, A. (2022b). Visual programming: Compositional visual reasoning without training. *ArXiv*, abs/2211.11559.
- Gupta, T. and Kembhavi, A. (2023). Visual programming: Compositional visual reasoning without training. In *CVPR*.

- Gupta, T., Marten, R., Kembhavi, A., and Hoiem, D. (2022c). Grit: General robust image task benchmark. *arXiv preprint arXiv:2204.13653*.
- Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., Luo, J., and Bigham, J. P. (2018). Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*.
- Gutmann, M. and Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*.
- Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M.-W. (2020). Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.
- Hafiz, A. M. and Bhat, G. M. (2020). A survey on instance segmentation: state of the art. *International journal of multimedia information retrieval*.
- Harley, A. W., Fang, Z., and Fragkiadaki, K. (2022). Particle video revisited: Tracking through occlusions using point trajectories. In *European Conference on Computer Vision*, pages 59–75. Springer.
- He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022a). Masked autoencoders are scalable vision learners. In *CVPR*.
- He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.
- He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *ICCV*.
- He, K., Sun, J., and Tang, X. (2010). Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*.
- He, P., Liu, X., Gao, J., and Chen, W. (2021). DeBERTa: Decoding-enhanced bert with disentangled attention. In *ICLR*.
- He, R., Sun, S., Yu, X., Xue, C., Zhang, W., Torr, P., Bai, S., and Qi, X. (2022b). Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*.
- Henaff, O. (2020). Data-efficient image recognition with contrastive predictive coding. In *ICML*.
- Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-or, D. (2022). Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*.
- Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. (2018). Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*.
- Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., et al. (2022). Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.
- Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In *NeurIPS*.
- Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. (2022). Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.
- Hong, Y., Zhen, H., Chen, P., Zheng, S., Du, Y., Chen, Z., and Gan, C. (2023). 3d-llm: Injecting the 3d world into large language models. *arXiv preprint arXiv:2307.12981*.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Hu, R., Rohrbach, M., and Darrell, T. (2016). Segmentation from natural language expressions. In *European Conference on Computer Vision*, pages 108–124. Springer.
- Hu, R. and Singh, A. (2021a). Unit: Multimodal multitask learning with a unified transformer. In *ICCV*.
- Hu, R. and Singh, A. (2021b). Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1439–1449.

- Hu, W., Xu, Y., Li, Y., Li, W., Chen, Z., and Tu, Z. (2023). Bliva: A simple multimodal llm for better handling of text-rich visual questions. *arXiv preprint arXiv:2308.09936*.
- Huang, L., You, S., Zheng, M., Wang, F., Qian, C., and Yamasaki, T. (2022a). Green hierarchical vision transformer for masked image modeling. *NeurIPS*.
- Huang, R., Li, M., Yang, D., Shi, J., Chang, X., Ye, Z., Wu, Y., Hong, Z., Huang, J., Liu, J., et al. (2023a). Audiogpt: Understanding and generating speech, music, sound, and talking head. *arXiv preprint arXiv:2304.12995*.
- Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O. K., Liu, Q., et al. (2023b). Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*.
- Huang, S., Jiang, Z., Dong, H., Qiao, Y., Gao, P., and Li, H. (2023c). Instruct2act: Mapping multi-modality instructions to robotic actions with large language model. *arXiv preprint arXiv:2305.11176*.
- Huang, W., Abbeel, P., Pathak, D., and Mordatch, I. (2022b). Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pages 9118–9147. PMLR.
- Huang, Y., Meng, Z., Liu, F., Su, Y., Collier, N., and Lu, Y. (2023d). Sparkles: Unlocking chats across multiple images for multimodal instruction-following models. *arXiv preprint arXiv:2308.16463*.
- Huang, Z., Zeng, Z., Huang, Y., Liu, B., Fu, D., and Fu, J. (2021). Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *CVPR*.
- Huang, Z., Zeng, Z., Liu, B., Fu, D., and Fu, J. (2020). Pixel-BERT: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*.
- Huynh, D., Kuen, J., Lin, Z., Gu, J., and Elhamifar, E. (2022). Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7031.
- Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., and Schmidt, L. (2021). Openclip. If you use this software, please cite it as below.
- Jain, J., Li, J., Chiu, M. T., Hassani, A., Orlov, N., and Shi, H. (2023). Oneformer: One transformer to rule universal image segmentation. In *CVPR*.
- Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., and Makedon, F. (2020). A survey on contrastive self-supervised learning. *Technologies*.
- Jerripothula, K. R., Cai, J., and Yuan, J. (2016). Image co-segmentation via saliency co-fusion. *IEEE Transactions on Multimedia*, 18(9):1896–1909.
- Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y., Li, Z., and Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*.
- Jing, L. and Tian, Y. (2020). Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*.
- Joulin, A., Bach, F., and Ponce, J. (2010). Discriminative clustering for image co-segmentation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1943–1950. IEEE.
- Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., and Carion, N. (2021). Mdetr-modulated detection for end-to-end multi-modal understanding. In *ICCV*.
- Kang, M., Zhu, J.-Y., Zhang, R., Park, J., Shechtman, E., Paris, S., and Park, T. (2023). Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10124–10134.
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

- Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., and Irani, M. (2023). Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017.
- Kazemzadeh, S., Ordonez, V., Matten, M., and Berg, T. (2014). Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*.
- Kim, W., Son, B., and Kim, I. (2021). ViLT: Vision-and-language transformer without convolution or region supervision. In *ICML*.
- Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- Kirillov, A., He, K., Girshick, R., Rother, C., and Dollár, P. (2019). Panoptic segmentation. In *CVPR*.
- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023). Segment anything. *arXiv preprint arXiv:2304.02643*.
- Koh, J. Y., Fried, D., and Salakhutdinov, R. (2023). Generating images with multimodal language models. *arXiv preprint arXiv:2305.17216*.
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.
- Kokkinos, I. (2017). Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6129–6138.
- Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N. (2020). Big transfer (bit): General visual representation learning. In *ECCV*.
- Kolesnikov, A., Pinto, A. S., Beyer, L., Zhai, X., Harmsen, J., and Houlsby, N. (2022). Uvim: A unified modeling approach for vision with learned guiding codes. *arXiv preprint arXiv:2205.10337*.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *NeurIPS*.
- Kumari, N., Zhang, B., Zhang, R., Shechtman, E., and Zhu, J.-Y. (2023). Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941.
- Kuo, W., Bertsch, F., Li, W., Piergiovanni, A., Saffar, M., and Angelova, A. (2022). Findit: Generalized localization with natural language queries. In *ECCV*.
- Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., and Jia, J. (2023). Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*.
- Lamb, A., Dumoulin, V., and Courville, A. (2016). Discriminative regularization for generative models. *arXiv preprint arXiv:1602.03220*.
- Lambert, J., Liu, Z., Sener, O., Hays, J., and Koltun, V. (2020). Mseg: A composite dataset for multi-domain semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2879–2888.
- Larsen, A. B. L., Sønderby, S. K., Larochelle, H., and Winther, O. (2016). Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning*, pages 1558–1566. PMLR.
- Laurençon, H., Saulnier, L., Tronchon, L., Bekman, S., Singh, A., Lozhkov, A., Wang, T., Karamcheti, S., Rush, A. M., Kiela, D., et al. (2023). Obelisc: An open web-scale filtered dataset of interleaved image-text documents. *arXiv preprint arXiv:2306.16527*.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*.
- Li, B., Liu, H., Chen, L., Lee, Y. J., Li, C., and Liu, Z. (2023a). Benchmarking and analyzing generative data for visual recognition. *arXiv preprint arXiv:2307.13697*.
- Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., and Shan, Y. (2023b). Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

- Li, B., Weinberger, K. Q., Belongie, S., Koltun, V., and Ranftl, R. (2022a). Language-driven semantic segmentation. In *ICLR*.
- Li, B., Zhang, Y., Chen, L., Wang, J., Pu, F., Yang, J., Li, C., and Liu, Z. (2023c). Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*.
- Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., and Liu, Z. (2023d). Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*.
- Li, C., Liu, H., Li, L. H., Zhang, P., Aneja, J., Yang, J., Jin, P., Lee, Y. J., Hu, H., Liu, Z., et al. (2022b). Elevater: A benchmark and toolkit for evaluating language-augmented visual models. In *NeurIPS, Track on Datasets and Benchmarks*.
- Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., and Gao, J. (2023e). Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*.
- Li, C., Yang, J., Zhang, P., Gao, M., Xiao, B., Dai, X., Yuan, L., and Gao, J. (2021a). Efficient self-supervised vision transformers for representation learning. *arXiv preprint arXiv:2106.09785*.
- Li, F., Zhang, H., Sun, P., Zou, X., Liu, S., Yang, J., Li, C., Zhang, L., and Gao, J. (2023f). Semantic-sam: Segment and recognize anything at any granularity. *arXiv preprint arXiv:2307.04767*.
- Li, F., Zhang, H., Zhang, Y.-F., Liu, S., Guo, J., Ni, L. M., Zhang, P., and Zhang, L. (2022c). Vision-language intelligence: Tasks, representation learning, and large models. *arXiv preprint arXiv:2203.01922*.
- Li, G., Duan, N., Fang, Y., Gong, M., and Jiang, D. (2020a). Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*.
- Li, H., Zhu, J., Jiang, X., Zhu, X., Li, H., Yuan, C., Wang, X., Qiao, Y., Wang, X., Wang, W., et al. (2023g). Uni-perceiver v2: A generalist model for large-scale vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2691–2700.
- Li, J., Li, D., Savarese, S., and Hoi, S. (2023h). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.
- Li, J., Li, D., Xiong, C., and Hoi, S. (2022d). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.
- Li, J., Selvaraju, R. R., Gotmare, A. D., Joty, S., Xiong, C., and Hoi, S. (2021b). Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*.
- Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., and Qiao, Y. (2023i). Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.
- Li, L., Gan, Z., Cheng, Y., and Liu, J. (2019a). Relation-aware graph attention network for visual question answering. In *ICCV*.
- Li, L., Yin, Y., Li, S., Chen, L., Wang, P., Ren, S., Li, M., Yang, Y., Xu, J., Sun, X., et al. (2023j). M3it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*.
- Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., and Chang, K.-W. (2019b). VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.
- Li, L. H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.-N., et al. (2022e). Grounded language-image pre-training. *CVPR*.
- Li, L. H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.-N., et al. (2022f). Grounded language-image pre-training. In *CVPR*.
- Li, M., Song, F., Yu, B., Yu, H., Li, Z., Huang, F., and Li, Y. (2023k). Api-bank: A benchmark for tool-augmented llms. *arXiv preprint arXiv:2304.08244*.
- Li, S. and Tajbakhsh, N. (2023). Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs. *arXiv preprint arXiv:2308.03349*.
- Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., and Gao, J. (2020b). Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*.

- Li, X. L. and Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.
- Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., and Wen, J.-R. (2023l). Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.
- Li, Y., Fan, H., Hu, R., Feichtenhofer, C., and He, K. (2023m). Scaling language-image pre-training via masking. In *CVPR*.
- Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., and Yan, J. (2022g). Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *ICLR*.
- Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., and Lee, Y. J. (2023n). Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521.
- Li, Y., Zhang, C., Yu, G., Wang, Z., Fu, B., Lin, G., Shen, C., Chen, L., and Wei, Y. (2023o). Stablellava: Enhanced visual instruction tuning with synthesized image-dialogue data. *arXiv preprint arXiv:2308.10253*.
- Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., and Marculescu, D. (2023a). Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070.
- Liang, Y., Wu, C., Song, T., Wu, W., Xia, Y., Liu, Y., Ou, Y., Lu, S., Ji, L., Mao, S., et al. (2023b). Taskmatrix. ai: Completing tasks by connecting foundation models with millions of apis. *arXiv preprint arXiv:2303.16434*.
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *ECCV*.
- Lin, Y., Wu, H., Wang, R., Lu, H., Lin, X., Xiong, H., and Wang, L. (2023). Towards language-guided interactive 3d generation: Llms as layout interpreter with generative feedback. *arXiv preprint arXiv:2305.15808*.
- Liu, C., Lin, Z., Shen, X., Yang, J., Lu, X., and Yuille, A. (2017). Recurrent multimodal interaction for referring image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1271–1280.
- Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., and Wang, L. (2023a). Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*.
- Liu, H., Jiang, X., Li, X., Guo, A., Hu, Y., Jiang, D., and Ren, B. (2023b). The devil is in the frequency: Geminated gestalt autoencoder for self-supervised visual pre-training. In *AAAI*.
- Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023c). Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.
- Liu, H., Son, K., Yang, J., Liu, C., Gao, J., Lee, Y. J., and Li, C. (2023d). Learning customized visual models with retrieval-augmented knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Liu, J., Ding, H., Cai, Z., Zhang, Y., Satzoda, R. K., Mahadevan, V., and Manmatha, R. (2023e). Polyformer: Referring image segmentation as sequential polygon generation.
- Liu, N., Li, S., Du, Y., Torralba, A., and Tenenbaum, J. B. (2022a). Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer.
- Liu, S., Fan, L., Johns, E., Yu, Z., Xiao, C., and Anandkumar, A. (2023f). Prism: A vision-language model with an ensemble of experts.
- Liu, S., Ye, J., and Wang, X. (2023g). Any-to-any style transfer. *arXiv preprint arXiv:2304.09728*.
- Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al. (2023h). Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

- Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I* 14, pages 21–37. Springer.
- Liu, W. and Zuo, Y. (2023). Stone needle: A general multimodal large-scale model framework towards healthcare. *arXiv preprint arXiv:2306.16034*.
- Liu, X., Zhu, Z., Liu, H., Yuan, Y., Cui, M., Huang, Q., Liang, J., Cao, Y., Kong, Q., Plumbley, M. D., et al. (2023i). Wavjourney: Compositional audio creation with large language models. *arXiv preprint arXiv:2307.14335*.
- Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al. (2023j). Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.
- Liu, Y., Li, Z., Li, H., Yu, W., Huang, M., Peng, D., Liu, M., Chen, M., Li, C., Jin, L., et al. (2023k). On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Liu, Z., He, Y., Wang, W., Wang, W., Wang, Y., Chen, S., Zhang, Q., Yang, Y., Li, Q., Yu, J., et al. (2023l). Internchat: Solving vision-centric tasks by interacting with chatbots beyond language. *arXiv preprint arXiv:2305.05662*.
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*.
- Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022b). A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986.
- Long, A., Yin, W., Ajanthan, T., Nguyen, V., Purkait, P., Garg, R., Blair, A., Shen, C., and van den Hengel, A. (2022). Retrieval augmented classification for long-tail visual recognition. In *CVPR*.
- Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *CVPR*.
- Lu, C.-Z., Jin, X., Hou, Q., Liew, J. H., Cheng, M.-M., and Feng, J. (2023a). Delving deeper into data scaling in masked image modeling. *arXiv preprint arXiv:2305.15248*.
- Lu, J., Batra, D., Parikh, D., and Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*.
- Lu, J., Clark, C., Zellers, R., Mottaghi, R., and Kembhavi, A. (2022a). Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*.
- Lu, J., Goswami, V., Rohrbach, M., Parikh, D., and Lee, S. (2020). 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10437–10446.
- Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., and Kalyan, A. (2022b). Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*.
- Lu, P., Peng, B., Cheng, H., Galley, M., Chang, K.-W., Wu, Y. N., Zhu, S.-C., and Gao, J. (2023b). Chameleon: Plug-and-play compositional reasoning with large language models. *arXiv preprint arXiv:2304.09842*.
- Lu, Q., Kuen, J., Tiancheng, S., Jiuxiang, G., Weidong, G., Jiaya, J., Zhe, L., and Ming-Hsuan, Y. (2023c). High-quality entity segmentation. In *ICCV*.
- Lu, Y., Li, C., Liu, H., Yang, J., Gao, J., and Shen, Y. (2023d). An empirical study of scaling instruction-tuned large multimodal models. *arXiv preprint*.
- Lüdecke, T. and Ecker, A. (2022). Image segmentation using text and image prompts. In *CVPR*.
- Luo, G., Zhou, Y., Ren, T., Chen, S., Sun, X., and Ji, R. (2023a). Cheap and quick: Efficient vision-language instruction tuning for large language models. *arXiv preprint arXiv:2305.15023*.

- Luo, H., Bao, J., Wu, Y., He, X., and Li, T. (2023b). Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *International Conference on Machine Learning*, pages 23033–23044. PMLR.
- Luo, R., Zhao, Z., Yang, M., Dong, J., Qiu, M., Lu, P., Wang, T., and Wei, Z. (2023c). Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*.
- Luo, W., Xing, J., Milan, A., Zhang, X., Liu, W., and Kim, T.-K. (2021). Multiple object tracking: A literature review. *Artificial intelligence*, 293:103448.
- Ma, J. and Wang, B. (2023). Segment anything in medical images. *arXiv preprint arXiv:2304.12306*.
- Ma, Z., Hong, X., and Shangguan, Q. (2023). Can sam count anything? an empirical study on sam counting. *arXiv preprint arXiv:2304.10817*.
- Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A. L., and Murphy, K. (2016). Generation and comprehension of unambiguous object descriptions. In *CVPR*.
- Margffoy-Tuay, E., Pérez, J. C., Botero, E., and Arbeláez, P. (2018). Dynamic multimodal instance segmentation guided by natural language queries. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 630–645.
- Marino, K., Chen, X., Parikh, D., Gupta, A., and Rohrbach, M. (2021). Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *CVPR*.
- Mazumder, M., Banbury, C., Yao, X., Karlaš, B., Rojas, W. G., Diamos, S., Diamos, G., He, L., Kiela, D., Jurado, D., et al. (2022). Dataperf: Benchmarks for data-centric ai development. *arXiv preprint arXiv:2207.10062*.
- McGuinness, K. and O’connor, N. E. (2010). A comparative evaluation of interactive segmentation algorithms. *Pattern Recognition*, 43(2):434–444.
- Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. (2021). Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*.
- Mertan, A., Duff, D. J., and Unal, G. (2022). Single image depth estimation: An overview. *Digital Signal Processing*, 123:103441.
- Miech, A., Zhukov, D., Alayrac, J.-B., Tapaswi, M., Laptev, I., and Sivic, J. (2019). Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., Wang, X., Zhai, X., Kipf, T., and Houlsby, N. (2022). Simple open-vocabulary object detection with vision transformers.
- Misra, I. and Maaten, L. v. d. (2020). Self-supervised learning of pretext-invariant representations. In *CVPR*.
- Misra, I., Shrivastava, A., Gupta, A., and Hebert, M. (2016). Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3994–4003.
- Monajatipoor, M., Li, L. H., Rouhsedaghat, M., Yang, L. F., and Chang, K.-W. (2023). Metavl: Transferring in-context learning ability from language models to vision-language models. *arXiv preprint arXiv:2306.01311*.
- Moor, M., Huang, Q., Wu, S., Yasunaga, M., Zakka, C., Dalmia, Y., Reis, E. P., Rajpurkar, P., and Leskovec, J. (2023). Med-flamingo: a multimodal medical few-shot learner. *arXiv preprint arXiv:2307.15189*.
- Mortensen, E. N. and Barrett, W. A. (1998). Interactive segmentation with intelligent scissors. *Graphical models and image processing*, 60(5):349–384.
- Mottaghi, R., Chen, X., Liu, X., Cho, N.-G., Lee, S.-W., Fidler, S., Urtasun, R., and Yuille, A. (2014). The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898.

- Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., and Qie, X. (2023). T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*.
- Mu, N., Kirillov, A., Wagner, D., and Xie, S. (2021). Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*.
- Mu, Y., Zhang, Q., Hu, M., Wang, W., Ding, M., Jin, J., Wang, B., Dai, J., Qiao, Y., and Luo, P. (2023). Embodiedgpt: Vision-language pre-training via embodied chain of thought. *arXiv preprint arXiv:2305.15021*.
- Musgrave, K., Belongie, S., and Lim, S.-N. (2020). A metric learning reality check. In *ECCV*.
- Nagaraja, V. K., Morariu, V. I., and Davis, L. S. (2016). Modeling context between objects for referring expression understanding. In *ECCV*.
- Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., et al. (2021). Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.
- Nguyen, T., Ilharco, G., Wortsman, M., Oh, S., and Schmidt, L. (2022). Quality not quantity: On the interaction between dataset design and robustness of clip. *NeurIPS*.
- Ning, J., Li, C., Zhang, Z., Geng, Z., Dai, Q., He, K., and Hu, H. (2023). All in tokens: Unifying output space of visual tasks via soft token. *arXiv preprint arXiv:2301.02229*.
- Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
- Oord, A. v. d., Vinyals, O., and Kavukcuoglu, K. (2017). Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*.
- OpenAI (2022). ChatGPT. <https://openai.com/blog/chatgpt/>.
- OpenAI (2023a). GPT-4 technical report. <https://arxiv.org/abs/2303.08774>.
- OpenAI (2023b). Gpt-4 technical report.
- Oquab, M., Dariseti, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. (2023). Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.
- Ordonez, V., Kulkarni, G., and Berg, T. (2011). Im2text: Describing images using 1 million captioned photographs. In *NeurIPS*.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- Ozbulak, U., Lee, H. J., Boga, B., Anzaku, E. T., Park, H., Van Messem, A., De Neve, W., and Vankerschaver, J. (2023). Know your self-supervised learning: A survey on image-based generative and discriminative training. *arXiv preprint arXiv:2305.13689*.
- Paranjape, B., Lundberg, S., Singh, S., Hajishirzi, H., Zettlemoyer, L., and Ribeiro, M. T. (2023). Art: Automatic multi-step reasoning and tool-use for large language models. *arXiv preprint arXiv:2303.09014*.
- Patil, S. G., Zhang, T., Wang, X., and Gonzalez, J. E. (2023). Gorilla: Large language model connected with massive apis. *arXiv preprint arXiv:2305.15334*.
- Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., and Launay, J. (2023). The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.
- Peng, B., Li, C., He, P., Galley, M., and Gao, J. (2023a). Instruction tuning with GPT-4. *arXiv preprint arXiv:2304.03277*.
- Peng, Z., Dong, L., Bao, H., Ye, Q., and Wei, F. (2022a). Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*.
- Peng, Z., Dong, L., Bao, H., Ye, Q., and Wei, F. (2022b). A unified view of masked image modeling. *arXiv preprint arXiv:2210.10615*.

- Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., and Wei, F. (2023b). Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.
- Peters, M. E., Neumann, M., Logan IV, R. L., Schwartz, R., Joshi, V., Singh, S., and Smith, N. A. (2019). Knowledge enhanced contextual word representations. *arXiv preprint arXiv:1909.04164*.
- Pham, H., Dai, Z., Ghiasi, G., Liu, H., Yu, A. W., Luong, M.-T., Tan, M., and Le, Q. V. (2021). Combined scaling for zero-shot transfer learning. *arXiv preprint arXiv:2111.10050*.
- Pi, R., Gao, J., Diao, S., Pan, R., Dong, H., Zhang, J., Yao, L., Han, J., Xu, H., and Zhang, L. K. T. (2023). Detgpt: Detect what you need via reasoning. *arXiv preprint arXiv:2305.14167*.
- Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*.
- Pont-Tuset, J., Uijlings, J., Changpinyo, S., Soricut, R., and Ferrari, V. (2020). Connecting vision and language with localized narratives. In *ECCV*.
- Qian, C., Han, C., Fung, Y. R., Qin, Y., Liu, Z., and Ji, H. (2023). Creator: Disentangling abstract and concrete reasonings of large language models through tool creation. *arXiv preprint arXiv:2305.14318*.
- Qian, R., Li, Y., Xu, Z., Yang, M.-H., Belongie, S., and Cui, Y. (2022). Multimodal open-vocabulary video classification via pre-trained vision and language models. *arXiv preprint arXiv:2207.07646*.
- Qin, C., Zhang, S., Yu, N., Feng, Y., Yang, X., Zhou, Y., Wang, H., Niebles, J. C., Xiong, C., Savarese, S., et al. (2023a). Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*.
- Qin, J., Wu, J., Yan, P., Li, M., Yuxi, R., Xiao, X., Wang, Y., Wang, R., Wen, S., Pan, X., and Wang, X. (2023b). Freeseg: Unified, universal and open-vocabulary image segmentation.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *ICML*.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*.
- Rahman, S., Khan, S., and Barnes, N. (2020). Improved visual-semantic alignment for zero-shot object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11932–11939.
- Rajić, F., Ke, L., Tai, Y.-W., Tang, C.-K., Danelljan, M., and Yu, F. (2023). Segment anything meets point tracking. *arXiv preprint arXiv:2307.01197*.
- Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.
- Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. (2021a). Zero-Shot Text-to-Image Generation. In *ICML*.
- Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. (2021b). Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.
- Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., and Lu, J. (2022). Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091.
- Razavi, A., Van den Oord, A., and Vinyals, O. (2019). Generating diverse high-fidelity images with vqvae-2. In *NeurIPS*.
- Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.

- Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S. G., Novikov, A., Barth-Maron, G., Gimenez, M., Sulsky, Y., Kay, J., Springenberg, J. T., et al. (2022). A generalist agent. *arXiv preprint arXiv:2205.06175*.
- Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*.
- Ridnik, T., Ben-Baruch, E., Noy, A., and Zelnik-Manor, L. (2021). Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2021). High-resolution image synthesis with latent diffusion models.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *CVPR*.
- Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*, pages 234–241. Springer.
- Roy, S., Wald, T., Koehler, G., Rokuss, M. R., Disch, N., Holzschuh, J., Zimmerer, D., and Maier-Hein, K. H. (2023). Sam. md: Zero-shot medical image segmentation capabilities of the segment anything model. *arXiv preprint arXiv:2304.05396*.
- Ruan, L. and Jin, Q. (2022). Survey: Transformer based video-language pre-training. *AI Open*.
- Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. (2023). 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*, pages 22500–22510.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *IJCV*.
- Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., et al. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*.
- Sariyildiz, M. B., Perez, J., and Larlus, D. (2020). Learning visual representations with caption annotations. In *ECCV*.
- Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., and Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*.
- Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. (2022). Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*.
- Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. (2021). Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.
- Schenk, D., Khandelwal, A., Clark, C., Marino, K., and Mottaghi, R. (2022). A-okvqa: A benchmark for visual question answering using world knowledge. *arXiv preprint arXiv:2206.01718*.
- Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *ACL*.
- Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., and Sun, J. (2019). Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*.
- Shao, W., Hu, Y., Gao, P., Lei, M., Zhang, K., Meng, F., Xu, P., Huang, S., Li, H., Qiao, Y., et al. (2023). Tiny l1vm-ehub: Early multimodal experiments with bard. *arXiv preprint arXiv:2308.03729*.
- ShareGPT (2023). <https://sharegpt.com/>.
- Sharma, P., Ding, N., Goodman, S., and Soricut, R. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*.
- Shen, Q., Yang, X., and Wang, X. (2023a). Anything-3d: Towards single-view anything reconstruction in the wild. *arXiv preprint arXiv:2304.10261*.

- Shen, S., Li, C., Hu, X., Xie, Y., Yang, J., Zhang, P., Rohrbach, A., Gan, Z., Wang, L., Yuan, L., et al. (2022a). K-lite: Learning transferable visual models with external knowledge. In *NeurIPS*.
- Shen, S., Li, L. H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.-W., Yao, Z., and Keutzer, K. (2022b). How much can clip benefit vision-and-language tasks? In *ICLR*.
- Shen, Y., Song, K., Tan, X., Li, D., Lu, W., and Zhuang, Y. (2023b). Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*.
- Sheynin, S., Ashual, O., Polyak, A., Singer, U., Gafni, O., Nachmani, E., and Taigman, Y. (2022). Knn-diffusion: Image generation via large-scale retrieval. *arXiv preprint arXiv:2204.02849*.
- Shi, J., Xiong, W., Lin, Z., and Jung, H. J. (2023a). Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*.
- Shi, P., Qiu, J., Abaxi, S. M. D., Wei, H., Lo, F. P.-W., and Yuan, W. (2023b). Generalist vision foundation models for medical imaging: A case study of segment anything model on zero-shot medical segmentation. *Diagnostics*.
- Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., and Singh, S. (2020). Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.
- Sidorov, O., Hu, R., Rohrbach, M., and Singh, A. (2020). Textcaps: a dataset for image captioning with reading comprehension. In *ECCV*.
- Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. (2012). Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part V 12*, pages 746–760. Springer.
- Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al. (2022). Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*.
- Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., and Kiela, D. (2022a). Flava: A foundational language and vision alignment model. In *CVPR*.
- Singh, M., Duval, Q., Alwala, K. V., Fan, H., Aggarwal, V., Adcock, A., Joulin, A., Dollár, P., Feichtenhofer, C., Girshick, R., et al. (2023). The effectiveness of mae pre-pretraining for billion-scale pretraining. *arXiv preprint arXiv:2303.13496*.
- Singh, M., Gustafson, L., Adcock, A., de Freitas Reis, V., Gedik, B., Kosaraju, R. P., Mahajan, D., Girshick, R., Dollár, P., and Van Der Maaten, L. (2022b). Revisiting weakly supervised pre-training of visual perception models. In *CVPR*.
- Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR.
- Song, Y. and Ermon, S. (2020). Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448.
- Song, Y., Xiong, W., Zhu, D., Li, C., Wang, K., Tian, Y., and Li, S. (2023). Restgpt: Connecting large language models with real-world applications via restful apis. *arXiv preprint arXiv:2306.06624*.
- Srinivasan, K., Raman, K., Chen, J., Bendersky, M., and Najork, M. (2021). Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. *arXiv preprint arXiv:2103.01913*.
- Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., and Dai, J. (2019). VL-BERT: Pre-training of generic visual-linguistic representations. In *ICLR*.
- Su, Y., Lan, T., Li, H., Xu, J., Wang, Y., and Cai, D. (2023). Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*.
- Sun, C., Shrivastava, A., Singh, S., and Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*.
- Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., and Wang, X. (2023a). Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*.

- Sun, Y., Yang, Y., Peng, H., Shen, Y., Yang, Y., Hu, H., Qiu, L., and Koike, H. (2023b). Imagebrush: Learning visual in-context instructions for exemplar-based image manipulation. *arXiv preprint arXiv:2308.00906*.
- Sun, Y., Zhu, C., Zheng, S., Zhang, K., Shui, Z., Yu, X., Zhao, Y., Li, H., Zhang, Y., Zhao, R., et al. (2023c). Pathasst: Redefining pathology through generative foundation ai assistant for pathology. *arXiv preprint arXiv:2305.15072*.
- Surís, D., Menon, S., and Vondrick, C. (2023). Vipergrpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*.
- Tan, H. and Bansal, M. (2019). LXMERT: Learning cross-modality encoder representations from transformers. In *EMNLP*.
- Tang, L., Xiao, H., and Li, B. (2023a). Can sam segment anything? when sam meets camouflaged object detection. *arXiv preprint arXiv:2304.04709*.
- Tang, Z., Yang, Z., Zhu, C., Zeng, M., and Bansal, M. (2023b). Any-to-any generation via composable diffusion.
- Tao, C., Zhu, X., Su, W., Huang, G., Li, B., Zhou, J., Qiao, Y., Wang, X., and Dai, J. (2023). Siamese image modeling for self-supervised vision representation learning. In *CVPR*.
- Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. (2023). Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.
- Team, M. N. (2023). Introducing mpt-7b: A new standard for open-source, ly usable llms. Accessed: 2023-03-28.
- Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., and Li, L.-J. (2016). Yfcc100m: The new data in multimedia research. *Communications of the ACM*.
- Tian, Y., Krishnan, D., and Isola, P. (2020a). Contrastive multiview coding. In *ECCV*.
- Tian, Z., Shen, C., and Chen, H. (2020b). Conditional convolutions for instance segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 282–298. Springer.
- Tong, Z., Song, Y., Wang, J., and Wang, L. (2022). Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *NeurIPS*.
- Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In *ICML*.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Tschannen, M., Kumar, M., Steiner, A., Zhai, X., Houlsby, N., and Beyer, L. (2023). Image captioners are scalable vision learners too. *arXiv preprint arXiv:2306.07915*.
- Tu, T., Azizi, S., Driess, D., Schaeckermann, M., Amin, M., Chang, P.-C., Carroll, A., Lau, C., Tanno, R., Ktena, I., et al. (2023). Towards generalist biomedical ai. *arXiv preprint arXiv:2307.14334*.
- Vahdat, A. and Kautz, J. (2020). Nvae: A deep hierarchical variational autoencoder.
- van den Oord, A., Vinyals, O., and Kavukcuoglu, K. (2017). Neural discrete representation learning. In *NeurIPS*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *NeurIPS*.
- Vicuna (2023). Vicuna: An open-source chatbot impressing GPT-4 with 90%* chatgpt quality. <https://vicuna.lmsys.org/>.
- Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *CVPR*.
- Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2016). Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):652–663.

- Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, volume 1, pages I–I. Ieee.
- Wang, B. and Komatsuzaki, A. (2021). GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. <https://github.com/kingoflolz/mesh-transformer-jax>.
- Wang, B., Wu, F., Han, X., Peng, J., Zhong, H., Zhang, P., Dong, X., Li, W., Li, W., Wang, J., et al. (2023a). Vigc: Visual instruction generation and correction. *arXiv preprint arXiv:2308.12714*.
- Wang, F., Kong, T., Zhang, R., Liu, H., and Li, H. (2023b). Self-supervised learning by estimating twin class distribution. *TIP*.
- Wang, H., Wang, Y., Zhou, Z., Ji, X., Li, Z., Gong, D., Zhou, J., and Liu, W. (2018). Cosface: Large margin cosine loss for deep face recognition. *CVPR*.
- Wang, H., Zhu, Y., Adam, H., Yuille, A., and Chen, L.-C. (2021a). Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5463–5474.
- Wang, J., Chen, D., Luo, C., Dai, X., Yuan, L., Wu, Z., and Jiang, Y.-G. (2023c). Chatvideo: A tracklet-centric multimodal and versatile video understanding system. *arXiv preprint arXiv:2304.14407*.
- Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., and Wang, L. (2022a). Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*.
- Wang, J., Zhou, Y., Xu, G., Shi, P., Zhao, C., Xu, H., Ye, Q., Yan, M., Zhang, J., Zhu, J., et al. (2023d). Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*.
- Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., and Qiao, Y. (2023e). Video-mae v2: Scaling video masked autoencoders with dual masking. In *CVPR*.
- Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., and Yang, H. (2022b). Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*.
- Wang, R., Chen, D., Wu, Z., Chen, Y., Dai, X., Liu, M., Jiang, Y.-G., Zhou, L., and Yuan, L. (2022c). Bevt: Bert pretraining of video transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14733–14743.
- Wang, T., Li, L., Lin, K., Lin, C.-C., Yang, Z., Zhang, H., Liu, Z., and Wang, L. (2023f). Disco: Disentangled control for referring human dance generation in real world. *arXiv preprint arXiv:2307.00040*.
- Wang, T., Zhang, J., Fei, J., Ge, Y., Zheng, H., Tang, Y., Li, Z., Gao, M., Zhao, S., Shan, Y., et al. (2023g). Caption anything: Interactive image description with diverse multimodal controls. *arXiv preprint arXiv:2305.02677*.
- Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O. K., Singhal, S., Som, S., and Wei, F. (2022d). Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*.
- Wang, W., Bao, H., Dong, L., and Wei, F. (2021b). Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*.
- Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., et al. (2023h). VisionLLM: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*.
- Wang, X., Wang, W., Cao, Y., Shen, C., and Huang, T. (2023i). 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*, pages 6830–6839.
- Wang, X., Zhang, X., Cao, Y., Wang, W., Shen, C., and Huang, T. (2023j). Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*.
- Wang, Y., Ivison, H., Dasigi, P., Hessel, J., Khot, T., Chandu, K. R., Wadden, D., MacMillan, K., Smith, N. A., Beltagy, I., et al. (2023k). How far can camels go? exploring the state of instruction tuning on open resources. *arXiv preprint arXiv:2306.04751*.

- Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. (2022e). Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.
- Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Arunkumar, A., Ashok, A., Dhanasekaran, A. S., Naik, A., Stap, D., et al. (2022f). Benchmarking generalization via in-context instructions on 1,600+ language tasks. *arXiv preprint arXiv:2204.07705*.
- Wang, Z., Chen, J., and Hoi, S. C. (2020). Deep learning for image super-resolution: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3365–3387.
- Wang, Z., Huang, H., Zhao, Y., Zhang, Z., and Zhao, Z. (2023l). Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes. *arXiv preprint arXiv:2308.08769*.
- Wang, Z., Jiang, Y., Lu, Y., Shen, Y., He, P., Chen, W., Wang, Z., and Zhou, M. (2023m). In-context learning unlocked for diffusion models.
- Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., and Cao, Y. (2022g). Simvlm: Simple visual language model pretraining with weak supervision. In *ICLR*.
- Weers, F., Shankar, V., Katharopoulos, A., Yang, Y., and Gunter, T. (2023). Masked autoencoding does not help natural language supervision at scale. In *CVPR*.
- Wei, C., Fan, H., Xie, S., Wu, C.-Y., Yuille, A., and Feichtenhofer, C. (2021). Masked feature prediction for self-supervised visual pre-training. In *CVPR*.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., and Zhou, D. (2022a). Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.
- Wei, L., Xie, L., Zhou, W., Li, H., and Tian, Q. (2022b). Mvp: Multimodality-guided visual pre-training. In *ECCV*.
- Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., and Zuo, W. (2023). Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*.
- Weng, L. (2023). Llm-powered autonomous agents. *lilianweng.github.io*.
- Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., and Duan, N. (2023a). Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.
- Wu, C., Zhang, X., Zhang, Y., Wang, Y., and Xie, W. (2023b). Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*.
- Wu, J., Jiang, Y., Sun, P., Yuan, Z., and Luo, P. (2022a). Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4984.
- Wu, J., Lu, J., Sabharwal, A., and Mottaghi, R. (2021). Multi-modal answer validation for knowledge-based VQA. *arXiv preprint arXiv:2103.12248*.
- Wu, J., Wang, J., Yang, Z., Gan, Z., Liu, Z., Yuan, J., and Wang, L. (2022b). Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*.
- Wu, S., Fei, H., Qu, L., Ji, W., and Chua, T.-S. (2023c). Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519.
- Wu, W., Timofeev, A., Chen, C., Zhang, B., Duan, K., Liu, S., Zheng, Y., Shlens, J., Du, X., Gan, Z., et al. (2023d). Mofi: Learning image representations from noisy entity annotated images. *arXiv preprint arXiv:2306.07952*.
- Wu, Y., Lim, J., and Yang, M.-H. (2013). Online object tracking: A benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2411–2418.
- Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. (2018). Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*.
- Xian, Y., Lampert, C. H., Schiele, B., and Akata, Z. (2018). Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *TPAMI*.
- Xiao, Z., Chen, Y., Zhang, L., Yao, J., Wu, Z., Yu, X., Pan, Y., Zhao, L., Ma, C., Liu, X., et al. (2023). Instruction-vit: Multi-modal prompts for instruction learning in vit. *arXiv preprint arXiv:2305.00201*.

- Xie, D., Wang, R., Ma, J., Chen, C., Lu, H., Yang, D., Shi, F., and Lin, X. (2023a). Edit everything: A text-guided generative system for images editing. *arXiv preprint arXiv:2304.14006*.
- Xie, X., Fu, L., Zhang, Z., Wang, Z., and Bai, X. (2022a). Toward understanding wordart: Corner-guided transformer for scene text recognition.
- Xie, Z., Lin, Y., Yao, Z., Zhang, Z., Dai, Q., Cao, Y., and Hu, H. (2021). Self-supervised learning with swin transformers. *arXiv preprint arXiv:2105.04553*.
- Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., and Hu, H. (2022b). Simmim: A simple framework for masked image modeling. In *CVPR*.
- Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Wei, Y., Dai, Q., and Hu, H. (2023b). On data scaling in masked image modeling. In *CVPR*.
- Xu, D., Ouyang, W., Wang, X., and Sebe, N. (2018). Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 675–684.
- Xu, H., Yan, M., Li, C., Bi, B., Huang, S., Xiao, W., and Huang, F. (2021). E2e-vlp: End-to-end vision-language pre-training enhanced by visual learning.
- Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., and Wang, X. (2022a). Groupvit: Semantic segmentation emerges from text supervision. In *CVPR*.
- Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., and De Mello, S. (2023a). Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966.
- Xu, P., Shao, W., Zhang, K., Gao, P., Liu, S., Lei, M., Meng, F., Huang, S., Qiao, Y., and Luo, P. (2023b). Lvilm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*.
- Xu, R., Wang, X., Wang, T., Chen, Y., Pang, J., and Lin, D. (2023c). Pointllm: Empowering large language models to understand point clouds. *arXiv preprint arXiv:2308.16911*.
- Xu, Z., Shen, Y., and Huang, L. (2022b). Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. *arXiv preprint arXiv:2212.10773*.
- Yan, B., Jiang, Y., Wu, J., Wang, D., Luo, P., Yuan, Z., and Lu, H. (2023). Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15325–15336.
- Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., and Wen, F. (2023a). Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391.
- Yang, J., Gao, M., Li, Z., Gao, S., Wang, F., and Zheng, F. (2023b). Track anything: Segment anything meets videos. *arXiv preprint arXiv:2304.11968*.
- Yang, J., Li, C., Zhang, P., Xiao, B., Liu, C., Yuan, L., and Gao, J. (2022a). Unified contrastive learning in image-text-label space. In *CVPR*.
- Yang, J., Li, C., Zhang, P., Xiao, B., Yuan, L., Liu, C., and Gao, J. (2022b). Unicl: Unified contrastive learning in image-text-label space. In *CVPR*.
- Yang, R., Song, L., Li, Y., Zhao, S., Ge, Y., Li, X., and Shan, Y. (2023c). Gpt4tools: Teaching large language model to use tools via self-instruction. *arXiv preprint arXiv:2305.18752*.
- Yang, Z., Gan, Z., Wang, J., Hu, X., Ahmed, F., Liu, Z., Lu, Y., and Wang, L. (2021). Crossing the format boundary of text and boxes: Towards unified vision-language modeling. *arXiv preprint arXiv:2111.12085*.
- Yang, Z., Gan, Z., Wang, J., Hu, X., Ahmed, F., Liu, Z., Lu, Y., and Wang, L. (2022c). Unitab: Unifying text and box outputs for grounded vision-language modeling. In *European Conference on Computer Vision*, pages 521–539. Springer.
- Yang, Z., Gan, Z., Wang, J., Hu, X., Lu, Y., Liu, Z., and Wang, L. (2022d). An empirical study of gpt-3 for few-shot knowledge-based vqa. In *AAAI*.
- Yang*, Z., Li*, L., Wang*, J., Lin*, K., Azarnasab*, E., Ahmed*, F., Liu, Z., Liu, C., Zeng, M., and Wang, L. (2023). Mm-react: Prompting chatgpt for multimodal reasoning and action.

- Yang, Z., Ping, W., Liu, Z., Korthikanti, V., Nie, W., Huang, D.-A., Fan, L., Yu, Z., Lan, S., Li, B., et al. (2023a). Re-vilm: Retrieval-augmented visual language model for zero and few-shot image captioning. *arXiv preprint arXiv:2302.04858*.
- Yang, Z., Wang, J., Gan, Z., Li, L., Lin, K., Wu, C., Duan, N., Liu, Z., Liu, C., Zeng, M., et al. (2023b). Reco: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14246–14255.
- Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., and Torr, P. H. (2022e). Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165.
- Yao, L., Han, J., Liang, X., Xu, D., Zhang, W., Li, Z., and Xu, H. (2023). Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23497–23506.
- Yao, L., Han, J., Wen, Y., Liang, X., Xu, D., Zhang, W., Li, Z., Xu, C., and Xu, H. (2022a). Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *arXiv preprint arXiv:2209.09407*.
- Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., and Xu, C. (2022b). Filip: Fine-grained interactive language-image pre-training. In *ICLR*.
- Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. (2022c). React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.
- Yasunaga, M., Aghajanyan, A., Shi, W., James, R., Leskovec, J., Liang, P., Lewis, M., Zettlemoyer, L., and Yih, W.-t. (2022). Retrieval-augmented multimodal language modeling. *arXiv preprint arXiv:2211.12561*.
- Ye, J., Hu, A., Xu, H., Ye, Q., Yan, M., Dan, Y., Zhao, C., Xu, G., Li, C., Tian, J., et al. (2023a). mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*.
- Ye, L., Rochan, M., Liu, Z., and Wang, Y. (2019a). Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10502–10511.
- Ye, M., Zhang, X., Yuen, P. C., and Chang, S.-F. (2019b). Unsupervised embedding learning via invariant and spreading instance feature. In *CVPR*.
- Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al. (2023b). mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.
- Yi, K., Ge, Y., Li, X., Yang, S., Li, D., Wu, J., Shan, Y., and Qie, X. (2022). Masked image modeling with denoising contrast. *arXiv preprint arXiv:2205.09616*.
- Yilmaz, A., Javed, O., and Shah, M. (2006). Object tracking: A survey. *Acm computing surveys (CSUR)*, 38(4):13–es.
- Yin, D., Dong, L., Cheng, H., Liu, X., Chang, K.-W., Wei, F., and Gao, J. (2022). A survey of knowledge-intensive nlp with pre-trained language models. *arXiv preprint arXiv:2202.08772*.
- Yin, Z., Wang, J., Cao, J., Shi, Z., Liu, D., Li, M., Sheng, L., Bai, L., Huang, X., Wang, Z., et al. (2023). Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *arXiv preprint arXiv:2306.06687*.
- Yoon, S., Kang, W. Y., Jeon, S., Lee, S., Han, C., Park, J., and Kim, E.-S. (2021). Image-to-image retrieval by learning similarity between scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10718–10726.
- Yu, J., Li, X., Koh, J. Y., Zhang, H., Pang, R., Qin, J., Ku, A., Xu, Y., Baldridge, J., and Wu, Y. (2021). Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*.
- Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. (2022a). Coca: Contrastive captioners are image-text foundation models. *TMLR*.
- Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B. K., et al. (2022b). Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*.

- Yu, L. and et al (2023). Scaling autoregressive multi-modal models: Pretraining and instruction tuning.
- Yu, L., Poirson, P., Yang, S., Berg, A. C., and Berg, T. L. (2016). Modeling context in referring expressions. In *ECCV*.
- Yu, Q., He, J., Deng, X., Shen, X., and Chen, L.-C. (2023a). Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *arXiv preprint arXiv:2308.02487*.
- Yu, Q., Li, J., Ye, W., Tang, S., and Zhuang, Y. (2023b). Interactive data synthesis for systematic vision adaptation via llms-aiges collaboration. *arXiv preprint arXiv:2305.12799*.
- Yu, T., Feng, R., Feng, R., Liu, J., Jin, X., Zeng, W., and Chen, Z. (2023c). Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*.
- Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., and Wang, L. (2023d). Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.
- Yu, Z., Yu, J., Cui, Y., Tao, D., and Tian, Q. (2019). Deep modular co-attention networks for visual question answering. In *CVPR*.
- Yuan, L., Chen, D., Chen, Y.-L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., Liu, C., Liu, M., Liu, Z., Lu, Y., Shi, Y., Wang, L., Wang, J., Xiao, B., Xiao, Z., Yang, J., Zeng, M., Zhou, L., and Zhang, P. (2021). Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.
- Zamir, A. R., Sax, A., Shen, W., Guibas, L. J., Malik, J., and Savarese, S. (2018). Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722.
- Zang, Y., Li, W., Han, J., Zhou, K., and Loy, C. C. (2023). Contextual object detection with multi-modal large language models. *arXiv preprint arXiv:2305.18279*.
- Zang, Y., Li, W., Zhou, K., Huang, C., and Loy, C. C. (2022). Open-vocabulary detr with conditional matching. *arXiv preprint arXiv:2203.11876*.
- Zareian, A., Rosa, K. D., Hu, D. H., and Chang, S.-F. (2021). Open-vocabulary object detection using captions. In *CVPR*.
- Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. (2021). Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*.
- Zellers, R., Bisk, Y., Farhadi, A., and Choi, Y. (2019). From recognition to cognition: Visual commonsense reasoning. In *CVPR*.
- Zeng, Y., Lin, Z., Zhang, J., Liu, Q., Collomosse, J., Kuen, J., and Patel, V. M. (2023). Scenecomposer: Any-level semantic image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22468–22478.
- Zeng, Y., Zhang, X., and Li, H. (2022). Multi-grained vision language pre-training: Aligning texts with visual concepts. In *ICML*.
- Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. (2022a). Scaling vision transformers. In *CVPR*.
- Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. (2023). Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*.
- Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., and Beyer, L. (2022b). Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*.
- Zhang, C., Liu, L., Cui, Y., Huang, G., Lin, W., Yang, Y., and Hu, Y. (2023a). A comprehensive survey on segment anything model for vision and beyond. *arXiv preprint arXiv:2305.08196*.
- Zhang, C., Yang, Z., He, X., and Deng, L. (2020). Multimodal intelligence: Representation learning, information fusion, and applications. *JSTSP*.
- Zhang, C., Zhang, C., Zhang, M., and Kweon, I. S. (2023b). Text-to-image diffusion model in generative ai: A survey. *arXiv preprint arXiv:2303.07909*.
- Zhang, C., Zheng, S., Li, C., Qiao, Y., Kang, T., Shan, X., Zhang, C., Qin, C., Rameau, F., Bae, S.-H., et al. (2023c). A survey on segment anything model (sam): Vision foundation model meets prompt engineering. *arXiv preprint arXiv:2306.06211*.

- Zhang, D., Li, S., Zhang, X., Zhan, J., Wang, P., Zhou, Y., and Qiu, X. (2023d). Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.
- Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L. M., and Shum, H.-Y. (2022a). Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.
- Zhang, H., Li, F., Zou, X., Liu, S., Li, C., Gao, J., Yang, J., and Zhang, L. (2023e). A simple framework for open-vocabulary segmentation and detection. *arXiv preprint arXiv:2303.08131*.
- Zhang, H., Li, X., and Bing, L. (2023f). Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.
- Zhang, H., Zhang, P., Hu, X., Chen, Y.-C., Li, L. H., Dai, X., Wang, L., Yuan, L., Hwang, J.-N., and Gao, J. (2022b). Glipv2: Unifying localization and vision-language understanding. In *ECCV*.
- Zhang, J., Huang, J., Jin, S., and Lu, S. (2023g). Vision-language models for vision tasks: A survey. *arXiv preprint arXiv:2304.00685*.
- Zhang, J., Shen, F., Liu, L., Zhu, F., Yu, M., Shao, L., Shen, H. T., and Van Gool, L. (2018). Generative domain-migration hashing for sketch-to-image retrieval. In *Proceedings of the European conference on computer vision (ECCV)*, pages 297–314.
- Zhang, L. and Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*.
- Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., and Gao, J. (2021). VinVL: Revisiting visual representations in vision-language models. In *CVPR*.
- Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Dong, H., Gao, P., and Li, H. (2023h). Personalize segment anything model with one shot. *arXiv preprint arXiv:2305.03048*.
- Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., Li, J., Hu, R., Zhang, T., Wu, F., et al. (2023i). Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.
- Zhang, S., Gong, C., Wu, L., Liu, X., and Zhou, M. (2023j). Automl-gpt: Automatic machine learning with gpt. *arXiv preprint arXiv:2305.02499*.
- Zhang, S., Sun, P., Chen, S., Xiao, M., Shao, W., Zhang, W., Chen, K., and Luo, P. (2023k). Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*.
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.
- Zhang, W., Aljunied, S. M., Gao, C., Chia, Y. K., and Bing, L. (2023l). M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *arXiv preprint arXiv:2306.05179*.
- Zhang, X., Chen, J., Yuan, J., Chen, Q., Wang, J., Wang, X., Han, S., Chen, X., Pi, J., Yao, K., et al. (2022c). Cae v2: Context autoencoder with clip target. *arXiv preprint arXiv:2211.09799*.
- Zhang, X., Tian, Y., Huang, W., Ye, Q., Dai, Q., Xie, L., and Tian, Q. (2022d). Hivit: Hierarchical vision transformer meets masked image modeling. *arXiv preprint arXiv:2205.14949*.
- Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., and Xie, W. (2023m). Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.
- Zhang, Y., Huang, X., Ma, J., Li, Z., Luo, Z., Xie, Y., Qin, Y., Luo, T., Li, Y., Liu, S., et al. (2023n). Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*.
- Zhang, Y. and Jiao, R. (2023). How segment anything model (sam) boost medical image segmentation? *arXiv preprint arXiv:2305.03678*.
- Zhang, Y., Zhang, R., Gu, J., Zhou, Y., Lipka, N., Yang, D., and Sun, T. (2023o). Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*.
- Zhao, B., Wu, B., and Huang, T. (2023a). Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*.
- Zhao, S., Chen, D., Chen, Y.-C., Bao, J., Hao, S., Yuan, L., and Wong, K.-Y. K. (2023b). Uni-controlnet: All-in-one control to text-to-image diffusion models. *arXiv preprint arXiv:2305.16322*.

- Zhao, Y., Lin, Z., Zhou, D., Huang, Z., Feng, J., and Kang, B. (2023c). Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581*.
- Zhao, Y., Pang, T., Du, C., Yang, X., Li, C., Cheung, N.-M., and Lin, M. (2023d). On evaluating adversarial robustness of large vision-language models. *arXiv preprint arXiv:2305.16934*.
- Zhao, Z., Guo, L., Yue, T., Chen, S., Shao, S., Zhu, X., Yuan, Z., and Liu, J. (2023e). Chatbridge: Bridging modalities with large language model as a language catalyst. *arXiv preprint arXiv:2305.16103*.
- Zhao, Z., Wallace, E., Feng, S., Klein, D., and Singh, S. (2021). Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.
- Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L. H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al. (2022a). Regionclip: Region-based language-image pretraining. In *CVPR*.
- Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L. H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al. (2022b). Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803.
- Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., and Torralba, A. (2017). Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641.
- Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L., et al. (2023a). Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.
- Zhou, C., Loy, C. C., and Dai, B. (2022a). Extract free dense labels from clip. In *ECCV*.
- Zhou, G., Hong, Y., and Wu, Q. (2023b). Navgpt: Explicit reasoning in vision-and-language navigation with large language models. *arXiv preprint arXiv:2305.16986*.
- Zhou, J., Dong, L., Gan, Z., Wang, L., and Wei, F. (2023c). Non-contrastive learning meets language-image pre-training. In *CVPR*.
- Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., and Kong, T. (2021). ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*.
- Zhou, T., Zhang, Y., Zhou, Y., Wu, Y., and Gong, C. (2023d). Can sam segment polyps? *arXiv preprint arXiv:2304.07583*.
- Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., and Misra, I. (2022b). Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer.
- Zhou, Y., Li, C., Chen, C., Gao, J., and Xu, J. (2022c). Lafite2: Few-shot text-to-image generation. *arXiv preprint arXiv:2210.14124*.
- Zhou, Y. and Shimada, N. (2023). Vision + language applications: A survey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 826–842.
- Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. (2023a). Minigpt-4: Enhancing vision-language understanding with advanced large language models.
- Zhu, P., Wang, H., and Saligrama, V. (2019). Zero shot detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(4):998–1010.
- Zhu, P., Wang, H., and Saligrama, V. (2020). Don’t even look once: Synthesizing features for zero-shot detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11693–11702.
- Zhu, W., Hessel, J., Awadalla, A., Gadre, S. Y., Dodge, J., Fang, A., Yu, Y., Schmidt, L., Wang, W. Y., and Choi, Y. (2023b). Multimodal c4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*.
- Zong, Z., Song, G., and Liu, Y. (2023). Detrs with collaborative hybrid assignments training.
- Zou, X., Dou, Z.-Y., Yang, J., Gan, Z., Li, L., Li, C., Dai, X., Behl, H., Wang, J., Yuan, L., et al. (2023a). Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Zou, X., Liu, H., and Lee, Y. J. (2022). End-to-end instance edge detection. *arXiv preprint arXiv:2204.02898*.

Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Gao, J., and Lee, Y. J. (2023b). Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*.