Foundation Models Defining a New Era in Vision: A Survey and Outlook

Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, *Life Fellow, IEEE*, Ming-Hsuan Yang, *Fellow, IEEE*, and Fahad Shahbaz Khan

(Survey paper)

***Abstract*—Vision systems that see and reason about the compositional nature of visual scenes are fundamental to understanding our world. The complex relations between objects and their locations, ambiguities, and variations in the real-world environment can be better described in human language, naturally governed by grammatical rules and other modalities such as audio and depth. The models learned to bridge the gap between such modalities and large-scale training data facilitate contextual reasoning, generalization, and prompt capabilities at test time. These models are referred to as *foundation models*. The output of such models can be modified through human-provided prompts without retraining, e.g., segmenting a particular object by providing a bounding box, having interactive dialogues by asking questions about an image or video scene or manipulating the robot's behavior through language instructions. In this survey, we provide a comprehensive review of such emerging foundation models, including typical architecture designs to combine different modalities (vision, text, audio, etc.), training objectives (contrastive, generative), pre-training datasets, fine-tuning mechanisms, and the common prompting patterns; textual, visual, and heterogeneous. We discuss the open challenges and research directions for foundation models in computer vision, including difficulties in their evaluations and benchmarking, gaps in their real-world understanding, limitations of contextual understanding, biases, vulnerability to adversarial attacks, and interpretability issues. We review recent developments in this field, covering a wide range of applications of foundation models systematically and comprehensively.**

***Index Terms*—Contrastive learning, language and vision, large language models, masked modeling, self-supervised learning.**

I. Introduction

RECENT years have witnessed remarkable success towards developing *foundation models*, that are trained on a large-scale broad data. Once trained, they operate as a basis and can be adapted (e.g., fine-tuned) to a wide range of downstream tasks related to the originally trained model [1]. While the essential ingredients of the foundation models, such as deep neural networks and self-supervised learning, have been around for many years, the recent surge, specifically through large language models (LLMs), can be mainly attributed to massively scaling up both data and model size [2]. For instance, recent models with billion parameters such as GPT-3 [3] have been effectively utilized for zero/few-shot learning, achieving impressive performance without requiring large-scale task-specific data or model parameter updating. Similarly, the recent 540-billion parameter Pathways Language Model (PaLM) has demonstrated state-of-the-art capabilities on numerous challenging problems ranging from language understanding and generation to reasoning and code-related tasks [4], [5].

Concurrent to LLMs in natural language processing, large foundation models for different perception tasks have also recently been explored in the literature. For instance, pre-trained vision-language models (VL) such as CLIP [6] have demonstrated promising zero-shot performance on different downstream vision tasks, including image classification and object detection. These VL foundation models are typically trained using millions of image-text pairs collected from the web and provide representations with generalization and transfer capabilities. These pre-trained VL foundation models can then be adapted to a downstream task by presenting a natural description of the given task and prompts. For instance, the seminal CLIP model utilizes carefully designed prompts to operate on different downstream tasks, including zero-shot classification, where the text encoder dynamically constructs the classifiers via class names or other free-form texts. Here, the textual prompts are handcrafted templates, e.g., `A photo of a {label}`," that aid in specifying the text as corresponding to the visual image

Received 2 October 2023; revised 2 October 2024; accepted 7 November 2024. Date of publication 9 January 2025; date of current version 6 March 2025. Recommended for acceptance by A. Kovashka. *(Corresponding author: Muhammad Awais.)*

Muhammad Awais is with the MBZ University of AI, Abu Dhabi, UAE, and also with the Computer Science Department, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: iawaisrauf@gmail.com).

Muzammal Naseer is with the Computer Science Department and Center of Secure Cyber-Physical Security Systems, Khalifa University, Abu Dhabi, UAE, and also with the CECS, Australian National University, Canberra, ACT 0200, Australia.

Salman Khan is with the MBZ University of AI, Abu Dhabi, UAE, and also with the CECS, Australian National University, Canberra ACT 0200, Australia.

Rao Muhammad Anwer and Hisham Cholakkal are with the MBZ University of AI, Abu Dhabi, UAE.

Mubarak Shah is with the Center for Research in Computer Vision, University of Central Florida, Orlando, FL 32816 USA.

Ming-Hsuan Yang is with the University of California, Merced, CA 95344 USA, also with the Yonsei University, Seoul 03722, South Korea, and also with the Google Research, Mountain View, CA 94043 USA.

Fahad Shahbaz Khan is with the MBZ University of AI, Abu Dhabi, UAE, and also with the Computer Vision Laboratory, Linköping University, 581 83 Linkoping, Sweden.

A comprehensive list of foundation models studied in this work is available at https://github.com/awaisrauf/Awesome-CV-Foundational-Models; awaisrauf.github.io

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2024.3506283, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2024.3506283

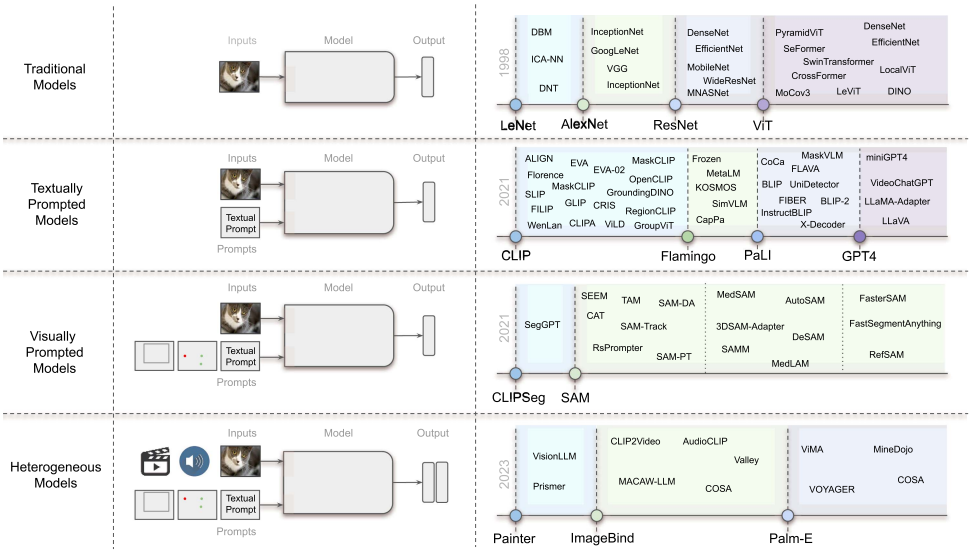

Fig. 1. Overview of the evolution of foundation models in computer vision. (left) We show the progression of models in computer vision, starting from traditional single modality models with a pre-determined number of outputs to textually prompted, visually prompted, and heterogenous models. (right) We show the evolution of these models with major milestones reported in the literature, as shown by dotted lines.

content. Recently, numerous works have also explored adding conversational capabilities to the VL models by fine-tuning them on a specific instruction set [7], [8], [9], [10].

Besides large VL foundation models, several research efforts have been devoted to developing large foundation models that visual inputs can prompt. For instance, the recently introduced SAM [11] performs a class-agnostic segmentation given an image and a visual prompt such as box, point, or mask, which specifies what to segment in an input. Such a model is trained on billions of object masks following a model-in-the-loop (semi-automated) dataset annotation setting. Further, such a generic visual prompt-based segmentation model can be adapted for specific downstream tasks such as medical image segmentation [12], [13], video object segmentation [14], robotics [15], and remote sensing [16]. In addition to textual and visual prompt-based foundation models, research works have explored developing models that strive to align multiple paired modalities (e.g., image-text, video-audio, or image-depth) to learn meaningful representations helpful for different downstream tasks [17], [18], [19].

This work presents a systematic review of foundation models (Fig. 1) in computer vision. As depicted in Fig 1, we distinguish these models according to their input modalities. This categorization criterion allows a seamless depiction of these models' evolution and progress while facilitating meaningful grouping. Based on this categorization, models are divided into textually prompted (Section III–III-D), visually prompted (Section IV), heterogeneous modality-based (Section V-B) and embodied foundation models (Section V-C). Within the textually prompted foundation models, we further distinguish them into contrastive, generative, hybrid (contrastive and generative), and conversational VL models. Finally, we discuss open challenges and research directions based on our analysis (Section VI). In addition, we provide essential background, comprehensive preliminaries, and extended discussions for each section of Appendix (Section A). Next, we review other surveys related to ours and discuss the differences and uniqueness.

Related Reviews and Differences: In the literature, few recent works have reviewed large language models (LLMs) in natural language processing [2], [20], [21], [22], [23]. Zhao et al. [2] review recent advances in LLMs, distinguishing different aspects of LLMs such as pre-training, adaptation tuning, LLM utilization, and evaluation. This survey also summarizes resources available to develop LLMs and discusses potential future directions. In the context of VLMs, the work of [24] performs a preliminary review of vision-language pre-trained models regarding task definition and general architecture. Similarly, [25] discusses different techniques to encode images and texts to embeddings before the pre-training step and reviews different pre-training architectures. The work of [26] reviews transformer techniques for multimodal data with a survey of vanilla transformers, vision transformers, and multimodal transformers from a geometrically topological perspective. In the context of multimodal learning, the recent review [27] focuses on self-supervised multimodal learning techniques to effectively utilize supervision from raw multimodal data. The survey distinguishes existing approaches based on objective functions, data alignment, and architectures. The work of [28], [29] summarizes different vision-language pre-training network architectures, objectives, and downstream tasks and categorizes vision-language pre-training frameworks. Similarly, [30] reviews the utility of vision-language models across a wide range of vision tasks, including classification, segmentation, and action recognition. Along the same lines, [31] offers a systematic review of visual instruction tuning for designing models capable of following arbitrary instructions thereby solving general vision tasks. Recently, the work of [32] reviews the visually prompted foundation segmentation model, segmenting anything, and discusses its potential downstream tasks.

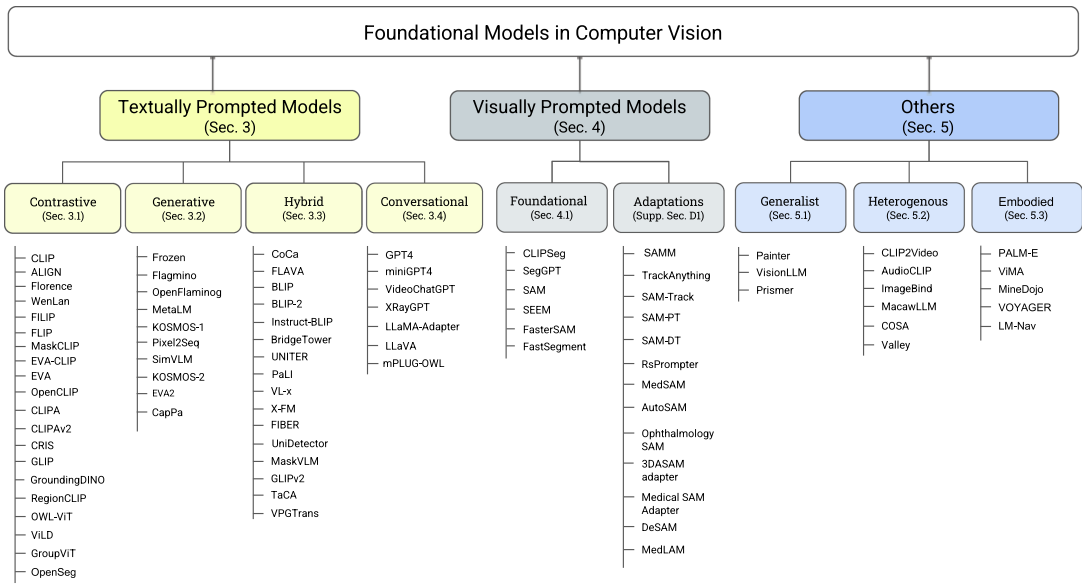

Fig. 2. Overview of our taxonomy for vision-language foundation models. We categorize these foundation models into five main groups based on their inputs, outputs, and utilization.

The main differences between this survey and the works mentioned above are as follows. Unlike previous surveys that primarily focus on textual prompt-based vision-language models, our work focuses on the three different classes of foundation models (Fig. 2): textually prompted models (contrastive, generative, hybrid, and conversational), visually prompted models (e.g., SegGPT [33], SAM [11]) and heterogeneous modalities-based models (e.g., ImageBind [17], Valley [34]). Our work provides a comprehensive and up-to-date overview of the recent vision foundation models (Sections III, IV, V-B, and V-C). Finally, we present a detailed discussion on open challenges and potential research directions of foundation models in computer vision (Section VI).

II. Essential Background

We briefly discuss the scope of our survey and foundational concepts essential for understanding the rest of the paper. We provide more discussions in the Appendix.

Foundational Models and Scope of the Survey: The term "foundational model" was introduced by [1] at Stanford Institute for Human-Centered AI. Foundational models are defined as "the base models trained on large-scale data in a self-supervised or semi-supervised manner that can be adapted for several other downstream tasks." The paradigm shift towards foundational models is significant as it allows the replacement of several narrow task-specific models with broader and generic base models that can be once trained and quickly adapted for multiple applications. It not only enables rapid model development and provides better performance for both in-domain and out-domain scenarios but also leads to the "emergent properties" of intelligence from large-scale foundational models trained on massive datasets [35], [36].

Computer vision has recently witnessed significant progress fueled by foundational models [11], [37] with an extensive body of literature encompassing discriminative and generative models. In this survey, we focus on multimodal (vision and language) foundational models trained on large-scale data that can be adapted for several computer vision tasks involving non-image outputs (e.g., generated text and segmentation masks). Note that we do not cover image generative models aimed at model data distribution such as GANs, VAEs, and Diffusion models owing to dedicated surveys already existing in this area [38], [39], [40], [41] and because the former model class can cover a broader range of downstream applications.

Architecture Types: As depicted in Fig. 3, Vision-Language models primarily use four architectural designs. We first introduce the Dual-Encoder architecture, wherein separate encoders process visual and textual modalities. The output of these encoders is subsequently optimized through an objective function. The second architecture type, fusion, incorporates an additional fusion encoder, which takes the representations generated by the vision and text encoders and learns fused representations. The third type, Encoder-Decoder, consists of an encoder-decoder-based language model and a visual encoder. Lastly, the fourth architecture type, Adapted LLM, leverages a Large Language Model (LLM) as its core component, with a visual encoder employed to convert images into a format compatible with the LLM. For a more comprehensive understanding of these architectures, we refer the readers to the corresponding survey sections where each work is discussed. Next, we discuss the loss functions used to train different architectures.

Training Objectives: In the pre-training of foundational models, the aim is to learn useful representations from image-text data. Broadly, this objective is achieved by two types of losses: contrastive and generative. The contrastive objectives transform the learning into matching image-text data. For instance, to learn from unlabeled image-text data, [42], [43] utilize a simple in Image-Text Contrastive (**ITC**) loss, which aims to learn representations by learning to predict correct image-text

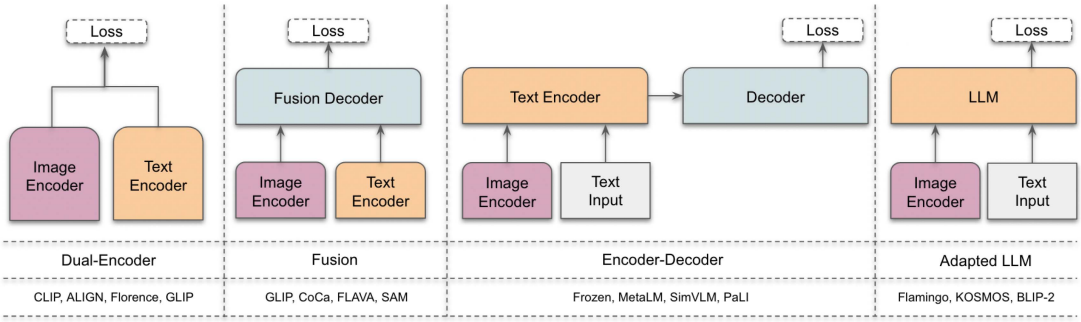

Fig. 3. Overview of four different architecture styles described in this survey. *Left* to *right:* (a) Dual-encoder (b) Fusion (c) Encoder-decoder (d) Adapter LLM. Examples for each category are shown in the bottom row. Appendix (Section A) presents more details about these architectures.

TABLE I
OVERVIEW OF DIFFERENT SETTINGS UNDER WHICH DATASETS ARE UTILIZED FOR TRAINING, FINE-TUNING, AND PROMPTING IN FOUNDATIONAL MODELS

Data Type	Examples
Pre-training	
Image-Text	WIT [42], LAION [46], [47]
w/ Pseudo Labels	Cap24M [48], SA-1B [11]
Benchmark Combination	UNITER [49], PMD [50]
Fine-tuning	
Task Specific	ImageNet [51]
Capability Specific	OWL-ViT [52]
Instruction-Following	InstructBLIP [9]
Prompt Engineering	
Train Time	GLIP [48]
Evaluation Time	CLIP [42]

More details are discussed in appendix sec A.

pairs. Similarly, Image-Text Matching (**ITM**) loss [44] aims to correctly predict whether a pair of images and text is positively or negatively matched. Generative objectives turn the problem into conditional generation. Standard Captioning (**Cap**) loss [45] also aims to predict the next token given previous tokens and images. We provide a more detailed description of various objectives used in the Appendix.

Large-scale Training: Large-scale training followed by effective prompting at inference has been a crucial ingredient of vision and language foundational models. We discuss the role of pre-training, fine-tuning, and prompting techniques here (see Table I).

Pre-training Data: Large-scale data is at the heart of modern vision-language foundational models. The datasets that have been utilized to pre-train these models can be divided into three broad categories: image-text datasets (such as WebImageText used in CLIP [42]), partially synthetic datasets (such as SA-1B used in SAM [11]), and mixture datasets (such as PMD used in FLAVA [50]). For further details, refer to Section A of the Appendix.

Fine-tuning: Fine-tuning is employed under three primary settings: to improve a model's performance on a specific task (e.g., open-world object detection), to improve a model for a particular capability (e.g., visual grounding), and to instruction-tune a model to make it solve different downstream vision tasks (e.g., InstructBLIP [9]).

Prompt Engineering: Prompt engineering has primarily been used with Large Language Models (LLMs) to make them do certain tasks [3], [53]. In the context of vision-language models or visually-prompted models, prompt engineering is predominately used for two purposes: to convert vision datasets to image-text training data (e.g., CLIP for image classification) to provide human intractability to the foundational models and use vision-language models for vision tasks.

III. TEXTUALLY PROMPTED MODELS

In this section, we explore textually prompted models, which are categorized into four types based on their training objectives and learning mechanism: contrastive (Section III-A), generative (Section III-B), hybrid (Section III-C) and conversational models (Section III-D). For an overview of these methods, refer to Fig. 3 and Table II. We relegate a detailed discussion on textually prompted models for visual-grounding tasks to Appendix B.

A. Contrastive Learning (CL)

Contrastive foundation methods, e.g., CLIP [42], have evolved with improved datasets, better architectures, modified training recipes, expanded utility, open-source reproduction, and study of scaling laws. We divide contrastive learning methods into general purpose and visual grounding-based foundation models.

Methods Solely based on CL: Radford et al. [42] introduce the **CLIP** model, which involves joint training of an image encoder (ViT or scaled CNN) and a text encoder (GPT-like transformer) through a contrastive pre-training task. This task focuses on correctly pairing images with their captions in a batch. The model generates multi-modal embeddings for N image-text pairs and is trained to maximize the cosine similarity between embeddings of N correct pairs while minimizing the similarity between embeddings of $N^2 - N$ incorrect pairs using symmetric cross-entropy loss. This work shows the significance of the natural language supervision data scale and curates a dataset of 400 million image-text pairs from the internet for large-scale model training. The CLIP framework exhibits impressive performance on many datasets, demonstrating good zero-shot generalization, higher robustness to natural and synthetic distribution shifts, and compatibility with linear probe-based fine-tuning.

TABLE II
TEXTUALLY PROMPTED MODELS

Method	Data		Pretraining Objectives			Architecture		Information	
	Public	Size	Contrastive	Generative	Others	Type	Base	Link	Venue
CLIP [42]	✗	400M	ITC	-	-	Dual-Enc	ResNet [83], ViT [84], GPT2 [6]	Link	arXiv'21
ALIGN [43]	✗	1800M	ITC	-	-	Dual-Enc	EffNet-L2 [85], BERT-Large [86]	-	ICML'21
WenLan [62]	✗	30M	InfoCE	-	-	Dual-Enc	RoBERTAa-Large[1], Faster-RCNN [87], EffNet-B7 [85]	Link	arXiv'23
Florence [37]	✗	900M	UniCL	-	-	Dual-Enc	CoSwnT [57], GPT2 [6]	-	ECCV'22
FILIP [64]	✗	340M	FILIP	-	-	Dual-Enc	ViT [84], GPT2 [6]	-	ICLR'22
SLIP [58]	✓	15M	ITC, SimCLR	-	-	Dual-Enc	ViT [84], ViT-S [88], GPT2 [6]	Link	ECCV'22
FLIP [66]	✓	400M	ITC	-	-	Dual-Enc	ViT [84], Transformer [89]	Link	arXiv'23
MaskCLIP [67]	✓	15M	ITC	MLM	Distil	Dual-Enc	ViT [89], GPT2 [6]	Link	CVPR'23
CLIPA [75]	✓	400M	ITC	-	-	Dual-Enc	ViT [84], Transformer [89]	Link	arXiv'23
CLIPAv2 [76]	✓	3000M	ITC	-	-	Dual-Enc	ViT [84], Transformer [89]	Link	arXiv'23
EVA [73]	✓	29.6M	ITC	-	-	Dual-Enc	ViT-G [90], BEiT-3 [91]	Link	CVPR'23
EVA-CLIP [70]	✓	2000M	ITC	-	-	Dual-Enc	ViT-G [90], BEiT-3 [91]	Link	arXiv'23
EVA-02 [92]	✓	2000M	ITC	-	-	Enc-Dec	TrV [92], -	Link	arXiv'23
OpenCLIP [74]	✓	5400M	ITC	-	-	Dual-Enc	ViT [84], GPT2 [6]	Link	CVPR'23
CRIS [93]	✓	0.4M	TPC	-	-	Fusion	ResNet [83], GPT2 [6]	Link	CVPR'22
MaskCLIP [94]	✓	0.17M	-	-	Task	Dual-Enc	ResNet [83], ViT [84], GPT2 [6]	Link	ECCV'22
GLIP [48]	✓	-	RWA	-	-	Fusion	ViT [84], GPT2 [6]	Link	CVPR'22
G-DINO [95]	✓	-	GLIP	-	Task	Fusion	Swin-{T, L} [96], BERT-base [69]	Link	arXiv'23
OWL-ViT [52]	✓	2M	-	-	DETR	Dual-Enc	ViT [84], Transformer [89]	Link	ECCV'22
ViLD [97]	✓	-	-	-	Task	Dual-Enc	Mask-RCNN [98], FPN [99], CLIP-GPT2 [42]	Link	ICLR'22
GroupViT [100]	✓	26M	ITC, MITC	-	-	Dual-Enc	ViT [89], [88], GPT2 [6]	Link	CVPR'22
OpenSeg [101]	✓	0.77M	-	-	Task	Dual-Enc	EffNet-B7 [85], ResNet [83], BERT-Large [86]	Link	ECCV'22
Frozen [102]	✓	3M	-	Cap	-	Enc-Dec	NF-ResNet [103], GPT2 [6]	-	NeurIPS'21
Flamingo [104]	✗	43M	-	Flamingo	-	AdaptedLLm	NF-ResNet [103], Chinchilla [105]	-	NeurIPS'22
OpenFlamingo [106]	✓	2571M	-	Flamingo	-	AdaptedLLm	CLIPL-ViT [6], LLMs	Link	github'23
MetaLM [107]	✓	10M	-	SemiCasualLM	-	Enc-Dec	ViT [84], GPT2 [6]	-	arXiv'23
KOSMOS-1 [108]	✓	3115M	-	SemiCasualLM	-	AdaptedLLm	CLIP-ViT [42], MAG-NETO [109]	-	arXiv'23
KOSMOS-v2[110]	-	115M	-	SemiCasualLM	-	AdaptedLLm	CLIP-ViT [42], MAG-NETO [109]	Link	arXiv'23
SimVLM [111]	✓	1800M	-	PrefixLM	-	Enc-Dec	ViT [84], BERT [86]	-	ICLR'22
MaskVLM [112]	✓	4M	ITC, ITM	MVLM	-	Fusion	ViT [84], RoBERTa [113]	-	ICLR23
mPLUG-OWL [10]	✓	1100+M	-	LM	-	AdaptedLLm	CLIP-ViT [42], LLaMA-7B [114]	Link	arXiv'23
CapPa [115]	✗	12B	-	Cap, CapPa	-	Dual-Enc	-	-	arXiv'23
UNITER [49]	✓	9.5M	ITM	MLM	-	Fusion	Faster-RCNN [87], BERT [86]	Link	ECCV'20
Pixel2Seqv2 [116]	✓	0.12M	-	VLM	-	Enc-Dec	ViT-B [84], Transformer [89]	-	NeurIPS'22
VL-x [117]	✓	9.18M	ITM	MLM	Task	Enc-Dec	BEiTv2 [118], RoBERTa [113]	-	ICML'21
CoCa [119]	✗	4800M	NSL	Cap	-	Fusion	ViT [84], Transformer [89]	-	TMLR'23
FLAVA [50]	✓	70M	ITC,ITM	MMM,MIM,MLM	-	Fusion	ViT [84], Transformer [89]	Link	CVPR'22
PaLI [120]	✗	1600M	-	-	-	Enc-Dc	ViT-e [120], mT5 [121]	-	arXiv'23
BLIP [122]	✓	129M	ITC,ITM	LM	-	Fusion	ViT [84], BERT [86]	Link	arXiv'22
BridgeTower [123]	✓	4M	ITM	MLM	-	Fusion	CLIP-ViT [42], RoBERTa [113]	Link	AAAI'23
X-FM [124]	✓	20M	ITC,ITM	MLM,MIM,IMLM	BBP	Fusion	BEiTv2 [91], RoBERTa [113]	Link	arXiv'23
BLIP-2 [125]	✓	129M	ITC,ITM	ITG	-	AdaptedLLm	CLIP-ViT [42], EVA-CLIP [73], OPT [126], FlanT5 [127]	Link	arXiv'23
Instruct-BLIP [9]	✓	-	-	LM	-	AdaptedLLm	EVA-ViT [70], LLM [128], [127][2]	Link	arXiv'23
TaCA [129]	✓	400M	ITC	-	Distil	-	CLIP-ViT [6], GPT2 [6]	Link	arXiv'23
VPGTrans [130]	✓	1.4M	-	-	Sim	-	BLIP-2	Link	arXiv'23
FIBER [131]	✓	4.8M	ITC,ITM	MLM	Task	Fusion	Swin [96], RoBERTa [113]	Link	NeurIPS'22
UniDetector [132]	✓	-	-	-	Task	Dual-Enc	RegionCLIP [133]	Link	CVPR'23
X-Decoder [134]	✓	4.07M	ITC	Cap	Task	Fusion	Focal-T [135], DaViT [136], GPT2 [6]	Link	CVPR'23
GLIPv2 [137]	✓	-	RWC	-	Task	Fusion	Swin-T [96], Transformer [89]	Link	NeurIPS'22

We present different properties of these models that contrast them, including pre-training dataset and model size, pre-training objective, architecture, publication venue, and online information. More details are discussed in the appendix (section A).

The visual-language dataset used by Radford et al. [42] requires complex and computationally expensive pre-processing and cleaning, limiting its scale. To avoid these steps, Jia et al. in **ALIGN** [43] gather one billion noisy image-caption pairs from Conceptual Captions Dataset [54]. A dual encoder architecture with CLIP-like normalized contrastive objectives is trained on this dataset. To align visual and language embeddings, cosine similarity of image-text embeddings are optimized through normalized softmax loss [55]. This work shows that the dataset's scale compensates for its noisy nature. The aligned image-text representations exhibit state-of-the-art performance for cross-modal matching/retrieval tasks and zero-shot classification.

Yuan et al. [37] argue that a genuine foundation model should serve tasks involving different space, time, and modality aspects. Specifically, a foundation model should be able to handle representation from coarse to fine (Space), static to dynamic (Time), and from RGB to multi-modalities (Modality). To achieve such generalizability, the **Florence** model is developed based on CLIP-like pre-training on the large, curated dataset and improved contrastive objective and efficient training. A pre-trained model is then extended to have three different adapter heads for each space. The Dynamic DETR-based adapter learns representation for fine-grained dense tasks with large-scale object detection datasets. Similarly, a METER [56] head is used for vision language representation, and CSwin [57] is used for video-based understanding. This framework results in a foundation model that generalizes across domains. Mu et al. [58] explore the potential synergy between image-based self-supervised learning

and language supervision. To this end, **SLIP** adds an adaptation of SimCLR [59], [60] loss for self-supervision based on different views or augmentation of the input image. They show that a CLIP-like model trained on the YFCC15M dataset [61] with the SLIP method performs better compared with similar models trained with language supervision or self-supervision alone. This performance improvement is demonstrated across several tasks, including zero-shot and linear probe-based classification problems.

Existing text-image-based methods typically assume strong semantic correlations between the data pair. However, web-scale data is littered with pairs that have weak correlations (e.g., captions that do not reflect images accurately). Huo et al. [62] propose **WenLan** to address this issue by using two-tower architecture and cross-modal contrastive learning based on MoCo [63], which can leverage more negative samples in limited GPU resources. Specifically, this approach leverages negative and positive examples and text-to-image and image-to-text-based contrastive losses. This results in a better model that is also efficient in training and shows improved performance on downstream tasks. In addition, a large-scale Chinese image-text dataset with 500 million data points is constructed.

CLIP-style methods use separate encoders for each modality, ensuring efficient inference by allowing decoupled encoders and pre-computed representations. However, these models often emphasize global features for cross-modal interaction, neglecting finer-grained interplay between modalities. To address this, Yao et al. [64] introduce **FILIP** (Fine-grained Interactive Language Image Pre-training) with cross-modal late interaction to capture token-wise semantic alignment. FILIP maximizes the similarity between token-wise visual and text embeddings. It calculates the similarity of each visual token against all text tokens and for each text token. The resulting similarities are averaged to compute the loss. Notably, FILIP maintains fine-grained modality interaction while preserving CLIP's inference efficiency. In addition, a vast dataset of 340 million image-text pairs is also constructed for model training. FILIP performs better than CLIP and other approaches in zero-shot classification and image-text retrieval tasks.

Masked Contrastive Learning: Motivated by the success of Masked Auto Encoders [65], Li et al. [66] introduce an efficient variant of CLIP known as **FLIP**. This approach involves masking 50-75% of input pixels during CLIP training, leading to a 2-4$\times$ reduction in computation, enabling larger batches, and enhancing accuracy. FLIP achieves equivalent accuracy to CLIP in over three times less training time, saving approximately 1800 TPU days. This work further explores scaling across different models, dataset sizes, and training durations using this accelerated method. Similarly, albeit with a different motivation, Dong et al. [67] argue that the language description of an image cannot express complete information as an image is a continuous and fine-grained signal. To fully leverage images in contrastive vision-language training, they propose **MaskCLIP** that randomly masks the input image and performs mean teacher-based self-distillation [68] to learn local semantic features. Specifically, the representation of the whole image and masked image are obtained from the mean teacher and student, respectively, and cross-entropy loss between the two models is minimized. Similarly, BERT [69] pre-training is used in the language encoder. These two modifications in the contrastive learning framework facilitate the model learning local and fine-grained semantics. MaskCLIP significantly improves CLIP in zero-shot, linear probe, and fine-tuning settings on several vision datasets.

While the above-discussed approaches mainly focus on the efficiency aspect of CLIP via masking, **EVA-CLIP** addresses instability and optimization efficiency together with masking visual inputs. Specifically, Sun et al. [70] present a method to improve training stability and reduce computational costs, including improved initialization, better optimizer, and random masking of images [66]. The proposed EVA-CLIP model is trained on an open-source dataset (Merged-2B dataset consisting of LAION-2B [47] and COYO-700M [71]). It performs well compared with similar budget OpenCLIP models [72] for a wide variety of tasks such as ImageNet classification, zero-shot classification, and image retrieval. Fang et al. [73] scale this model to one billion parameters by masking out image-text inputs and using the CLIP loss. The scaled EVA model performs favorably on several downstream tasks, including COCO, LVIS, and ImageNet1k.

Scaling and Reproducing CLIP: While the pre-trained models and weights of CLIP [42] are released, the training mechanism and dataset are not publicly available. To increase the accessibility, several subsequent works open-source large-scale image-text datasets, reproduce CLIP, and study its properties. The state-of-the-art performance of CLIP hinges on the large-scale image-text dataset, which is not publicly available. To address this problem, Schuhmann et al. [46] release an image-text dataset, LAION-400M, consisting of 400 million data points curated after filtering Common Crawl. In a subsequent effort, Schuhmann et al. [47] further scale it up and release a multilingual, multi-modal dataset called LAION-5B, which contains 5.8 billion data points curated from Common Crawl after filtering through the existing CLIP model. Utilizing large-scale LAION datasets [46], [47], **Open-CLIP** [72] train and reproduce CLIP training experiments and analyze its properties. Cherti et al. [74] supplement this open-source effort by studying the scaling laws of CLIP. The LAION-5B [47] trained OpenCLIP demonstrates consistent improvements in performance as data, model, and compute are scaled. The scaling behavior observed in OpenCLIP differs from the one seen in CLIP [42]. This contrast could be attributed to differences in the training data of CLIP (private WebImageText dataset) and OpenCLIP (publicly available LAION dataset).

Improving the performance of CLIP requires larger models that require higher computing resources. Li et al. [75] demonstrate that larger image-text models can be trained effectively with shorter sequences of input tokens without substantial performance degradation. Based on this *inverse scaling law*, they introduce a new efficient training recipe and CLIP-like model trained on academic-scale resources, i.e., **CLIPA**. CLIPA can achieve 63.2%, 67.8%, and 69.3% zero-shot ImageNet accuracy in two, three, and four days of training on 8 A100 GPUs, respectively. Building on the *inverse scaling law* observed by

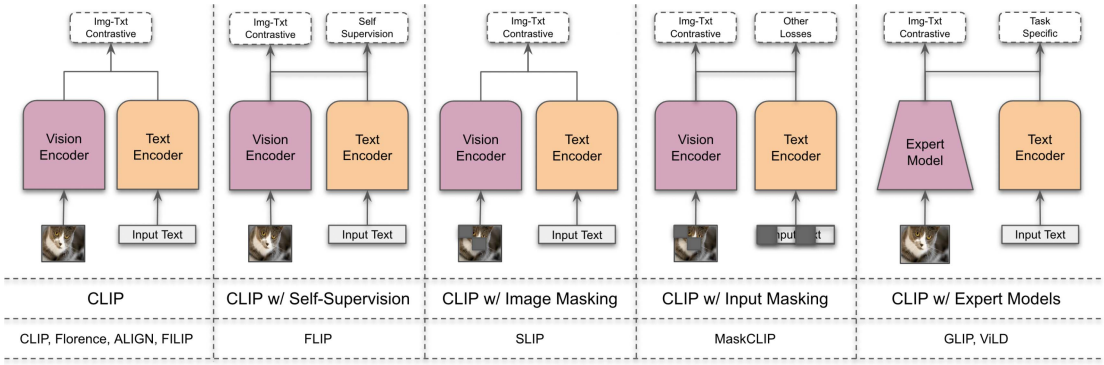

Fig. 4. Overview of CLIP and its variants. Several works after CLIP analyze the efficacy of different variants, including new losses, image-based self-supervision, masking of inputs, and expert models. Here, an expert model means a model that is pre-trained for a specific task.

CLIPA [75], Li et al. [76] train CLIP-like model on a large scale with significantly less computational budget and training costs. With their large-scale training, they demonstrate two interesting results. First, the inverse scaling law applies to fine-tuning: models can be fine-tuned on fewer input tokens. Second, larger models exhibit a smaller performance drop compared to smaller models when fine-tuned with the same number of input tokens. The trained CLIPA model achieves 69.3% zero-shot ImageNet classification accuracy in four training days on 8 A100 GPUs. Finally, several works have also explored CLIP (Fig. 4) and contrastive methods from different perspectives, such as remote sensing [77], visual document understanding [78] and healthcare [79].

Discussion: After the success of CLIP, the contrastive framework has shown impressive performance in a range of vision and vision-language tasks, with an extensive exploration into various aspects of the framework, such as data and training efficiency and adaptations for dense prediction tasks. This exploration has facilitated groundbreaking applications, including solving diverse multimodal tasks and integrating vision capabilities into LLMs. However, the processing of large-resolution input images remains under-explored, which restricts the potential applications of these models, as they can handle small details in large images. Moreover, recent studies have investigated issues of fairness and robustness in CLIP-like foundational models [80], [81], [82]. Given the extensive utility of these models across a wide variety of applications, a comprehensive understanding of these challenges is essential. Further, since pre-training these models demands substantial resources, designing effective post-training robustness and debiasing methods can bring significant advantages.

B. Generative Learning

Large-scale Language Models (LMs) demonstrate intriguing zero and few-shot performance in NLP tasks, and only recently have multimodal models emerged that combine vision and language. Contrastive vision-language models show promising generalizations but are limited in their problem-solving scope since they only offer text-image similarity scores. Here, we discuss efforts to empower LLMs with visual perception through training on vision-conditioned language generation tasks.

In-context Learning with Multimodal Inputs: Here, we explain methods that endow LMs with visual modality using interleaved image-text data. Tsimpoukelli et al. [102] propose **Frozen**, an efficient approach to add visual modality in the LMs without updating their weights. Frozen consists of an image encoder that encodes input images to the word embedding space of LMs such that these LMs can generate image captions. LM is kept frozen to learn joint embeddings, and the vision encoder is trained on captioning datasets with the task of the conditional generation of caption given an image. Although Frozen is trained on single image-text pairs, it can operate with an ordered set of multiple image-text pairs, enabling it to do few-shot tasks. The LM and vision encoder are prompted with the ordered textual and visual prompts during inference. The textual and visual embeddings are concatenated and fed to the LM decoder, which generates a textual output autoregressive. Frozen demonstrates few-shot visual-language capabilities across vision-language tasks.

Like Frozen, Alayrac et al. [104] build models that can adapt to new tasks using only a few examples. The family of **Flamingo** models leverages fixed pre-trained vision and language models along with a Perceiver Resampler-based bridge. The Perceiver Resampler connects the visual encoder to LM by producing a set number of visual tokens. These visual tokens condition LM's output using gated cross-attention dense blocks interleaved between LM's layers. These layers provide an efficient way for the LM to incorporate visual information. They are trained with next-token prediction tasks conditioned on preceding text and a set of images or videos. Flamingo models can handle large image and video inputs since the Perceiver Resample converts varying-sized visual inputs to a few visual tokens. During inference, interleaved support examples (image, text) or (video, text) are followed by a query visual input fed to the model for computation. Despite requiring significantly fewer annotated examples, Flamingo models outperform state-of-the-art fine-tuned methods for several few-shot vision-language tasks. **OpenFlamingo** [106] is an open-source version of the Flamingo model based on the same architecture but with a new Multimodal C4 dataset and 10M samples from LAION-2B. Their models utilize LLaMA-7B and CLIP's visual encoder and achieve 80% performance of the respective Flamingo model.

Language Models (LMs) as a General Interface for other Modalities: Several works propose to utilize language models

as a universal task layer and integrate other modalities into it through the pre-trained encoders. Here, we describe these methods. Hao et al. [107] develop **MetaLM**. This semi-casual model consists of a unidirectional transformer decoder and multiple bi-directional encoders connected with the decoder through the connector layers. MetaLM shows the casual language models' excellent zero and few-shot capabilities and better transferability of non-casual encoders [86], [138]. Jointly training encoders and decoders on a new semi-casual language modeling objective facilitates MetaLM learning to generate the next word given the previous tokens and encoded representations. This joint framework inherits in-context learning, instruction-following, and fine-tuning abilities. To understand the capabilities of MetaLM, authors perform a multitude of experiments. On NLP tasks, it outperforms GPT on numerous multi-task fine-tuning, single-task fine-tuning, instruction-tuned zero-shot, and in-context learning. Likewise, it shows improved performance compared with established baselines [49], [117] for zero-shot generalization, in-context learning, and fine-tuning.

Like MetaLM [107], Huang et al. [108] align visual inputs with LMs to create models that handle inputs of multiple modalities. The **KOSMOS-1** model consists of a Magneto-based Large Language Model (LLM) [109] as a general interface and xPos [139] encoders to encode different modalities. This model is trained on web-scale data consisting of text corpus, image-caption pairs, and interleaved image-caption pairs to generate the next token given context. Training data also consists of several language-only instruction-tuning datasets, which are also dealt with as language modeling tasks to further align it with the human interface. To demonstrate the capabilities of KOSMOS-1, a large set of experiments are performed across NLP, vision, cross-modal transfer, non-verbal reasoning, and vision-language tasks. These experimental results show the general-purpose nature of LLMs.

Peng et al. [110] extend KOSMOS-1 [108] for grounding capabilities in **KOSMOS-2**. The main difference with KOSMOS-1 is a different pipeline to extract text spans (i.e., noun phrases and referring expressions) and link them to corresponding regions in the images. This pipeline consists of two steps. First, non-chunks are extracted from the text and linked with regions in the image based on a pre-trained detector. Second, noun chunks are expanded to referring expressions by traversing nouns' dependency tree. Based on this pipeline, they curated GRIT (GRounded Image-Text pairs) from COYO-700M [71] and LIAON-2B [47] consisting of 91M images, 115 text spans, and 137M bounding boxes. The output is converted into a sequence of text where special tags represent the start and end of different elements such as text, location tokens for bounding boxes, image embeddings, etc. The model is trained on a combination of multi-modal corps from KOSMOS-1 [108] and GRIT for the next token prediction. After training, instruction tuning is carried out on language-only and grounded-instruction data. KOSMOS-2 performs well on language and vision, grounding, and referring tasks, expanding it to a more diverse set of downstream tasks.

Training with a General Generative Objective: Drawing inspiration from the success of LLMs for language tasks, numerous vision-language models have been developed. Here, we describe methods that train models on simple modeling tasks for pre-training vision-language models. Wang et al. [111] propose a minimalist pre-training framework for vision-language models. The proposed Simple Vision Language Modeling (**SimVLM**) framework trains encoder-decoder style models on the Prefix Language Modeling (PrefixLM) objective. The prefixLM considers the image to be a prefix for the textual description and, thereby, enforces the model to complete a description ($x_{\geq T_p}$) given an image and its partial description ($x_{<T_p}$) of randomly selected length (T_p). A simple transformer-based encoder-decoder architecture is utilized. Textual and visual embeddings (extracted by the first three blocks of a ResNet) are fed to the encoder, and the decoder outputs a text string. This model is trained on prefixLM on noisy image-text pairs dataset [43]. SimVLM does not require task-specific architecture or training and beats previous pre-training and state-of-the-art methods on several vision-language tasks.

Kwon et al. [140] develop a joint masked reconstruction language model where the masked input of one is reconstructed conditioned on other unmasked input. The proposed **MaskVLM** consists of an image and language encoder to encode corresponding modalities and a cross-modal decoder with cross-attention to align both modalities. Image and text are masked randomly following Devlin et al. [86] and He et al. [65], and the model is trained on joint conditional reconstruction task as well as Image-text Contrastive (ITC) [42], [43] and Image Text Matching (TIM) [49] task. This results in an efficient model outperforming similar models for vision-language tasks in a low data regime.

With the demonstrated success of CLIP [42], contrastive learning has gained widespread adoption in vision-language pre-training. Tschannen et al. [115] revisit the effectiveness of captioning for vision-language pre-training on webscale image-text pair datasets and comparisons with contrastive approaches. First, they compare the performance of Captioning-based models (Cap) with CLIP-style models on similar scales and compute budgets. They train a simple encoder-decoder architecture on a standard next-word prediction task. Experimental results show that captioning models a) generally lag behind CLIP-style models in zero-shot classification, but the gap reduces with scale, b) match or outperform them in few-shot classification, c) have competitive performance for classification tasks when fine-tuned with extensive labeled data, and d) with ViT backbone, outperforms CLIP-style models for multi-modal tasks. Second, they propose a generative pre-training method, **CapPa**, which alternates training between standard auto-regressive prediction (CaP) and parallel prediction (Pa), where the entire caption is predicted in a single pass. The CapPa pre-training improves ViT's performance. Third, they reveal the scaling properties of captioners by studying various architectures and training procedures and demonstrate performance gains when training data and architecture are scaled.

Discussion: Unlike the success of next-token prediction in NLP, generative modeling has not yet achieved comparable success in training scalable models for CV [115]. While some studies, such as [115], have explored this approach, it requires

further careful consideration. Generative-based objectives may learn finer-grained concepts than their contrastive counterparts and potentially overcome some of the limitations of contrastive modeling [141]. A significant obstacle in advancing this area is the lack of appropriate benchmarks that can directly compare the strengths and weaknesses of models trained with different paradigms, complicating performance comparisons. Thus, the development of new benchmarks for better comparison of the capabilities of foundational models across training mechanisms and architectures could substantially benefit vision-language models.

C. Hybrid Contrastive and Generative Learning

Unification of tasks: Inspired by the generalizability of BERT [69] in Natural Langauge Processing (NLP) tasks [86], Chen et al. [49] propose Universal Image-TExt Representation (**UNITER**), a method to leverage conventional image-text datasets (COCO [142], Visual Genome [143], Conceptual Captions [54], SBU Captions [144]) to train foundation models that can be used for heterogeneous vision-language tasks. Four pre-training tasks are involved, spanning generative (i.e., Masked Language Modeling (MLM), Masked Region Modeling (MRM)) and contrastive (Image-Text Matching (ITM), and Word Region Alignment (WRA)) objectives. The UNITER architecture consists of an image, text embedder, and a cross-modality contextualized embedding transformer. The text and images of these datasets are fed to respective embedders to extract embeddings. Individual embeddings are fed to the cross-modality transformer for a cross-modal representation. This model is trained on four different vision-language datasets to optimize the pre-training tasks mentioned earlier. UNITER demonstrates generalizability across nine different vision-language tasks, setting state-of-the-art for most tasks.

Chen et al. [116] reformulate and unify four core vision tasks (object detection, instance segmentation, key points prediction, and captioning) into a single pixel-to-sequence interface where both task description and outputs are converted into tokens. The proposed **Pixel2Seqv2** model utilizes an encoder-decoder architecture in that a vision encoder encodes image inputs, and a sequence decoder generates a single token conditioned on previous tokens and encoded images. This model is trained on a simple language modeling task conditioned on previous tokens and encoded images. At inference time, output tokens are sampled and given a task prompt and input image, and task-specific de-tokenization is performed. This Pixel2Seq can solve four vision tasks efficiently without requiring specialized architecture or losses.

In [117], Cho et al. present a unified framework to learn different computer vision tasks in a single architecture. The unification is achieved by reformulating these tasks into multi-modal conditional text generation tasks. The proposed Vision-Language (**VL-x**) employs a pre-trained encoder-decoder language model, such as BART or T5 [138], [145]. Text and visual embeddings are fed to the encoder of this language model. Visual embeddings are extracted from a pre-trained object detector model and consist of Region of Interest (RoI) object features, RoI bounding box coordinates, and image and region IDs. The output of the visual task is converted into a sequence of characters, and a task-specific prefix is added (e.g., classification: bird). This augmented text is encoded as learned embeddings and fed to the language model encoder along with the visual embeddings. This model is then trained with multi-modal language modeling and related vision-language pre-training tasks, such as visual question-answering, image-text matching, visual grounding, and grounded captioning. The multi-tasking framework achieves performance comparable to that of specialized models.

Universal Architectures: We discuss approaches aiming for strong performance across uni, cross, and multi-modal tasks via innovative architectures and diverse objectives, encompassing contrastive, generative, and task-specific losses. Yu et al. [119] propose a unified encoder-decoder-based model called Contrastive Captioner (**CoCa**) that has the capabilities of the single encoder, dual encoder, and encoder-decoder models. The CoCa model consists of an unimodal image and text encoder and a decoupled multi-modal decoder with a cross-attention layer. The unimodal encoders are trained on a contrastive loss like CLIP. This helps the model learn robust and aligned global representations. The decoupled decoder is trained with a generative approach to captioning loss, which allows it to learn detailed granularity and region-level information. A combination of these two approaches endows the model with both contrastive and generative capabilities. This model performs well across a set of diverse vision datasets. The CoCa model can perform multiple tasks and outperforms numerous specialized models under zero-shot, few-shot, and light fine-tuning settings.

Aiming to develop a foundation model performing well across vision, language, and vision-language tasks, Singh et al. propose **FLAVA** [50] consisting of image and text encoders and multi-modal encoders, vision-task heads, language task heads, and multi-modal task heads. The image and text encoders convert the input to a representation fed to a multi-modal encoder. The multi-modal encoder transformers apply cross-attention and fuse both modalities. This multi-modal representation is fed to modality-specific heads (vision, language, and vision language). To achieve better generalization, this model is trained on multiple uni and multi-modal losses, including a global contrastive loss similar to CLIP [42] for cross-modal alignment, masked multi-modal masking and image-text matching, masked-image modeling, masked-language modeling, etc. The training consists of an unimodal pre-training of image and text encoders on supervised datasets, followed by joint unimodal and multimodal training on image-text datasets. FLAVA is evaluated on 35 tasks across vision, language, and vision-language tasks and achieves state-of-the-art performance.

The **BridgeTower** [123] model combines information from the top layers of uni-modal encoders by introducing bridge layers. These bridge layers enable cross-modal alignment and fusion of different levels of semantic, visual, and textual features without affecting the individual encoder's ability to perform uni-modal tasks. Their architecture consists of a standard vision and language encoder and a cross-modal encoder with multiple bridge layers connecting both encoders' top layers by co-attention [69]. Despite using a smaller training dataset,

FLAVA performs significantly better on 35 vision, language, and vision-language tasks.

Chen et al. [120] propose **PaLI** and analyze the scaling effects of large image-text models through a jointly scaled architecture and a vast multilingual image-text dataset. The PaLI architecture combines a text encoder-decoder transformer (mT5) [121] with a ViT [90] for visual tokens. While both components are pre-trained, only the language element is updated and trained across various vision and language tasks. Additionally, the language model undergoes training on pure language understanding tasks to mitigate the risk of catastrophic forgetting. An expansive dataset, WebLI, containing 10 billion images and 12 billion alt-text descriptions, is constructed. A filtered 10% subset of WebLI, consisting of 1 billion samples, is utilized to balance quality and scale. The training set consists of filtered WebLI coupled with task-specific vision and language datasets, including span corruption and object detection. Vision-specific datasets are converted into appropriate format by using template-based prompts. In addition, comprehensive evaluations on scaling laws within vision-language models are carried out. PaLI's extensive pre-training across 100+ languages culminates in state-of-the-art performance across diverse vision, language, and vision-language tasks.

To address the disparity in performance between existing models that span modalities and individual-type foundation models, Zhang et al. [124] introduce **X-FM**. The X-FM architecture consists of three modular encoders: language, vision, and fusion. The language and vision encoders are a stack of BERT [86] and ViT [84] like transformer layers along with post-layer and pre-norm, respectively. In the fusion encoder's self-attention sub-layers, queries are from language, and keys and values are from vision. The learning methodology for X-FM involves training encoders through a blend of unimodal and multimodal objectives supplemented by two novel techniques. Language training employs a masked language model (MLM) and image-text contrastive learning (ITC) for the encoder. The vision encoder undergoes masked image modeling (MIM) and ITC training. Meanwhile, the fusion encoder is trained through image-text matching (ITM), image-conditioned masked language modeling (IMLM), and bounding box prediction (BBP). When training the language encoder, the first new technique involves halting the gradient from the vision-language path. This isolates the language encoder, enabling MLM and ITC-based training for language modeling and alignment. The second technique entails training the vision encoder with masked images, minimizing the discrepancy between masked and unmasked outputs using mean squared error (MSE) loss. This efficient MIM training enhances both vision and fusion encoders mutually. X-FM's performance surpasses other general foundation models across twenty-two tasks encompassing language, vision, and language-vision domains.

Most foundation models rely on the large-scale nature of noisy image-text datasets, which Li et al. [122] argue is a suboptimal strategy. To make better use of data, they propose **BLIP** framework. First, BLIP has the CapFilt (captioning and filtering) mechanism to bootstrap quality data from noisy web data. CapFilt model is initialized from a model pre-trained on noisy image-text data and fine-tuned on small curated data (e.g., COCO [146]). This fine-tuned model is utilized to synthesize quality data. Specifically, given web images, the Captioner synthesizes captions, and the filter eliminates noisy ones. Second, BLIP has a Multimodal mixture of Encoder-Decocer (MED) architecture consisting of unimodal encoders for image and text and image-grounded text encoder and decoder. This model is trained with image-text contrastive, image-text matching, and language modeling objectives. The BLIP framework achieves significant improvements and state-of-the-art performance across various tasks, such as image-text retrieval, captioning, and VQA.

Efficient Utilization of Pre-Trained Models: BLIP and similar models are computationally expensive, requiring large-scale, end-to-end image-text training, often from scratch. Here, we discuss methods that efficiently leverage pre-trained vision and language models for vision-language modeling. Li et al. [125] propose **BLIP-2** to efficiently align pre-trained and frozen unimodal text and image encoders on an image-caption dataset. BLIP-2 uses a querying transformer to bridge the modality gap of frozen uni-modal encoders. The parameters of the Q-former are trained to align two modalities by using image-text contrastive learning, image-grounded text generation, and image-text matching losses. Dai et al. [9] show that aligning pre-training models on image-caption allows broader generalization. The proposed **InstructBLIP** is a vision-language instruction-tuning framework that enables general foundation models to solve multi-modal tasks through a unified language interface. Like BLIP-2 [125], the architecture consists of a visual encoder, a Q-Former, and an LLM. Different from BLIP-2, an instruction-aware visual feature extraction module is developed. Specifically, the Q-former takes encoded images as well as instruction embeddings. This enables Q-Former to extract instruction-related visual features. Similar to BLIP-2, their model is trained in two phases: Q-former is first trained on image-text pairs, and then instruction-tuning is performed where both LLM and visual encoder are kept frozen. A suite of 26 datasets is converted into instruction-tuning format following set templates to train this model for multi-modal tasks. InstructBLIP achieves state-of-the-art zero-shot performance across a wide range of vision-language tasks.

Aiming to address the computational complexity of training visual components in multi-modal models, Zhang et al. [130] introduce **VPGTrans**. This approach streamlines the transfer of visual encoders across varying sizes and types of LLMs based on an extensive experimental investigation. The two-stage strategy involves initially freezing the trained VPG from the source LLM and fine-tuning the projection module on the target LLM. Subsequently, the VPG and projection layer are jointly trained with the target LLM. Empirical findings across diverse LLM configurations showcase impressive performance transfer with notably reduced training data and computational demands. Furthermore, in model enhancement, Zhang et al. [129] present an efficient framework for seamlessly upgrading legacy foundation models to new tasks. The proposed Task Agnostic Compatible Adapter **TaCA** operates as a compact module aligning representations from old and new encoders. By incorporating

distillation loss and cross-modal contrastive loss between feature sets of these encoders, this framework enables module upgrades without necessitating extensive re-training.

Discussion: In the last two sections, we explore research focused on training general models using web-based image-text datasets. Hybrid models, on the other hand, employ two main strategies: leveraging high-quality data from a variety of vision and vision-language tasks (such as segmentation and captioning) to train 'universal models,' or combining task data with image-text pairs to train generic models. These strategies facilitate the use of existing datasets and enable efficient training of large models with multiple task-specific objectives. However, these approaches may not yield an off-the-shelf image encoder that can be plugged into a wide range of applications like CLIP.

D. Conversational Vision-Language Models

Conversational VLMs are models equipped to hold human-like conversations based on multi-modal inputs. These models result from combining the success of Large Language Models (LLMs) and multi-modal models described in previous sections, such as CLIP. In this section, we review efforts to create such conversational VLMs.

OpenAI develops the **GPT4** model [161], which can hold multi-modal conversations, describe intricate images, and solve complex real-world problems. This model is based on transformer-based architecture [89], pre-trained to predict the next word token using public and private datasets. GPT4 is then fine-tuned with Reinforcement Learning from Human Feedback (RLHF) [162]. GPT4 performs well on conventional and real-world NLP, vision, and vision-language tasks. While GPT4 [161] shows state-of-the-art performance on multiple tasks, the model behind it is closed-sourced, and architectural details are unknown. Zhu et al. [8] aims to unravel this and develop an open-source **miniGPT4**, which consists of a pre-trained large language model (LLM), Vicuna [128], and a visual component that consists of ViT-G [70] and a Q-Former. MiniGPT-4 adds a single linear projection layer on the vision encoder and freezes all other parameters. A two-stage training-finetuning scheme is proposed to align visual features with the LLM. First, MiniGPT-4 is trained on a large set of multi-modal examples consisting of Conceptual Captions [54], SBU [144], and LAION [46]. Second, to improve the naturalness and usability, MiniGPT-4 is fine-tuned on a high-quality curated dataset of instructions and respective image and text pairs. MiniGPT-4 exhibits intriguing properties similar to GPT4, such as generating intricate image descriptions, creating a website from a sketch, and explaining visual scenarios.

Instruction-tuning has played an essential role in aligning LLMs to follow instructions and solve various tasks. Liu et al. [7] develop an open-source visual instruction-tuning framework and model dubbed **LLaVA**. They have two main contributions. First, a cost-effective method is developed to curate multi-modal instruction-following data by leveraging ChatGPT and GPT4 [161]. The curated dataset consists of conversations, detailed descriptions of visual inputs, and complex reasoning. Second, a large multi-modal framework based on a large pre-trained language model (LLaMA [114]) and CLIP vision encoder (ViT) is presented. The vision encoder processes images and feeds to a linear projection layer to make features compatible with the LLM. This model is trained using a two-phase strategy. First, it is trained for vision-language alignment, and then only the projection layer's parameters are updated. In the second phase, the LLMs and projection layer parameters are fine-tuned end-to-end on a curated dataset. Similarly, Zhang et al. [163] also present a visual-tuning dataset.

Zhang et al. [164] develop the efficient **LLaMA-Adapter** to convert LLaMA [114] for instruction-following. LLaMA-Adapter is primarily designed for text-based instruction fine-tuning, but it also incorporates visual knowledge. It appends a set of learnable adaption prompts as prefixes in the input tokens of the early transformer layers. LLaMA-Adapter can also handle input images by converting them into visual tokens using CLIP [42] based encoders. These adaptable visual prompts are also incorporated into LLaMA for vision-language tasks. In addition, it performs well in the large-scale multi-modal ScienceQA [165] dataset. However, due to a lack of instruction-following dataset, LLaMA-Adapter [164] can only operate with traditional vision-language tasks. Gao et al. [166] design a parameter-efficient visual instruction, LLaMA-Adapter V2, that achieves better performance on language-vision tasks and can conduct multi-run dialogs. Several improvements are introduced in this model. First, they unfreeze the normalization layers and add bias in the linear layer of the transformer to incorporate new knowledge in the LLaMA. Second, disjoint visual-language and language-only training processes are developed to leverage both data types. Third, visual and text tokens are fed to different transformer layers to avoid interference with visual and language representations. Finally, expert systems (e.g., captioning models and search engines) are utilized to improve the data further. LLaMA-Adapter V2 performs better on conventional vision-language tasks and visual instruction-following compared with V1.

Similarly, Ye et al. [10] present **mPLUG-OWL**, a modular vision-languages model trained on language modeling objectives. The model consists of an image encoder, an image abstractor, and a frozen LLM. This model is trained in two phases: the image encoder and visual abstractor are first trained on image-text pairs with language modeling tasks, and then a visual abstractor and a Low-Rank Adaptation module are fine-tuned with language-only and multi-modal datasets. The proposed mPLUG-OWL performs well on multi-turn dialogue and instruction understanding, visual understanding, and knowledge transfer.

Existing LVLMs struggle to process high-resolution images, often missing small details. This issue arises because they use pre-trained encoders like CLIP, which are pre-trained on lower-resolution images. To address this, Monkey [167] introduces a two-stage approach. First, it divides the high-resolution image into smaller patches and then processes each patch through a LoRA-adapted ViT encoder. Second, it utilizes a multi-level description of the image to improve the context for the LLM. This two-stage strategy significantly boosts Monkey's efficacy on high-resolution images.

TABLE III
A SUMMARY OF PUBLICLY AVAILABLE INFORMATION ABOUT VISUALLY PROMPTED, HETEROGENEOUS MODALITY-BASED MODELS AND EMBODIED FOUNDATIONAL AGENTS, THEIR PROMPT DESIGN DIFFERENCES, AND THE NATURE OF THEIR TRAINING DATA TYPE AND SIZE

Type	Foundational Model	Public	[Link]	Prompts	Training	Data Size
Visually Prompted Models	CLIPSeg [147]	✓	Link	Text, Image	✓	0.34M Images
	SegGPT [33]	✓	Link	Learnable Image Prompt	✓	≈ 0.3M Images
	SAM [11]	✓	Link	Points, Box, Text	✓	11M Images, 1.1B Masks
	SEEM [148]	✓	Link	Text, Points, Scribbles, Boxes, Images	✓	118K Images
	FasterSAM [149]	✓	Link	Points, Box, Text	✓	11K Images
	Fast Segment [150]	✓	Link	Points, Box, Text	✓	22K Images
Generalist Models	Painter [151]	✓	Link	Task-specific Prompts	✓	≈ 162K Images
	VisionLLM [152]	✓	Link	Task Descriptions and Categories	✓	238K Images
	Prismer [153]	✓	Link	Text	✓	11M Images, 12.7M (Image, Text)
Heterogeneous Modalities based Models	CLIP2Video [154]	✓	Link	Text	✓	≈ 33.7K Videos
	AudioCLIP [18]	✓	Link	Text, Audio	✓	1.8M (Video, Audio)
	Image Bind [17]	✓	Link	Text, Depth, Audio, Thermal	✓	1.8M (Video, Audio), 10.3K (Image, Depth), 15.4K (Image, Thermal), 3,670 hours (Video, IMU)
	MACAW-LLM [19]	✓	Link	Text generated using GPT4	✓	69K (Image, Text), 50k (Video, Text)
	COSA [155]	✓	Link	Text	✓	From 5M to 514M (Image, Text)
	Valley [34]	✓	Link	Text generated using ChatGPT	✓	42K Conversations and 5.8K QA about Videos
Embodied Foundational Agents	Palm-E [156]	✗	–	Multi-modal (Image, Text)	✓	Pushing Dynamics, Tasks in a Kitchen Environment,
	ViMA [157]	✓	Link	Multi-modal (Image, Text)	✓	600K+ expert trajectories for learning
	MineDojo [158]	✓	Link	Text	✓	730k Videos, 6k Transcripts, 340K Reddit Posts
	VOYAGER [159]	✓	Link	Iterative Prompting with Feedback	✓	Minecraft items, Library storing behavior
	LM-Nav [160]	✓	Link	LandMarks, Text	✗	–

Discussion: To supplement LLMs with visual understanding, open-source initiatives predominantly utilize image encoders from CLIP-style vision-language models. The image encoder and LLM integration are accomplished through an efficient alignment stage on a small image-text dataset. This integration benefits from advances in both LLMs and vision-language models. However, these models do not perform as well as proprietary models such as GPT-4 or Bard, which are trained end-to-end. Combining vision and language encoders natively coupled with end-to-end training on web-scale data could lead to more effective LVLMs.

IV. VISUALLY PROMPTED MODELS

This section discusses foundation models that can be prompted by non-textual prompts and are designed for various visual tasks (Table III). These models can leverage diverse prompt types such as text, points, bounding boxes, or even a mask of the desired region to obtain the target segmentation. Furthermore, we briefly touch upon adapting SAM for applications without large-scale datasets and relegate details to the Appendix (Section D).

Foundation Models for Segmentation: Segmentation involves grouping pixels in meaningful concepts within an image and pixel-wise object identification. Different types of segmentation are based on how pixels are grouped, including panoptic, instance, and semantic segmentation. The existing segmentation models are specialized based on the type of segmentation or dataset. The foundation segmentation models aim to develop models that can be generalized universally for segmentation tasks.

Standard segmentation models cannot generalize to new categories and are unable to incorporate further queries without retraining on a relevant dataset. **CLIPSeg [147]** exploits CLIP's [42] generalization capabilities to achieve zero-shot and one-shot segmentation. CLIPSeg accomplishes this feat by conditioning transformer-based decoder on joint text-visual CLIP embeddings. This model consists of CLIP-based image and text encoders and a transformer-based decoder with skip connections like U-net [168]. The relevant CLIP encoder processes the visual and text queries to construct embeddings, which are then input to the CLIPSeg decoder. Target segmentation can be prompted by text or an image using CLIP. As such, CLIPSeg can generate image segmentations based on arbitrary prompts at test time.

Diversifying Segmentation Tasks: **SegGPT [33]** provides an in-context learning paradigm and aims to train a single foundation model with a generalizable scheme for these diverse segmentation tasks. The challenge is accommodating different segmentation tasks and datasets in a single training framework. SegGPT accomplishes this by mapping different kinds of segmentation data into the same format of images (random color mapping for each data sample) using an in-context learning framework [151]. The goal is to color the appropriate areas, such as classes, object instances, and parts, according to the context. After training, the SegGPT model can perform few-shot semantic segmentation, video object segmentation, semantic

segmentation, and panoptic segmentation without fine-tuning for the downstream tasks.

SAM [11]: Given an image and visual prompt (box, points, text, or mask) that specifies what to segment in an image, SAM encodes the image and prompt embeddings, which are then combined in a lightweight mask decoder that predicts segmentation masks. SAM outputs a valid segmentation mask even when the given prompt is ambiguous. For example, given a point prompt on a person wearing a shirt, the model has to segment the shirt or the person wearing it. It is trained on over 1 billion masks with privacy-respecting images and model-in-the-loop dataset annotation settings. The data annotation has three stages: assisted-manual, semi-automatic, and fully automated. In the first stage, SAM assists annotators in annotating masks. By prompting SAM with likely object locations, SAM can generate masks for a subset of objects while annotators focus on annotating the remaining objects. The final step involves prompting SAM with a regular grid of foreground points, yielding an average of 100 high-quality masks per image.

Diversifying SAM's Prompting Mechanism: Zou et al. [148] propose the SEEM model as a universal interface for the segmentation of everything, everywhere, with multi-modal prompts. It can take multiple types of prompts, including points, masks, text, boxes, and refereed regions of another image, and thus has strong composability. SEEM comprises a text, image encoder, and visual sampler for such prompts. These encoded inputs are projected to a joint image-text representations space and fed to a decoder that outputs classes and mask embeddings.

Adapting SAM for other Applications: A visual foundation model like SAM [11] is trained on large-scale datasets that contain more than one billion masks and 11 million images. In other domains, such as medical image understanding, large-scale datasets of this scale may not be available. To alleviate these issues, SAM has been adapted for a variety of vision tasks, such as medical, tracking, remote sensing, and captioning. For a detailed discussion of the adaptations of SAM for these applications, please refer to Appendix D.

Discussion: The visual prompts and model-in-the-loop data approach taken by SAM have not only become pioneering for training foundational models but have also set new standards for data collection and training pipelines. Furthermore, it has attracted much attention to adopting SAM for various other relevant fields. However, SAM is computationally expensive, and its pipeline depends on large datasets. While some studies have begun addressing these challenges, there is a need for more efforts to develop a more efficient SAM-like approach suitable for a wider array of tasks across different fields.

V. Generalist, Heterogeneous and Embodied Models

A. Generalist Models

Using contextual learning, a model can be quickly adapted to various tasks with only a few prompts and examples in NLP [169]. The difficulty with in-context learning in vision is that output representations vary greatly among tasks (requiring different loss functions and architectures), making it challenging to define general-purpose task prompts or instructions for reconfiguring a visual model for out-of-domain tasks. **Painter [151]** is a general model that can perform multiple tasks simultaneously and adapt to a new one given a prompt and very few specific examples. Given an input and output image for a specific task, the output image's pixels are masked, and the objective of the Painter model is to paint the masked output image. This simple training objective unifies several vision tasks (without modifications to model architecture or loss function), including depth estimation, human key point detection, semantic segmentation, instance segmentation, image denoising, image deraining, and image enhancement. After training, the Painter can determine which task to perform during inference using the input/output paired images from the same task as the input condition. **VisionLLM [152]** is a general model that aligns vision and language modalities to solve open-ended tasks. Given an image, VisionLLM learns image features using a vision model; these features and language instruction, e.g., "describe the image in detail," are passed through a language-guided image tokenizer. The image tokenizer's output and language instructions are provided to an open-ended LLM-based task decoder designed to orchestrate various tasks according to language instructions. **Prismer** [153] is another vision-language model that leverages diverse pre-trained domain experts in semantic segmentation, object, text, and edge detection, surface normal and depth estimation to perform multiple reasoning tasks such as image captioning and visual question answering.

Discussion: Generalist models harness the power of in-context learning to tackle novel tasks, employing various methods to address the output-diversity challenge in vision. Further efforts in utilizing traditional models pre-trained on vision tasks (e.g., segmentation, detection) and foundational models (e.g., CLIP and GPT4) can result in generalist models that are efficient and able to solve diverse sets of complex tasks.

B. Heterogeneous Modalities Based Models

We discuss foundation models that align multiple paired modalities, e.g., image-text, video-audio, image-depth, etc., for representation learning. These heterogeneous modalities-based approaches can be divided into two categories: aligning CLIP with heterogeneous modalities and aligning LLMs with these modalities. The first category comprises approaches **CLIP2Video** [154] that align CLIP with videos, **Audio-CLIP** [18] which extends CLIP for audio, and **Image Bind** [17] which leverages multiple modalities. The second category includes approaches like **MACAW-LLM** [19], an instruction-tuned multi-modal LLM that integrates four different modalities, **COSA** [155], which aligns LLMs with videos, and **Valley** [34], a multi-modal framework capable of integrating video, image, and language perceptions. More in-depth explanations of these models and further discussion are provided in Section E of the Appendix.

C. Embodied Agents

The embodied agents encompass techniques for real-world robot manipulation, game-play-based continual learning, and

navigational agents. Notably, **Palm-E** [156] integrates continuous sensor input data into a language model, facilitating grounded inference. Similarly, **ViMA** [157] demonstrates the efficacy of multimodal prompting, interweaving textual and visual tokens, thereby enabling effective learning of robot manipulation through multimodal prompts. Furthermore, **MineDojo** [158] gathers and trains agents on Minecraft data, while **VOYAGER** [159] is developed for lifelong learning based on LLM. It is designed to drive exploration, refine diverse skills, and continuously discover new elements within the Minecraft environment. Finally, **LM-Nav** [160] utilizes VLMs to devise actionable plans in an environment.

VI. Challenges and Research Directions

Multimodal Open-source Models: In NLP tasks, the advances from GPT3 to ChatGPT show the importance of instruction-following and human-feedback-based reinforcement learning. For multimodal (text and image) inputs, a similar capability is claimed by GPT4 [161] to allow reasoning and understanding based on vision-language inputs. However, GPT4 is a closed-source model with restricted access to date, and its training details remain unknown. To bridge this gap, multimodal open-source foundation models such as BLIP2 [125], GIT [170], and Flamingo [104] can be extended with the instruction-following and human intent alignment to obtain ChatGPT-like in the multimodal space. While some efforts have been made, e.g., Intruct-BLIP, miniGPT4, LLaVA, and Video-ChatGPT, matching GPT4's capabilities with open-source frameworks remains a significant challenge for multi-modal foundation models.

Evaluation and Benchmarking: The open-ended nature of large-scale conversational Vision-Language Models makes their comprehensive evaluation challenging. This challenge is shared with the progress in LLM but more severe for visual inputs since the possible tasks and reasoning capabilities become diverse for a broad evaluation. One *quantitative* approach is to define a set of instructions covering multiple reasoning aspects and forward the responses from two competing chatbot VLMs to GPT4 to rate them on a scale of 1 to 10. Vicuna-Instruction-80 introduces this LLM-as-a-judge approach [128], [171] benchmark for LLM that comprises of 9 instruction categories: *generic, knowledge, math, counterfactual, Fermi, coding, writing, roleplay, common-sense*. This approach has also been extended for VLMs, e.g., [172] uses four criteria (*correctness of information, detail orientation, contextual understanding, temporal understanding, consistency*) scored by GPT4 for benchmarking a VLM tailored for videos. However, the use of an external GPT4 model as a gold standard is still debatable, and new efforts in LLM for benchmarking and identifying the corner cases have been reported to address the limitations of existing evaluation measures, e.g., [36], [173], [174]. Such efforts are likely to be extended to VLMs, with even more attention paid to their peculiar visual aspect.

Hallucination: Hallucination refers to the phenomenon where the output generated from a large VLM/LLM is unreal or nonsensical, often based on a hypothetical scenario. Foundation models for language and vision, e.g., those based on Generative Pretrained Models for open-ended conversations [7], [8], [161], [172], can sometimes fabricate answers, even if they are technically correct in specific contexts. This is because they are trained on massive datasets of text and images that are often noisy, and they may be unable to distinguish between what is real and what is not. Specifically for VLMs where visual data is provided as an input condition, e.g., image-based visual question answering, one form of hallucination ignores the visual input and can only give an answer based on the text prompt. Strategies to mitigate hallucination include providing explicit instructions (or so-called *system commands*), the chain of thought prompting [175], [176], [177], self-consistency (voting) [178], [179] and use of knowledge bases for retrieval augmented generation [180], [181], [182].

Object Hallucination: Another type of hallucination specific to vision-language-based FMs is object hallucination, whereby models generate inaccurate descriptions of non-existing objects. Efforts to mitigate this issue can be broadly categorized into three categories: 1) detection [183], [184], [185], [186] by quantifying hallucinations using various methods such as VQAs, classification of objects etc; 2) construction of fine-tuning data aimed at reducing object hallucinations [184], [187], and algorithmic approaches to minimize hallucinations post-training [188], [189], [190]. Curious readers may refer to [191] for a more comprehensive treatment of the subject.

Multimodal Alignment: Existing VLMs sometimes have poor alignment between vision-language (or other modalities). For instance, Segment anything [11] performance with text prompts is weaker than the visual prompts (points, boxes, or masks). For heterogeneous modalities, such an alignment can be even more challenging. Approaches like ImageBind [17] demonstrate viable ways to achieve alignment between several modalities. However, there is still much room to demonstrate strong alignment capabilities for a more comprehensive range of related inputs with the shared semantic space. For a unified representation space that can thoroughly understand the world around us, foundation models targeted at learning joint embedding spaces will be crucial for further development.

Large Data and Compute Requirements. Training large-scale vision and language models are data and compute-intensive. Acquiring labeled data on a large scale can be expensive and time-consuming, especially for specialized visual domains or low-resource languages. Similarly, their inference is expensive as many parameters are involved. The computational demands limit their accessibility and scalability in many real-world applications. For example, applications requiring real-time inference capability or deployment on the edge and mobile devices with limited on-device compute and restricted battery times. Similarly, visual prompt-based models like Segment Anything [11] would benefit from a real-time speed to ensure intractability [150], [192]. Efforts, such as retentive networks [193], can be integrated within VLMs for high throughput processing.

Adaptation of FMs: Several approaches have been developed for the efficient adaptation of foundational models for downstream tasks and applications. Although Parameter-efficient fine-tuning (PEFT) approaches are primarily explored for LLMs and Diffusion models, they are also directly applicable to adapting other vision foundation models. Some representative

approaches include Low-rank Adaptation (LoRA) [194] and its variants such as QLoRA [195] and GLoRA [196], Prefix tuning [197], [198], Adapters [164], [166], Prompt tuning [199], [200]. Reducing the compute and memory footprint for quick adaptation of textually and visually prompted foundation models is still an open research problem since the existing approaches require careful selection of hyper-parameters (e.g., rank in LoRA or the placement and dimensions of bottleneck adapters) and can result in loss of generalization performance.

Vulnerability to Adversarial Attacks: Foundation models, similar to other neural network-based models, are vulnerable to adversarial attacks. These attacks involve carefully crafted inputs that can cause the model to generate incorrect or harmful outputs [201], [202], [203]. However, there are specific ways for adversarial attacks on foundation models that make them susceptible to unwanted behavior. For example, models based on conversational LLMs have been shown vulnerable to adversarially prompt injection, which needs a direct interaction between the adversary and the LLM-based conversational agent [204], [205], [206]. Greshake et al. [207] demonstrate that even a direct interaction between the model and adversary is not needed in LLM-integrated applications, and an adversary can remotely poison the information retrieved by the conversational agent via indirect prompt injection. This leads to a spectrum of vulnerabilities for LLMs and VLMs, including manipulated content, fraud, malware, intrusion, personal information leakage, and denial of services via language and visual prompt injections. Carlini et al. [208] recently show that NLP-based attacks are weak due to discrete optimization involved, but these results should not be interpreted as showing that these foundation models are robust. For conversational VLMs, an attack can be easily launched by adversarially perturbing the inputs to get harmful responses from the model. In [209], Maus et al. shows how non-meaningful text can be appended with the prompts to fool generative text and image models. The exemplars (input-output pairs) provided for In-Context Learning of VLMs can also be changed to fool the model [210], [211]. Visual prompt-based models, e.g., SAM, have also been attacked by corrupting their inputs and associated prompts [212], [213]. Strategies to robustify VLMs against such attacks is an open research question of great importance.

Bias and Fairness: Foundation models in vision and language can inherit and amplify biases in the data used for training them. Biases, stereotypes, and prejudices related to race, underrepresented groups, minority cultures, and gender can make the models output biased predictions or exhibit skewed behavior. For example, the recent work by Shtedriski et al. [214] shows the sensitivity of the CLIP model towards red circles, where simply drawing a red circle around a person's face increases its chances of being misclassified as a murderer, suspect, or a missing person as data from new media consists of a red circle around criminal faces. New benchmarks have recently been developed to assess the capability of existing VLMs towards certain biases [215]. Addressing biases in foundation AI models is crucial to ensure fairness, inclusivity, and ethical deployment.

Interpretablity: Foundation models are often difficult to interpret, making it difficult to understand how and why they generate the output they do. Existing methods have applied chain-of-thought reasoning to explain the results generated by vision and language models [216]. New benchmarks are also developed to evaluate and train explicitly, focusing on providing detailed step-wise rationales for model choices, e.g., ScienceQA [165]. Recently, Visual Programming uses the interpretable neuro-symbolic representation to break down a complex task into more straightforward steps that explain the rationale of a particular output from GPT-3 [217]. While these approaches show promising results, numerous failure cases can be further improved, e.g., by allowing user feedback to improve the interpretation.

Limited Contextual Understanding: Understanding sarcasm, irony, or other figurative images and language inputs (e.g., memes) with nuanced context can be challenging for foundational models. While initial efforts on this topic have been made in language-only models, it remains an open problem in large multi-modal models.

Role of Visual Language FM Models in Generative Modeling: State-of-the-art (SOTA) generative models extensively rely on visual-language (VL) models for text-guided generation; e.g., open-source Latent Diffusion Models (LDMs) often integrate fixed CLIP [218], [219]. Similarly, GANs and diffusion models employ CLIP and its variants for text-driven image manipulation and editing [220], [221]. However, the impact of differently pre-trained VL models on generation results remains unclear and underexplored. We provide a detailed discussion in the Appendix (Section A.2).

VII. Conclusion

Foundation models with multiple modalities, including natural language and vision, are essential for developing AI systems that can effectively perceive and reason about the real world. This survey reviews vision and language foundation models, focusing on their architecture types, training objectives, downstream task adaption, and prompting designs. We provide systematic categorizations of textually-prompted, visually-prompted, and heterogeneous modality models. In addition, we provide broad coverage of their applications in various visual tasks, including zero-shot recognition and localization abilities, visual dialogue about an image or a video, cross-modal, and medical data understanding. We summarize how foundation models in vision can act as generalist models solving multiple tasks simultaneously. Their combination with large language models gives rise to embodied agents that can continually learn and navigate in a complex environment. We hope this effort will spur further research in leveraging the potential of foundation models and addressing their limitations, e.g., limited contextual understanding, biases, and vulnerability to malicious uses.

References

[1] R. Bommasani et al., "On the opportunities and risks of foundation models," 2021, *arXiv:2108.07258*.
[2] W. X. Zhao et al., "A survey of large language models," 2023, *arXiv:2303.18223*.
[3] T. Brown et al., "Language models are few-shot learners," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 7–8.
[4] A. Chowdhery et al., "PaLM: Scaling language modeling with pathways," 2022, *arXiv:2204.02311*.

[5] R. Anil et al., "PaLM 2 technical report," 2023, *arXiv:2305.10403*.
[6] A. Radford et al., "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, 2019.
[7] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," 2023, *arXiv:2304.08485*.
[8] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "MiniGPT-4: Enhancing vision-language understanding with advanced large language models," 2023, *arXiv:2304.10592*.
[9] W. Dai et al., "InstructBLIP: Towards general-purpose vision-language models with instruction tuning," 2023, *arXiv:2305.06500*.
[10] Q. Ye et al., "mPLUG-Owl: Modularization empowers large language models with multimodality," 2023, *arXiv:2304.14178*.
[11] A. Kirillov et al., "Segment anything," 2023, *arXiv:2304.02643*.
[12] J. Ma and B. Wang, "Segment anything in medical images," 2023, *arXiv:2304.12306*.
[13] J. Wu et al., "Medical sam adapter: Adapting segment anything model for medical image segmentation," 2023, *arXiv:2304.12620*.
[14] L. Yonglin, Z. Jing, T. Xiao, and L. Long, "RefSAM: Efficiently adapting segmenting anything model for referring video object segmentation," 2023, *arXiv:2307.00997*.
[15] J. Yang et al., "Pave the way to grasp anything: Transferring foundation models for universal pick-place robots," 2023, *arXiv:2306.05716*.
[16] K. Chen et al., "Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model," 2023.
[17] R. Girdhar et al., "ImageBind: One embedding space to bind them all," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023.
[18] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "AudioCLIP: Extending clip to image, text and audio," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 976–980.
[19] C. Lyu et al., "Macaw-LLM: Multi-modal language modeling with image, video, audio, and text integration," 2023. [Online]. Available: https://github.com/lyuchenyang/Macaw-LLM
[20] L. Fan, L. Li, Z. Ma, S. Lee, H. Yu, and L. Hemphill, "A bibliometric review of large language models research from 2017 to 2023," 2023, *arXiv:2304.02020*.
[21] J. Huang and K. C.-C. Chang, "Towards reasoning in large language models: A survey," in *Proc. Findings Assoc. Comput. Linguistics*, Jul. 2023, pp. 1049–1065.
[22] Q. Dong et al., "A survey for in-context learning," 2022, *arXiv:2301.00234*.
[23] C. Zhou et al., "A comprehensive survey on pretrained foundation models: A history from bert to chatgpt," 2023, *arXiv:2302.09419*.
[24] S. Long, F. Cao, S. C. Han, and H. Yang, "Vision-and-language pretrained models: A survey," in *Proc. Int. Joint Conf. Artif. Intell.*, 2022, pp. 5530–5537.
[25] Y. Du, Z. Liu, J. Li, and W. X. Zhao, "A survey of vision-language pre-trained models," in *Proc. Int. Joint Conf. Artif. Intell.*, 2022, pp. 5436–5443.
[26] P. Xu, X. Zhu, and D. Clifton, "Multimodal learning with transformers: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12113–12132, Oct. 2023.
[27] Y. Zong, O. M. Aodha, and T. Hospedales, "Self-supervised multimodal learning: A survey," 2023, *arXiv:2304.01008*.
[28] Z. Jingyi, H. Jiaxing, J. Sheng, and L. Shijian, "Vision-language models for vision tasks: A survey," 2023, *arXiv:2304.00685*.
[29] Z. Gan et al., "Vision-language pre-training: Basics, recent advances, and future trends," *Foundations Trends Comput. Graph. Vis.*, vol. 14, no. 3–4, pp. 163–352, 2022.
[30] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5625–5644, Aug. 2024.
[31] J. Huang, J. Zhang, K. Jiang, H. Qiu, and S. Lu, "Visual instruction tuning towards general-purpose multimodal model: A survey," 2023, *arXiv:2312.16602*.
[32] C. Zhang et al., "A comprehensive survey on segment anything model for vision and beyond," 2023, *arXiv:2305.08196*.
[33] X. Wang, X. Zhang, Y. Cao, W. Wang, C. Shen, and T. Huang, "SegGPT: Segmenting everything in context," 2023, *arXiv:2304.03284*.
[34] R. Luo et al., "Valley: Video assistant with large language model enhanced ability," 2023, *arXiv: 2306.07207*.
[35] J. Wei et al., "Emergent abilities of large language models," 2022, *arXiv:2206.07682*.
[36] S. Bubeck et al., "Sparks of artificial general intelligence: Early experiments with GPT-4," 2023, *arXiv:2303.12712*.
[37] L. Yuan et al., "Florence: A new foundation model for computer vision," 2021, *arXiv:2111.11432*.
[38] H. Cao, C. Tan, Z. Gao, G. Chen, P.-A. Heng, and S. Z. Li, "A survey on generative diffusion model," 2022, *arXiv:2209.02646*.
[39] C. Zhang, C. Zhang, M. Zhang, and I. S. Kweon, "Text-to-image diffusion model in generative AI: A survey," 2023, *arXiv:2303.07909*.
[40] L. Yang et al., "Diffusion models: A comprehensive survey of methods and applications," 2022, *arXiv:2209.00796*.
[41] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10850–10869, Sep. 2023.
[42] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
[43] C. Jia et al., "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 4904–4916.
[44] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 9694–9705.
[45] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Comput. Surv.*, vol. 51, no. 6, 2019, Art. no. 118.
[46] C. Schuhmann et al., "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs," 2021, *arXiv:2111.02114*.
[47] C. Schuhmann et al., "LAION-5B: An open large-scale dataset for training next generation image-text models," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 25278–25294.
[48] L. H. Li et al., "Grounded language-image pre-training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10955–10965.
[49] Y.-C. Chen et al., "Uniter: Universal image-text representation learning," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 104–120.
[50] A. Singh et al., "Flava: A foundational language and vision alignment model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 15617–15629.
[51] L. Fei-Fei, J. Deng, and K. Li, "ImageNet: Constructing a large-scale image database," *J. Vis.*, vol. 9, no. 8, p. 1037, 2009.
[52] M. Minderer et al., "Simple open-vocabulary object detection with vision transformers," 2022, *arXiv:2205.06230*.
[53] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process. (Volume 1: Long Papers)*, 2021, pp. 3816–3830.
[54] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2556–2565.
[55] A. Zhai and H.-Y. Wu, "Classification is a strong baseline for deep metric learning," 2018, *arXiv: 1811.12649*.
[56] Z.-Y. Dou et al., "An empirical study of training end-to-end vision-and-language transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18145–18155.
[57] X. Dong et al., "CSWin transformer: A general vision transformer backbone with cross-shaped windows," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12114–12124.
[58] N. Mu, A. Kirillov, D. Wagner, and S. Xie, "SLIP: Self-supervision meets language-image pre-training," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 529–544.
[59] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
[60] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 22243–22255.
[61] B. Thomee et al., "YFCC100M: The new data in multimedia research," *Commun. ACM*, vol. 59, no. 2, pp. 64–73, 2016.
[62] Y. Huo et al., "WenLan: Bridging vision and language by large-scale multi-modal pre-training," 2021, *arXiv:2103.06561*.
[63] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9726–9735.
[64] L. Yao et al., "Filip: Fine-grained interactive language-image pre-training," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 3–7.

[65] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 15979–15988.
[66] Y. Li, H. Fan, R. Hu, C. Feichtenhofer, and K. He, "Scaling language-image pre-training via masking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 23390–23400.
[67] X. Dong et al., "Maskclip: Masked self-distillation advances contrastive language-image pretraining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 10995–11005.
[68] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 1195–1204.
[69] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 13–23.
[70] Q. Sun, Y. Fang, L. Wu, X. Wang, and Y. Cao, "EVA-CLIP: Improved training techniques for clip at scale," 2023, *arXiv:2303.15389*.
[71] M. Byeon, B. Park, H. Kim, S. Lee, W. Baek, and S. Kim, "Coyo-700m: Image-text pair dataset," 2022. [Online]. Available: https://github.com/kakaobrain/coyo-dataset
[72] G. Ilharco et al., "Openclip," Zenodo, Jul. 2021.
[73] Y. Fang et al., "EVA: Exploring the limits of masked visual representation learning at scale," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19358–19369.
[74] M. Cherti et al., "Reproducible scaling laws for contrastive language-image learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 2818–2829.
[75] X. Li, Z. Wang, and C. Xie, "An inverse scaling law for clip training," 2023, *arXiv: 2305.07017*.
[76] X. Li, Z. Wang, and C. Xie, "CLIPA-v2: Scaling CLIP training with 81.1% zero-shot imagenet accuracy within a $10,000 budget; an extra $4,000 unlocks 81.8% accuracy," 2023, *arXiv:2306.15658*.
[77] F. Liu, D. Chen, Z. Guan, X. Zhou, J. Zhu, and J. Zhou, "RemoteCLIP: A vision language foundation model for remote sensing," 2023, *arXiv: 2306.11029*.
[78] G. Kim et al., "Cream: Visually-situated natural language understanding with contrastive reading model and frozen large language models," 2023, *arXiv:2305.15080*.
[79] W. Liu and Y. Zuo, "Stone needle: A general multimodal large-scale model framework towards healthcare," 2023, *arXiv: 2306.16034*.
[80] J. Wang, Y. Zhang, and J. Sang, "FairCLIP: Social bias elimination based on attribute prototype learning and representation neutralization," 2022, *arXiv:2210.14562*.
[81] I. Alabdulmohsin, X. Wang, A. Steiner, P. Goyal, A. D'Amour, and X. Zhai, "Clip the bias: How useful is balancing data in multimodal learning?," 2024, *arXiv: 2403.04547*.
[82] W. Tu, W. Deng, and T. Gedeon, "A closer look at the robustness of contrastive language-image pre-training (CLIP)," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2024, pp. 13–36.
[83] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[84] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv: 2010.11929*.
[85] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
[86] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv: 1810.04805*.
[87] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 1137–1149.
[88] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.
[89] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[90] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1204–1213.
[91] W. Wang et al., "Image as a foreign language: Beit pretraining for all vision and vision-language tasks," 2022, *arXiv:2208.10442*.
[92] Y. Fang, Q. Sun, X. Wang, T. Huang, X. Wang, and Y. Cao, "EVA-02: A visual representation for neon genesis," 2023, *arXiv:2303.11331*.
[93] Z. Wang et al., "CRIS: Clip-driven referring image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11676–11685.
[94] C. Zhou, C. C. Loy, and B. Dai, "Extract free dense labels from clip," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 696–712.
[95] S. Liu et al., "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," 2023, *arXiv:2303.05499*.
[96] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.
[97] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," 2021, *arXiv:2104.13921*.
[98] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
[99] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 936–944.
[100] J. Xu et al., "GroupViT: Semantic segmentation emerges from text supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18113–18123.
[101] G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin, "Scaling open-vocabulary image segmentation with image-level labels," in *Proc. Eur. Conf. Comput. Vis.*, pp. 540–557, 2022.
[102] M. Tsimpoukelli, J. L. Menick, S. Cabi, S. Eslami, O. Vinyals, and F. Hill, "Multimodal few-shot learning with frozen language models," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 200–212.
[103] A. Brock, S. De, S. L. Smith, and K. Simonyan, "High-performance large-scale image recognition without normalization," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 1059–1071.
[104] J.-B. Alayrac et al., "Flamingo: A visual language model for few-shot learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 23716–23736.
[105] J. Hoffmann et al., "Training compute-optimal large language models," 2022, *arXiv:2203.15556*.
[106] A. Awadalla et al., "OpenFlamingo: An open-source framework for training large autoregressive vision-language models," 2023.
[107] Y. Hao et al., "Language models are general-purpose interfaces," 2022, *arXiv:2206.06336*.
[108] S. Huang et al., "Language is not all you need: Aligning perception with language models," 2023, *arXiv:2302.14045*.
[109] H. Wang et al., "Foundation transformers," 2022, *arXiv:2210.06423*.
[110] Z. Peng et al., "Kosmos-2: Grounding multimodal large language models to the world," 2023, *arXiv:2306.14824*.
[111] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "SimVLM: Simple visual language model pretraining with weak supervision," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 4–5.
[112] G. Kwon, Z. Cai, A. Ravichandran, E. Bas, R. Bhotika, and S. Soatto, "Masked vision and language modeling for multi-modal representation learning," in *Proc. 11th Int. Conf. Learn. Representations*, 2023, pp. 8948–8957.
[113] Y. Liu et al., "RoBERTa: A robustly optimized bert pretraining approach," *ArXiv*, vol. abs/1907.11692, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:198953378
[114] H. Touvron et al., "Llama: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.
[115] M. Tschannen, M. Kumar, A. Steiner, X. Zhai, N. Houlsby, and L. Beyer, "Image captioners are scalable vision learners too," 2023, *arXiv:2306.07915*.
[116] T. Chen, S. Saxena, L. Li, T.-Y. Lin, D. J. Fleet, and G. E. Hinton, "A unified sequence interface for vision tasks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 31333–31346.
[117] J. Cho, J. Lei, H. Tan, and M. Bansal, "Unifying vision-and-language tasks via text generation," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 6666–6676.
[118] Z. Peng, L. Dong, H. Bao, Q. Ye, and F. Wei, "BEiT v2: Masked image modeling with vector-quantized visual tokenizers," 2022, *arXiv:2208.06366*.
[119] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "CoCa: Contrastive captioners are image-text foundation models," 2022, *arXiv:2205.01917*.

[120] X. Chen et al., "PaLI: A jointly-scaled multilingual language-image model," in *Proc. 11th Int. Conf. Learn. Representations*, 2022, pp. 4–7.

[121] L. Xue et al., "mT5: A massively multilingual pre-trained text-to-text transformer," 2020, *arXiv: 2010.11934*.

[122] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 12888–12900.

[123] X. Xu, C. Wu, S. Rosenman, V. Lal, and N. Duan, "Bridge-tower: Building bridges between encoders in vision-language representation learning," 2022, *arXiv:2206.08657*.

[124] X. Zhang, Y. Zeng, J. Zhang, and H. Li, "Toward building general foundation models for language, vision, and vision-language understanding tasks," 2023, *arXiv:2301.05065*.

[125] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," 2023, *arXiv:2301.12597*.

[126] S. Zhang et al., "OPT: Open pre-trained transformer language models," 2022, *arXiv:2205.01068*.

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

[128] W.-L. Chiang et al., "Vicuna: An open-source chatbot impressing GPT-4 with 90% chatGPT quality," Mar. 2023. [Online]. Available: https://lmsys.org/blog/2023-03-30-vicuna/

[129] B. Zhang, Y. Ge, X. Xu, Y. Shan, and M. Z. Shou, "Taca: Upgrading your visual foundation model with task-agnostic compatible adapter," 2023, *arXiv:2306.12642*.

[130] A. Zhang et al., "Transfer visual prompt generator across LLMs," 2023, *arXiv:2305.01278* .

[131] Z.-Y. Dou et al., "Coarse-to-fine vision-language pre-training with fusion in the backbone," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 32942–32956.

[132] Z. Wang et al., "Detecting everything in the open world: Towards universal object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 11433–11443.

[133] Y. Zhong et al., "Regionclip: Region-based language-image pretraining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022.

[134] X. Zou et al., "Generalized decoding for pixel, image, and language," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 15116–15127.

[135] J. Yang, C. Li, X. Dai, and J. Gao, "Focal modulation networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 4203–4217.

[136] M. Ding, B. Xiao, N. Codella, P. Luo, J. Wang, and L. Yuan, "Davit: Dual attention vision transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 74–92.

[137] H. Zhang et al., "GLIPv2: Unifying localization and vision-language understanding," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 36067–36080.

[138] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 1, 2020, Art. no. 140.

[139] Y. Sun et al., "A length-extrapolatable transformer," 2022, *arXiv:2212.10554*.

[140] G. Kwon, Z. Cai, A. Ravichandran, E. Bas, R. Bhotika, and S. Soatto, "Masked vision and language modeling for multi-modal representation learning," in *Proc. 11th Int. Conf. Learn. Representations*, 2022, pp. 4–8.

[141] S. Tong, E. Jones, and J. Steinhardt, "Mass-producing failures of multimodal systems with language models," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2024, pp. 29292–29322.

[142] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[143] R. Krishna et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, pp. 32–73, 2017.

[144] V. Ordonez, G. Kulkarni, and T. Berg, "Im2text: Describing images using 1 million captioned photographs," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2011, pp. 3–4.

[145] M. Lewis et al., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," 2019, *arXiv: 1910.13461*.

[146] H. Caesar, J. Uijlings, and V. Ferrari, "COCO-stuff: Thing and stuff classes in context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1209–1218.

[147] T. Lüddecke and A. Ecker, "Image segmentation using text and image prompts," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7076–7086.

[148] X. Zou et al., "Segment everything everywhere all at once," 2023, *arXiv:2304.06718*.

[149] C. Zhang et al., "Faster segment anything: Towards lightweight sam for mobile applications," 2023, *arXiv:2306.14289*.

[150] X. Zhao et al., "Fast segment anything," *arXiv:2306.12156*, 2023.

[151] X. Wang, W. Wang, Y. Cao, C. Shen, and T. Huang, "Images speak in images: A generalist painter for in-context visual learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 6830–6839 .

[152] W. Wang et al., "VisionLLM: Large language model is also an open-ended decoder for vision-centric tasks," 2023, *arXiv:2305.11175*.

[153] S. Liu, L. Fan, E. Johns, Z. Yu, C. Xiao, and A. Anandkumar, "Prismer: A vision-language model with an ensemble of experts," 2023, *arXiv:2303.02506*.

[154] H. Fang, P. Xiong, L. Xu, and Y. Chen, "Clip2video: Mastering video-text retrieval via image clip," 2021, *arXiv:2106.11097*.

[155] S. Chen, X. He, H. Li, X. Jin, J. Feng, and J. Liu, "COSA: Concatenated sample pretrained vision-language foundation model," 2023, *arXiv:2306.09085*.

[156] D. Driess et al., "Palm-e: An embodied multimodal language model," 2023, *arXiv:2303.03378*.

[157] Y. Jiang et al., "VIMA: General robot manipulation with multimodal prompts," 2022, *arXiv:2210.03094*.

[158] L. Fan et al., "MineDojo: Building open-ended embodied agents with internet-scale knowledge," 2022, *arXiv:2206.08853*.

[159] G. Wang et al., "Voyager: An open-ended embodied agent with large language models," 2023, *arXiv:2305.16291*.

[160] D. Shah, B. Osinski, B. Ichter, and S. Levine, "LM-Nav: Robotic navigation with large pre-trained models of language, vision, and action," 2022, *arXiv:2207.04429*.

[161] OpenAI, "GPT-4 technical report," *PREPRINT*, 2023.

[162] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, Art. no. 120036.

[163] Y. Zhang et al., "LLaVAR: Enhanced visual instruction tuning for text-rich image understanding," 2023, *arXiv:2306.17107*.

[164] R. Zhang et al., "Llama-adapter: Efficient fine-tuning of language models with zero-init attention," 2023, *arXiv:2303.16199*.

[165] P. Lu et al., "Learn to explain: Multimodal reasoning via thought chains for science question answering," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 2507–2521.

[166] P. Gao et al., "LLaMA-Adapter V2: Parameter-efficient visual instruction model," 2023, *arXiv:2304.15010*.

[167] Z. Li et al., "Monkey: Image resolution and text label are important things for large multi-modal models," 2023, *arXiv:2311.06607*.

[168] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2015, pp. 234–241.

[169] L. Ouyang et al., "Training language models to follow instructions with human feedback," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 27730–27744.

[170] J. Wang et al., "GIT: A generative image-to-text transformer for vision and language," *Trans. Mach. Learn. Res.*, 2022. [Online]. Available: https://openreview.net/forum?id=b4tMhpN0JC

[171] L. Zheng et al., "Judging LLM-as-a-judge with MT-bench and chatbot arena," 2023, *arXiv:2306.05685*.

[172] M. Maaz, H. Rasheed, S. Khan, and F. S. Khan, "Video-chatGPT: Towards detailed video understanding via large vision and language models," 2023, *arXiv:2306.05424*.

[173] R. Schaeffer, B. Miranda, and S. Koyejo, "Are emergent abilities of large language models a mirage?," 2023, *arXiv:2304.15004*.

[174] Y.-T. Lin and Y.-N. Chen, "LLM-Eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models," 2023, *arXiv:2305.13711*.

[175] P. Lu et al., "Chameleon: Plug-and-play compositional reasoning with large language models," 2023, *arXiv:2304.09842*.

[176] Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, and A. Smola, "Multimodal chain-of-thought reasoning in language models," 2023, *arXiv:2302.00923*.

[177] H. Lovenia, W. Dai, S. Cahyawijaya, Z. Ji, and P. Fung, "Negative object presence evaluation (NOPE) to measure object hallucination in vision-language models," *arXiv:2310.05338*, 2023.

[178] N. Mündler, J. He, S. Jenko, and M. Vechev, "Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation," 2023, *arXiv:2305.15852*.

[179] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J.-R. Wen, "Evaluating object hallucination in large vision-language models," 2023, *arXiv:2305.10355*.
[180] M. Ranjit, G. Ganapathy, R. Manuel, and T. Ganu, "Retrieval augmented chest X-ray report generation using openai GPT models," 2023, *arXiv:2305.03660*.
[181] J. Liu, J. Jin, Z. Wang, J. Cheng, Z. Dou, and J.-R. Wen, "RETA-LLM: A retrieval-augmented large language model toolkit," 2023, *arXiv:2306.05212*.
[182] J. Pan et al., "Retrieving-to-answer: Zero-shot video question answering with frozen large language models," 2023, *arXiv:2306.11732*.
[183] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J. Wen, "Evaluating object hallucination in large vision-language models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2023, pp. 292–305.
[184] F. Liu, K. Lin, L. Li, J. Wang, Y. Yacoob, and L. Wang, "Mitigating hallucination in large multi-modal models via robust instruction tuning," 2023, *arXiv:2306.14565*.
[185] J. Wang et al., "Evaluation and analysis of hallucination in large vision-language models," 2023, *arXiv:2308.15126*.
[186] X. Chen et al., "Unified hallucination detection for multimodal large language models," 2024, *arXiv:2402.03190*.
[187] A. Gunjal, J. Yin, and E. Bas, "Detecting and preventing hallucinations in large vision language models," *arXiv:2308.06394*, 2023.
[188] S. Leng et al., "Mitigating object hallucinations in large vision-language models through visual contrastive decoding," 2023, *arXiv:2311.16922*.
[189] A. Deng, Z. Chen, and B. Hooi, "Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding," *arXiv:2402.15300*, 2024.
[190] Z. Han, Z. Bai, H. Mei, Q. Xu, C. Zhang, and M. Z. Shou, "Skip n: A simple method to reduce hallucination in large vision-language models," 2024, *arXiv:2402.01345*.
[191] Z. Bai et al., "Hallucination of multimodal large language models: A survey," 2024, *arXiv:2404.18930*.
[192] C. Zhang et al., "Faster segment anything: Towards lightweight sam for mobile applications," 2023, *arXiv:2306.14289*.
[193] Y. Sun et al., "Retentive network: A successor to transformer for large language models," 2023, *arXiv:2307.08621*.
[194] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 3–5.
[195] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLORA: Efficient finetuning of quantized LLMs," 2023, *arXiv:2305.14314*.
[196] A. Chavan, Z. Liu, D. Gupta, E. Xing, and Z. Shen, "One-for-all: Generalized LoRA for parameter-efficient fine-tuning," 2023, *arXiv:2306.07967*.
[197] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," 2021, *arXiv:2101.00190*.
[198] X. Liu et al., "P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks," 2021, *arXiv:2110.07602*.
[199] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," 2021, *arXiv:2104.08691*.
[200] T. Chen et al., "Sam fails to segment anything?–sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, and more," 2023, *arXiv:2304.09148*.
[201] C. Szegedy, "Intriguing properties of neural networks," 2013, *arXiv:1312.6199*.
[202] A. Madry, "Towards deep learning models resistant to adversarial attacks," 2017, *arXiv: 1706.06083*.
[203] A. Muhammad, F. Zhou, C. Xie, J. Li, S.-H. Bae, and Z. Li, "MixACM: Mixup-based robustness transfer via distillation of activated channel maps," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 4555–4569.
[204] "How to jailbreak chatGPT," 2023. Accessed: Jul. 4, 2023. [Online]. Available: https://watcher.guru/news/how-to-jailbreak-chatgpt
[205] F. Perez and I. Ribeiro, "Ignore previous prompt: Attack techniques for language models," 2022, *arXiv:2211.09527*.
[206] "Red-teaming large-language models," 2023. Accessed: Jul. 4, 2023. [Online]. Available: https://huggingface.co/blog/red-teaming
[207] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz, "More than you've asked for: A comprehensive analysis of novel prompt injection threats to application-integrated large language models," 2023, *arXiv:2302.12173*.
[208] N. Carlini et al., "Are aligned neural networks adversarially aligned?," 2023, *arXiv:2306.15447*.
[209] N. Maus, P. Chao, E. Wong, and J. Gardner, "Adversarial prompting for black box foundation models," 2023, *arXiv:2302.04237*.
[210] J. Wang, Z. Liu, K. H. Park, M. Chen, and C. Xiao, "Adversarial demonstration attacks on large language models," 2023, *arXiv:2305.14950*.
[211] A. Hanif et al., "Baple: Backdoor attacks on medical foundational models using prompt learning," 2024, *arXiv:2408.07440*.
[212] C. Zhang, C. Zhang, T. Kang, D. Kim, S.-H. Bae, and I. S. Kweon, "Attack-sam: Towards evaluating adversarial robustness of segment anything model," 2023, *arXiv:2305.00866*.
[213] Y. Huang et al., "On the robustness of segment anything," 2023, *arXiv:2305.16220*.
[214] A. Shtedritski, C. Rupprecht, and A. Vedaldi, "What does CLIP know about a red circle? Visual prompt engineering for VLMS," 2023, *arXiv:2304.06712*.
[215] S. M. Hall, F. G. Abrantes, H. Zhu, G. Sodunke, A. Shtedritski, and H. R. Kirk, "VisoGender: A dataset for benchmarking gender bias in image-text pronoun resolution," 2023, *arXiv:2306.12424*.
[216] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 24824–24837.
[217] T. Gupta and A. Kembhavi, "Visual programming: Compositional visual reasoning without training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 14953–14962.
[218] "Stable diffusion v1–1 model card," Accessed: May 9, 2024. [Online]. Available: https://huggingface.co/CompVis/stable-diffusion-v1--1
[219] D. Podell et al., "SDXL: Improving latent diffusion models for high-resolution image synthesis," 2023, *arXiv:2307.01952*.
[220] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "StyleCLIP: Text-driven manipulation of stylegan imagery," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 2085–2094.
[221] G. Kim, T. Kwon, and J. C. Ye, "DiffusionCLIP: Text-guided diffusion models for robust image manipulation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2416–2425.

Muhammad Awais received the PhD degree from Kyung-Hee University. He is currently working toward the MS degree with the Georgia Institute of Technology. He is a research associate with the MBZ University of Artificial Intelligence (MBZUAI). Before his role at MBZUAI, he worked as a researcher in various industrial and academic labs, including the AI Theory Group with Huawei's Noah Ark Lab in Hong Kong, Sony AI in Japan, and Kyung-Hee University in South Korea. His contributions to the fields of artificial intelligence and computer vision have been published in prestigious venues such as NeurIPS, ICCV, ECCV, MICCAI and TNNLS.

Muzammal Naseer received the PhD degree from the Australian National University (ANU), in 2022, where he received a competitive postgraduate scholarship. He is an assistant professor with Computer Science Department, College of Computing and Mathematical Sciences, Khalifa University. He served as a Researcher at Data61, CSIRO, Inception Institute of Artificial Intelligence, and Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI) from 2018-2024. He is interested in building Robust Intelligent Systems. His research focuses on robust visual-spatial and temporal perception, understanding and explaining AI behavior through adversarial machine learning, representation learning through self-learning (self-supervision, self-distillation, self-critique, self-reflection), and configuring the role of large language models (LLMs) in building robust AI systems across applications of life sciences and security. He has published at well-recognized machine learning and computer vision venues, including *IEEE Transactions on Pattern Analysis and Machine Intelligence*, CVPR, NeurIPS, ICLR, ICCV, and AAAI.

Salman Khan received the PhD degree from the University of Western Australia, in 2016. He is an associate professor with the MBZ University of Artificial Intelligence. He has been an adjunct faculty with the Australian National University since 2016. He has been awarded the Outstanding Reviewer award at IEEE CVPR multiple times, won the Best Paper award at 9th ICPRAM 2020, and won 2nd prize in the NTIRE Image Enhancement Competition alongside CVPR 2019. He served as a (senior) program committee member for several premier conferences, including CVPR, ICCV, ICML, ECCV, and NeurIPS. His thesis received an honorable mention on the Dean's List Award. He has published more than 100 papers in high-impact scientific journals and conferences. His research interests include computer vision and machine learning.

Rao Muhammad Anwer is an assistant professor of computer vision with MBZUAI. Before joining MBZUAI, Anwer was with the Inception Institute of Artificial Intelligence (IIAI) in Abu Dhabi, United Arab Emirates, working as a research scientist. Before joining IIAI, he was a researcher with Aalto University, Finland. His research interests include visual object recognition, pedestrian detection and action recognition, efficient and robust deep learning models for comprehensive scene understanding, and human visual relationship detection.

Hisham Cholakkal received the MTech degree from IIT Guwahati, India, and the PhD degree from Nanyang Technological University, Singapore. He is an assistant professor with the MBZ University of Artificial Intelligence (MBZUAI), UAE. Previously, from 2018 to 2020, he worked as a research scientist with the Inception Institute of Artificial Intelligence, UAE; from 2016 to 2018, he worked as a senior technical lead with Mercedes-Benz R&D India. He has also worked as a researcher with BEL-Central Research Lab, India, and Advanced Digital Sciences Center, Singapore. He has organized workshops at top venues such as ICCV 2023, NeurIPS 2022, and ACCV 2022 and regularly serves as a program committee member for top conferences, including CVPR, ICCV, NeurIPS, ICLR, and ECCV.

Mubarak Shah (Life Fellow, IEEE) is currently the Trustee chair professor of computer science and the founding director of the Center for Research in Computer Vision, University of Central Florida. He is an editor of an international book series on video computing, an editor-in-chief of the *Journal of Machine Vision and Applications*, and an associate editor of *ACM Computing Surveys journal*. His research interests include video surveillance, visual tracking, human activity recognition, visual analysis of crowded scenes, video registration, and UAV video analysis. He is a fellow of the AAAS, IAPR, and SPIE.

Ming-Hsuan Yang (Fellow, IEEE) is a professor of computer science and engineering with the University of California at Merced, and an adjunct professor of computer science with Yonsei University. He is a co-editor-in-chief of Computer Vision and Image Understanding, an associate editor-in-chief of *IEEE Transactions on Pattern Analysis and Machine Intelligence*, and an associate editor of the *International Journal of Computer Vision*. He received the Longuet-Higgins Prize in 2023, Best Paper Honorable Mention in CVPR 2018, NSF CAREER Award in 2012, and Google Faculty Award in 2009. He is a fellow of the ACM.

Fahad Shahbaz Khan received the MSc degree in intelligent systems design from the Chalmers University of Technology, Sweden, and the PhD degree in computer vision from Computer Vision Center Barcelona and Autonomous University of Barcelona, Spain. He is currently a professor of computer vision with MBZUAI, United Arab Emirates. He also holds a faculty position at Computer Vision Laboratory, Linköping University, Sweden. He has achieved top ranks on various international challenges (Visual Object Tracking VOT: 1st 2014 and 2018, 2nd 2015, 1st 2016; VOT-TIR: 1st 2015 and 2016; OpenCV Tracking: 1st 2015; 1st PASCAL VOC Segmentation and Action Recognition tasks 2010). He received the best paper award in the computer vision track at IEEE ICPR 2016. He has published more than 100 peer-reviewed conference papers, journal articles, and book contributions. His research interests include a wide range of topics within computer vision and machine learning. He serves as a regular senior program committee member for leading conferences such as CVPR, ICCV, and NeurIPS.