Multimodal Foundation Models: From Specialists to General-Purpose Assistants

Other titles in Foundations and Trends® in Computer Graphics and Vision

Computational Imaging Through Atmospheric Turbulence
Stanley H. Chan and Nicholas Chimitt
ISBN: 978-1-63828-999-9

Towards Better User Studies in Computer Graphics and Vision
Zoya Bylinskii, Laura Herman, Aaron Hertzmann, Stefanie Hutka and Yile Zhang
ISBN: 978-1-63828-172-6

An Introduction to Neural Data Compression
Yibo Yang, Stephan Mandt and Lucas Theis
ISBN: 978-1-63828-174-0

Learning-based Visual Compression
Ruolei Ji and Lina J. Karam
ISBN: 978-1-63828-112-2

Vision-Language Pre-Training: Basics, Recent Advances, and Future Trends
Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu and Jianfeng Gao
ISBN: 978-1-63828-132-0

Semantic Image Segmentation: Two Decades of Research
Gabriela Csurka, Riccardo Volpi and Boris Chidlovskii
ISBN: 978-1-63828-076-7

Multimodal Foundation Models: From Specialists to General-Purpose Assistants

Chunyuan Li
Microsoft Corporation
chunyl@microsoft.com

Zhe Gan
Microsoft Corporation
zhgan@microsoft.com

Zhengyuan Yang
Microsoft Corporation
zhengyang@microsoft.com

Jianwei Yang
Microsoft Corporation
jianwyan@microsoft.com

Linjie Li
Microsoft Corporation
linjli@microsoft.com

Lijuan Wang
Microsoft Corporation
lijuanw@microsoft.com

Jianfeng Gao
Microsoft Corporation
jfgao@microsoft.com

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Computer Graphics and Vision

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

C. Li *et al.*. *Multimodal Foundation Models: From Specialists to General-Purpose Assistants*. Foundations and Trends® in Computer Graphics and Vision, vol. 16, no. 1-2, pp. 1–214, 2024.

ISBN: 978-1-63828-337-9

Foundations and Trends® in Computer Graphics and Vision

Volume 16, Issue 1-2, 2024

Editorial Scope

Topics

Foundations and Trends® in Computer Graphics and Vision publishes survey and tutorial articles in the following topics:

- Rendering
- Shape
- Mesh simplification
- Animation
- Sensors and sensing
- Image restoration and enhancement
- Segmentation and grouping
- Feature detection and selection
- Color processing
- Texture analysis and synthesis
- Illumination and reflectance modeling
- Shape representation
- Tracking
- Calibration
- Structure from motion
- Motion estimation and registration
- Stereo matching and reconstruction
- 3D reconstruction and image-based modeling
- Learning and statistical methods
- Appearance-based matching
- Object and scene recognition
- Face detection and recognition
- Activity and gesture recognition
- Image and video retrieval
- Video analysis and event recognition
- Medical image analysis
- Robot localization and navigation

Information for Librarians

Foundations and Trends® in Computer Graphics and Vision, 2024, Volume 16, 4 issues. ISSN paper version 1572-2740. ISSN online version 1572-2759. Also available as a combined paper and online subscription.

Contents

Multimodal Foundation Models: From Specialists to General-Purpose Assistants

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang and Jianfeng Gao

Microsoft Corporation, USA; chunyl@microsoft.com, zhgan@microsoft.com, zhengyang@microsoft.com, jianwyan@microsoft.com, linjli@microsoft.com, lijuanw@microsoft.com, jfgao@microsoft.com

ABSTRACT

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

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang and Jianfeng Gao (2024), "Multimodal Foundation Models: From Specialists to General-Purpose Assistants", Foundations and Trends® in Computer Graphics and Vision: Vol. 16, No. 1-2, pp 1–214. DOI: 10.1561/0600000110.

monograph are researchers, graduate students, and professionals in computer vision and vision-language multimodal communities who are eager to learn the basics and recent advances in multimodal foundation models.

1

Introduction

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

In Figure 1.1, the evolution of AI is depicted, beginning with a specialized entity symbolized by a desktop computer and progressing to a versatile, general-purpose assistant represented by Doraemon. The choice of Doraemon serves a dual purpose: to highlight AI's capability to perform a broad spectrum of tasks and to emphasize its constant readiness to support human needs.

Over the last decade, the field of AI has experienced a fruitful trajectory in the development of models. We divide them into four categories, as illustrated in Figure 1.2. The categorization can be shared among different fields in AI, including language, vision and multimodality. We

Figure 1.1: The visual illustration of AI evolution from specialists to general-purpose assistants.

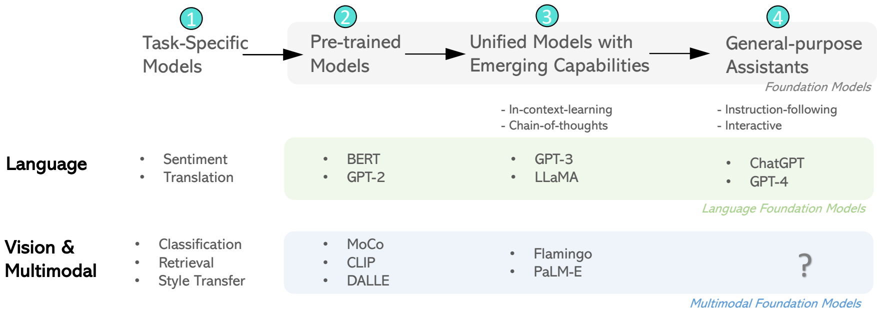

Figure 1.2: Illustration of foundation model development trajectory for language and vision/multi-modality. Among the four categories, the first category is the task-specific model, and the last three categories belong to foundation models, where these foundation models for language and vision are grouped in green and blue blocks, respectively. Some prominent properties of models in each category are highlighted. By comparing the models between language and vision, we are foreseeing that the transition of multimodal foundation models follows a similar trend: from the pre-trained model for specific purpose, to unified models and general-purpose assistants. However, research exploration is needed to figure out the best recipe, which is indicated as the question mark in the figure, as technical details of GPT-4V and Gemini [426] stay private.

first use language models in NLP to illustrate the evolution process. (*i*) At the early years, task-specific models are developed for individual datasets and tasks, typically being trained from scratch. (*ii*) With large-scale pre-training, language models achieve state-of-the-art performance on many established language understanding and generation tasks, such as BERT [98], RoBERTa [277], T5 [363], DeBERTa [167] and GPT-2 [362]). These pre-trained models serve the basis for downstream task adaptation. (*iii*) Exemplified by GPT-3 [37], large language models (LLMs) unify various language understanding and generation tasks into one model. With web-scale training and unification, some emerging capabilities appear, such as in-context-learning and chain-of-thoughts. (*iv*) With recent advances in human-AI alignment, LLMs start to play the role of general-purpose assistants to follow human intents to complete a wide range of language tasks in the wild, such as ChatGPT [333] and GPT-4 [334]. These assistants exhibit interesting capabilities, such as interaction and tool use, and lay a foundation for developing general-purpose AI agents. It is important to note that the latest iterations of foundation models build upon the noteworthy features of their earlier counterparts while also providing additional capabilities.

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

In this monograph, we limit the scope of multimodal foundation models to the vision and vision-language domains. Recent survey papers on related topics include (*i*) *image understanding models* such as self-supervised learning [193], [196], [339], segment anything (SAM) [552], [554], (*ii*) *image generation models* [553], [592], and (*iii*) *vision-language pre-training (VLP)*. Existing VLP survey papers cover VLP methods for task-specific VL problems before the era of pre-training, image-text tasks, core vision tasks, and/or video-text tasks [54], [109], [129], [231], [379], [551], [560]. Two recent survey papers cover the integration of vision models with LLM [16], [525].

Among them, [129] is a survey on VLP that covers the CVPR tutorial series on *Recent Advances in Vision-and-Language Research* in 2022 and before. This work summarizes the CVPR tutorial on *Recent Advances in Vision Foundation Models* in 2023. Different from the aforementioned survey papers that focus on literature review of a given research topic, this monograph presents our perspectives on the role transition of multimodal foundation models from specialists to general-purpose visual assistants, in the era of large language models. The contributions of this survey are summarized as follows.

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

1.1 What are Multimodal Foundation Models?

As elucidated in the Stanford foundation model paper [35], AI has been undergoing a paradigm shift with the rise of models (*e.g.*, BERT, GPT family, CLIP [360] and DALL-E [367]) trained on broad data that can be adapted to a wide range of downstream tasks. They call these models *foundation models* to underscore their critically central yet incomplete character: homogenization of the methodologies across research communities and emergence of new capabilities. From a technical perspective, it is *transfer learning* that makes foundation models possible, and it is *scale* that makes them powerful. The emergence of foundation models has been predominantly observed in the NLP domain, with examples ranging from BERT to ChatGPT. This trend has gained traction in recent years, extending to computer vision and other fields. In NLP, the introduction of BERT in late 2018 is considered as the inception of the foundation model era. The remarkable success of BERT rapidly stimulates interest in self-supervised learning in the computer vision community, giving rise to models such as SimCLR [62], MoCo [163], BEiT [26], and MAE [162]. During the same time period, the success of pre-training also significantly promotes the vision-and-language multimodal field to an unprecedented level of attention.

In this monograph, *we focus on multimodal foundation models, which inherit all properties of foundation models discussed in the Stanford paper [35], but with an emphasis on models with the capability to deal with vision and vision-language modalities.* Among the ever-growing literature, we categorize multimodal foundation models in Figure 1.3, based on their functionality and generality. For each category, we present exemplary models that demonstrate the primary capabilities inherent to these multimodal foundation models.

- **Visual Understanding Models.** (Highlighted with orange in Figure 1.3) Learning general visual representations is essential to build vision foundation models, as pre-training a strong vision backbone is fundamental to all types of computer vision downstream tasks, ranging from image-level (*e.g.*, image classification, retrieval, and captioning), region-level (*e.g.*, detection and grounding) to pixel-level tasks (*e.g.*, segmentation). We group the methods into three categories, depending on the types of supervision signals used to train the models.

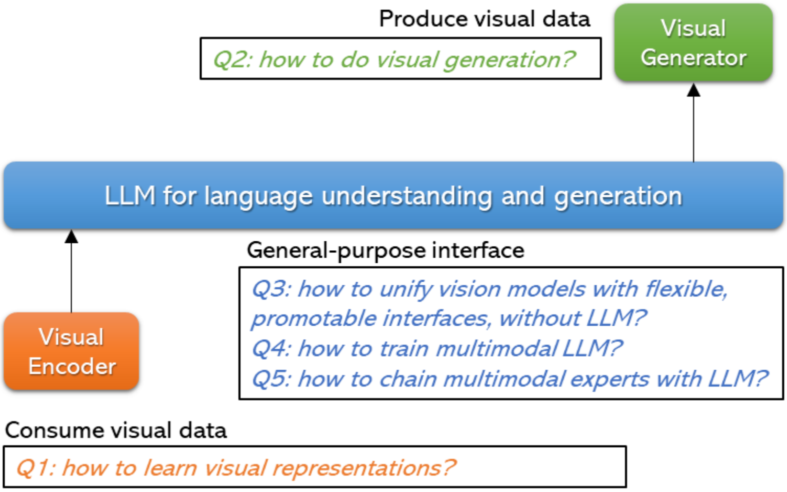

Figure 1.3: Illustration of three representative problems that multimodal foundation models aim to solve in this monograph: visual understanding tasks (orange), visual generation tasks (green), and general-purpose interface (blue) with language understanding and generation.

- **Label supervision.** Datasets like ImageNet [212] and ImageNet21K [374] have been popular for supervised learning, and larger-scale proprietary datasets are also used in industrial labs [406], [415], [548].
- **Language supervision.** Language is a richer form of supervision. Models like CLIP [360] and ALIGN [195] are pre-trained using a contrastive loss over millions or even billions of noisy image-text pairs mined from the Web. These models enable zero-shot image classification, and make traditional computer vision (CV) models to perform open-vocabulary CV tasks. We advocate the concept of *computer vision in the wild*,[1] and encourage the development and evaluation of future foundation models for this.
- **Image-only self-supervision.** This line of work aims to learn image representations from supervision signals mined from the images themselves, ranging from contrastive learning [62], [163],

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

non-contrastive learning [46], [77], [144], to masked image modeling [26], [162].

  – **Multimodal fusion, region-level and pixel-level pre-training.** Besides the methods of pre-training image backbones, we will also discuss pre-training methods that allow multimodal fusion (*e.g.*, CoCa [530], Flamingo [4]), region-level and pixel-level image understanding, such as open-set object detection (*e.g.*, GLIP [241]) and promptable semgentation (*e.g.*, SAM [206]). These methods typically rely on a pre-trained image encoder or a pre-trained image-text encoder pair.

- **Visual Generation Models.** (Highlighted with green in Figure 1.3) Recently, foundation image generation models have been built, due to the emergence of large-scale image-text data. The techniques that make it possible include the vector-quantized VAE methods [370], diffusion-based models [99] and auto-regressive models.

  – **Text-conditioned visual generation.** This research area focuses on generating faithful visual content, including images, videos, and more, conditioned on open-ended text descriptions/prompts. Text-to-image generation develops generative models that synthesize images of high fidelity to follow the text prompt. Prominent examples include DALL-E [367], DALL-E 2 [366], Stable Diffusion [375], [412], Imagen [382], and Parti [531]. Building on the success of text-to-image generation models, text-to-video generation models generate videos based on text prompts, such as Imagen Video [172] and Make-A-Video [403].

  – **Human-aligned visual generator.** This research area focuses on improving the pre-trained visual generator to better follow human intentions. Efforts have been made to address various challenges inherent to base visual generators. These include improving spatial controllability [511], [562], ensuring better adherence to text prompts [31], supporting flexible text-based editing [36], and facilitating visual concept customization [380].

- **General-purpose Interface.** (Highlighted with blue in Figure 1.3) The aforementioned multimodal foundation models are designed for specific purposes – tackling a specific set of CV problems/tasks. Recently, we see an emergence of general-purpose models that lay the basis of AI agents. Existing efforts focus on three research topics. The first topic aims to unify models for visual understanding and generation. These models are inspired by the unification spirit of LLMs in NLP, but do not explicitly leverage pre-trained LLM in modeling. In contrast, the other two topics embrace and involve LLMs in modeling, including training and chaining with LLMs, respectively.

– **Unified vision models for understanding and generation.** In computer vision, several attempts have been made to build a general-purpose foundation model by combining the functionalities of specific-purpose multimodal models. To this end, a unified model architecture is adopted for various downstream computer vision and vision-language (VL) tasks. There are different levels of unification. First, a prevalent effort is to bridge vision and language by converting all closed-set vision tasks to open-set ones, such as CLIP [360], GLIP [242], OpenSeg [137], *etc.* Second, the unification of different VL understanding tasks across different granularity levels is also actively explored, such as I/O unification methods like UniTAB [508], Unified-IO [287]), Pix2Seq-v2 [65] and functional unification methods like GPV [151], GLIP-v2 [559]) and X-Decoder [599]. In the end, it is also necessitated to make the models more interactive and promptable like ChatGPT, and this has been recently studied in SAM [206] and SEEM [601].
– **Training with LLMs.** Similar to the behavior of LLMs, which can address a language task by following the instruction and processing examples of the task in their text prompt, it is desirable to develop a visual and text interface to steer the model towards solving a multimodal task. By extending the capability of LLMs to multimodal settings and training the model end-to-end, multimodal LLMs or large multimodal models are developed, including Flamingo [4] and Multimodal GPT-4 [334].
– **Chaining tools with LLM.** Exploiting the tool use capabilities of LLMs, an increasing number of studies integrate LLMs such as ChatGPT with various multimodal foundation models

to facilitate image understanding and generation through a conversation interface. This interdisciplinary approach combines the strengths of NLP and computer vision, enabling researchers to develop more robust and versatile AI systems that are capable of processing visual information and generating human-like responses via human-computer conversations. Representative works include Visual ChatGPT [477] and MM-ReAct [513].

1.2 Definition and Transition from Specialists to General-Purpose Assistants

Based on the model development history and taxonomy in NLP, we group multimodal foundation models in Figure 1.3 into two categories.

- **Specific-Purpose Pre-trained Vision Models** cover most existing multimodal foundation models, including visual understanding models (*e.g.*, CLIP [360], SimCLR [62], BEiT [26], SAM [206]) and visual generation models (*e.g.*, Stable Diffusion [375], [412]), as they present powerful transferable ability for specific vision problems.
- **General-Purpose Assistants** refer to AI agents that can follow human intents to complete various computer vision tasks in the wild. The meanings of general-purpose assistants are two-fold: (*i*) generalists with unified architectures that could complete tasks across different problem types, and (*ii*) easy to follow human instruction, rather than replacing humans. To this end, several research topics have been actively explored, including unified vision modeling [287], [559], [599], training and chaining with LLMs [264], [477], [513], [593].

1.3 Who Should Read this Monograph?

This monograph is based on our CVPR 2023 tutorial,[2] with researchers in the computer vision and vision-language multimodal communities as our primary target audience. It reviews the literature and explains topics to those who seek to learn the basics and recent advances in

[2] https://vlp-tutorial.github.io/2023/index.html

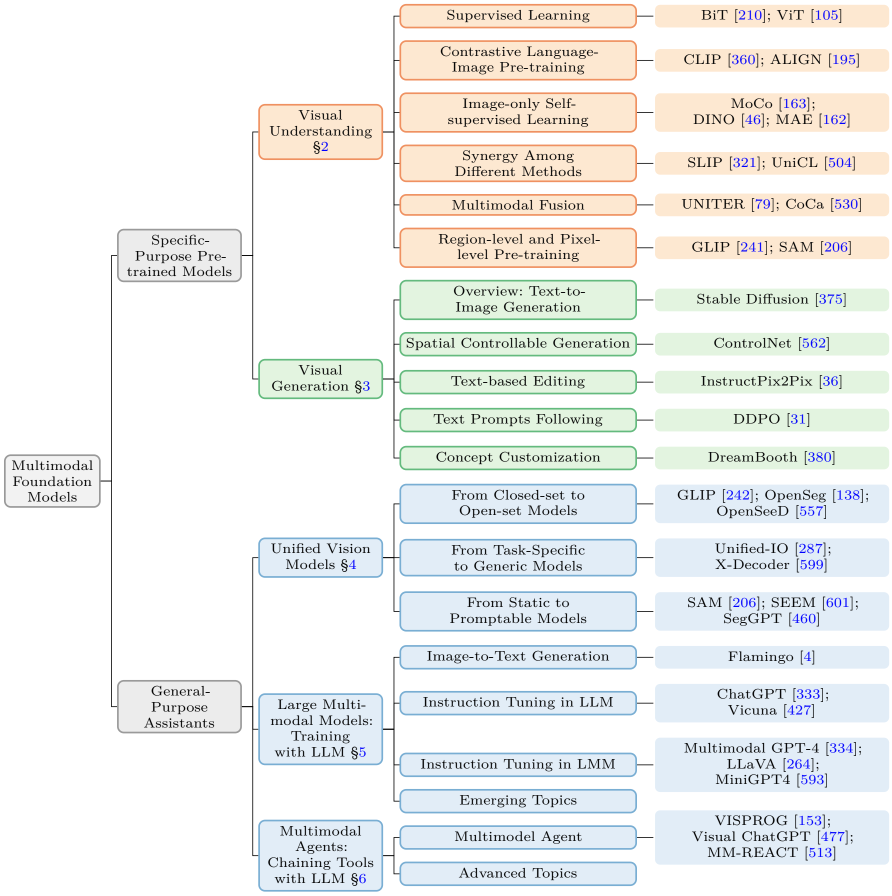

Figure 1.4: An overview of the monograph's structure, detailing Sections 2-6.

multimodal foundation models. The target audiences are graduate students, researchers and professionals who are not experts of multimodal foundation models but are eager to develop perspectives and learn the trends in the field. The structure of this monograph is illustrated in Figure 1.4. It consists of 7 sections.

- Section 1 introduces the landscape of multimodal foundation model research, and presents a historical view on the transition of research from specialists to general-purpose assistants.

- Section 2 introduces different ways to consume visual data, with a focus on how to learn a strong image backbone.
- Section 3 describes how to produce visual data that aligns with human intents.
- Section 4 describes how to design unified vision models, with an interface that is interactive and promptable, especially when LLMs are not employed.
- Section 5 describes how to train an LLM in an end-to-end manner to consume visual input for understanding and reasoning.
- Section 6 describes how to chain multimodal tools with an LLM to enable new capabilities.
- Section 7 concludes the monograph and discusses research trends.

Relations among Sections 2-6. Sections 2-6 are the core sections of this survey. An overview of the structure for these sections are provided in Figure 1.3. We start with a discussion of two typical multimodal foundation models for specific tasks, including visual understanding in Section 2 and visual generation in Section 3. As the notion of multimodal foundation models are originally based on visual backbone/representation learning for understanding tasks, we first present a comprehensive review to the transition of image backbone learning methods, evolving from early supervised methods to the recent language-image contrastive methods, and extend the discussion on image representations from image-level to region-level and pixel-level (Section 2). Recently, generative AI is becoming increasingly popular, where vision generative foundation models have been developed. In Section 3, we discuss large pre-trained text-to-image models, and various ways that the community leverage the generative foundation models to develop new techniques to make them better aligned with human intents. Inspired by the recent advances in NLP that LLMs serve as general-purpose assistants for a wide range of language tasks in daily life, the computer vision community has been anticipating and attempting to build general-purpose visual assistants. We discuss three different ways to build general-purpose assistants. Inspired by the spirit of LLMs, Section 4 focuses on unifying different vision models of understanding and generation without explicitly incorporating LLMs in modeling. In contrast, Section 5 and Section 6 focus on embracing LLMs to build general-purpose visual assistants, by explicitly augmenting LLMs in modeling. Specifically, Section 5 describes end-to-end training methods, and Section 6 focuses on training-free approaches that chain various vision models to LLMs.

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

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

1.4 Related Materials: Slide Decks and Pre-recorded Talks

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

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

References

[1] A. Agarwal, S. Karanam, K. Joseph, A. Saxena, K. Goswami, and B. V. Srinivasan, “A-star: Test-time attention segregation and retention for text-to-image synthesis,” *arXiv preprint arXiv:2306.14544*, 2023.

[2] H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee, and P. Anderson, “Nocaps: Novel object captioning at scale,” in *ICCV*, 2019.

[3] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, *et al.*, “Do as i can, not as i say: Grounding language in robotic affordances,” *arXiv preprint arXiv:2204.01691*, 2022.

[4] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, *et al.*, “Flamingo: A visual language model for few-shot learning,” *arXiv preprint arXiv:2204.14198*, 2022.

[5] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut, “Text summarization techniques: A brief survey,” *arXiv preprint arXiv:1707.02268*, 2017.

[6] E. Amrani, L. Karlinsky, and A. Bronstein, “Self-supervised classification network,” in *ECCV*, 2022.

[7] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *CVPR*, 2018.

[8] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *ICCV*, 2015.

[9] S. Arora, H. Khandeparkar, M. Khodak, O. Plevrakis, and N. Saunshi, "A theoretical analysis of contrastive unsupervised representation learning," *arXiv preprint arXiv:1902.09229*, 2019.

[10] M. Assran, M. Caron, I. Misra, P. Bojanowski, F. Bordes, P. Vincent, A. Joulin, M. Rabbat, and N. Ballas, "Masked siamese networks for label-efficient learning," in *ECCV*, 2022.

[11] O. Avrahami, K. Aberman, O. Fried, D. Cohen-Or, and D. Lischinski, "Break-a-scene: Extracting multiple concepts from a single image," *arXiv preprint arXiv:2305.16311*, 2023.

[12] O. Avrahami, O. Fried, and D. Lischinski, "Blended latent diffusion," *arXiv preprint arXiv:2206.02779*, 2022.

[13] O. Avrahami, T. Hayes, O. Gafni, S. Gupta, Y. Taigman, D. Parikh, D. Lischinski, O. Fried, and X. Yin, "Spatext: Spatio-textual representation for controllable image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18 370–18 380, 2023.

[14] O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18 208–18 218, 2022.

[15] A. Awadalla, I. Gao, J. Gardner, J. Hessel, Y. Hanafy, W. Zhu, K. Marathe, Y. Bitton, S. Gadre, J. Jitsev, S. Kornblith, P. W. Koh, G. Ilharco, M. Wortsman, and L. Schmidt, *Openflamingo*, version v0.1.1, Mar. 2023. DOI: 10.5281/zenodo.7733589.

[16] M. Awais, M. Naseer, S. Khan, R. M. Anwer, H. Cholakkal, M. Shah, M.-H. Yang, and F. S. Khan, "Foundational models defining a new era in vision: A survey and outlook," *arXiv preprint arXiv:2307.13721*, 2023.

[17] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," *NeurIPS*, 2019.

[18] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *ICML*, 2022.

[19] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *ICLR*, 2015.

[20] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A frontier large vision-language model with versatile abilities," *arXiv preprint arXiv:2308.12966*, 2023.

[21] S. Bai, S. Yang, J. Bai, P. Wang, X. Zhang, J. Lin, X. Wang, C. Zhou, and J. Zhou, *Touchstone: Evaluating vision-language models by language models*, 2023. URL: https://arxiv.org/abs/2308.16890.

[22] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, K. Kreis, M. Aittala, T. Aila, S. Laine, B. Catanzaro, *et al.*, "Ediffi: Text-to-image diffusion models with an ensemble of expert denoisers," *arXiv preprint arXiv:2211.01324*, 2022.

[23] I. Balažević, D. Steiner, N. Parthasarathy, R. Arandjelović, and O. J. Hénaff, "Towards in-context scene understanding," *arXiv preprint arXiv:2306.01667*, 2023.

[24] A. Bansal, K. Sikka, G. Sharma, R. Chellappa, and A. Divakaran, "Zero-shot object detection," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 384–400, 2018.

[25] A. Bansal, H.-M. Chu, A. Schwarzschild, S. Sengupta, M. Goldblum, J. Geiping, and T. Goldstein, "Universal guidance for diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 843–852, 2023.

[26] H. Bao, L. Dong, and F. Wei, "BEiT: Bert pre-training of image transformers," in *ICLR*, 2022.

[27] A. Bar, Y. Gandelsman, T. Darrell, A. Globerson, and A. Efros, "Visual prompting via image inpainting," *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 25 005–25 017.

[28] A. Bardes, J. Ponce, and Y. LeCun, "Vicreg: Variance-invariance-covariance regularization for self-supervised learning," *arXiv preprint arXiv:2105.04906*, 2021.

[29] J. Berant, A. Chou, R. Frostig, and P. Liang, "Semantic parsing on freebase from question-answer pairs," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1533–1544, 2013.

[30] Y. Bitton, H. Bansal, J. Hessel, R. Shao, W. Zhu, A. Awadalla, J. Gardner, R. Taori, and L. Schimdt, *Visit-bench: A benchmark for vision-language instruction following inspired by real-world use*, 2023. arXiv: 2308.06595.

[31] K. Black, M. Janner, Y. Du, I. Kostrikov, and S. Levine, "Training diffusion models with reinforcement learning," *arXiv preprint arXiv:2305.13301*, 2023.

[32] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, V. Jampani, and R. Rombach, "Stable video diffusion: Scaling latent video diffusion models to large datasets," 2023.

[33] A. Blattmann, R. Rombach, K. Oktay, and B. Ommer, "Retrieval-augmented diffusion models," *arXiv preprint arXiv:2204.11824*, 2022.

[34] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact: Real-time instance segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9157–9166, 2019.

[35] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[36] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18 392–18 402, 2023.

[37] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," in *NeuIPS*, 2020.

[38] M. Byeon, B. Park, H. Kim, S. Lee, W. Baek, and S. Kim, *Coyo-700m: Image-text pair dataset*, 2022. URL: https://github.com/kakaobrain/coyo-dataset.

[39] M. Cai, H. Liu, S. K. Mustikovela, G. P. Meyer, Y. Chai, D. Park, and Y. J. Lee, "Making large multimodal models understand arbitrary visual prompts," in *arXiv:2312.00784*, 2023.

[40] T. Cai, X. Wang, T. Ma, X. Chen, and D. Zhou, "Large language models as tool makers," *arXiv preprint arXiv:2305.17126*, 2023.

[41] Z. Cai, G. Kwon, A. Ravichandran, E. Bas, Z. Tu, R. Bhotika, and S. Soatto, "X-detr: A versatile architecture for instance-wise vision-language tasks," in *ECCV*, 2022.

[42] L. Cao, B. Zhang, C. Chen, Y. Yang, X. Du, W. Zhang, Z. Lu, and Y. Zheng, "Less is more: Removing text-regions improves clip training efficiency and robustness," *arXiv preprint arXiv:2305.05095*, 2023.

[43] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *ECCV*, 2020.

[44] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *ECCV*, 2018.

[45] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *NeurIPS*, 2020.

[46] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *ICCV*, 2021.

[47] L. Castrejon, K. Kundu, R. Urtasun, and S. Fidler, "Annotating object instances with a polygon-rnn," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5230–5238, 2017.

[48] H. Chang, H. Zhang, J. Barber, A. Maschinot, J. Lezama, L. Jiang, M.-H. Yang, K. Murphy, W. T. Freeman, M. Rubinstein, *et al.*, "Muse: Text-to-image generation via masked generative transformers," *arXiv preprint arXiv:2301.00704*, 2023.

[49] H. Chang, H. Zhang, L. Jiang, C. Liu, and W. T. Freeman, "Maskgit: Masked generative image transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11 315–11 325, 2022.

[50] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *CVPR*, 2021.

[51] H. Chefer, Y. Alaluf, Y. Vinker, L. Wolf, and D. Cohen-Or, "Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models," *arXiv preprint arXiv:2301.13826*, 2023.

[52] C. Chen, B. Zhang, L. Cao, J. Shen, T. Gunter, A. M. Jose, A. Toshev, J. Shlens, R. Pang, and Y. Yang, "Stair: Learning sparse text and image representation in grounded tokens," *arXiv preprint arXiv:2301.13081*, 2023.

[53] D. Chen, J. Liu, W. Dai, and B. Wang, "Visual instruction tuning with polite flamingo," *arXiv preprint arXiv:2307.01003*, 2023.

[54] F. Chen, D. Zhang, M. Han, X. Chen, J. Shi, S. Xu, and B. Xu, "Vlp: A survey on vision-language pre-training," *arXiv preprint arXiv:2202.09061*, 2022.

[55] K. Chen, Z. Zhang, W. Zeng, R. Zhang, F. Zhu, and R. Zhao, "Shikra: Unleashing multimodal llm's referential dialogue magic," *arXiv preprint arXiv:2306.15195*, 2023.

[56] L. Chen, M. Zhai, J. He, and G. Mori, "Object grounding via iterative context reasoning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

[57] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[58] L. Chen, J. Li, X. Dong, P. Zhang, C. He, J. Wang, F. Zhao, and D. Lin, "Sharegpt4v: Improving large multi-modal models with better captions," *arXiv preprint arXiv:2311.12793*, 2023.

[59] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, *et al.*, "Evaluating large language models trained on code," *arXiv preprint arXiv:2107.03374*, 2021.

[60] M. Chen, I. Laina, and A. Vedaldi, "Training-free layout control with cross-attention guidance," *arXiv preprint arXiv:2304.03373*, 2023.

[61] Q. Chen, X. Chen, G. Zeng, and J. Wang, "Group detr: Fast training convergence with decoupled one-to-many label assignment," *arXiv preprint arXiv:2207.13085*, 2022.

[62] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020.

[63] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," *NeurIPS*, 2020.

[64] T. Chen, S. Saxena, L. Li, D. J. Fleet, and G. Hinton, "Pix2seq: A language modeling framework for object detection," in *ICLR*, 2022.

[65] T. Chen, S. Saxena, L. Li, T.-Y. Lin, D. J. Fleet, and G. Hinton, "A unified sequence interface for vision tasks," *arXiv preprint arXiv:2206.07669*, 2022.

[66] W.-G. Chen, I. Spiridonova, J. Yang, J. Gao, and C. Li, "Llava-interactive: An all-in-one demo for image chat, segmentation, generation and editing," 2023.

[67] W. Chen, H. Hu, X. Chen, P. Verga, and W. W. Cohen, "Murag: Multimodal retrieval-augmented generator for open question answering over images and text," *arXiv preprint arXiv:2210.02928*, 2022.

[68] W. Chen, H. Hu, Y. Li, N. Rui, X. Jia, M.-W. Chang, and W. W. Cohen, "Subject-driven text-to-image generation via apprenticeship learning," *arXiv preprint arXiv:2304.00186*, 2023.

[69] W. Chen, H. Hu, C. Saharia, and W. W. Cohen, "Re-imagen: Retrieval-augmented text-to-image generator," *arXiv preprint arXiv:2209.14491*, 2022.

[70] X. Chen, J. Djolonga, P. Padlewski, B. Mustafa, S. Changpinyo, J. Wu, C. R. Ruiz, S. Goodman, X. Wang, Y. Tay, *et al.*, "Pali-x: On scaling up a multilingual vision and language model," *arXiv preprint arXiv:2305.18565*, 2023.

[71] X. Chen, X. Wang, S. Changpinyo, A. Piergiovanni, P. Padlewski, D. Salz, S. Goodman, A. Grycner, B. Mustafa, L. Beyer, *et al.*, "Pali: A jointly-scaled multilingual language-image model," *arXiv preprint arXiv:2209.06794*, 2022.

[72] X. Chen, Z. Zhao, F. Yu, Y. Zhang, and M. Duan, "Conditional diffusion for interactive segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7345–7354, 2021.

[73] X. Chen, Z. Zhao, Y. Zhang, M. Duan, D. Qi, and H. Zhao, "Focalclick: Towards practical interactive image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1300–1309, 2022.

[74] X. Chen, M. Ding, X. Wang, Y. Xin, S. Mo, Y. Wang, S. Han, P. Luo, G. Zeng, and J. Wang, "Context autoencoder for self-supervised representation learning," *arXiv preprint arXiv:2202.03026*, 2022.

[75] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[76] X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft COCO captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[77] X. Chen and K. He, "Exploring simple siamese representation learning," in *CVPR*, 2021.

[78] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *ICCV*, 2021.

[79] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "UNITER: Universal image-text representation learning," in *ECCV*, 2020.

[80] Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, and Y. Qiao, "Vision transformer adapter for dense predictions," *arXiv preprint arXiv:2205.08534*, 2022.

[81] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1290–1299, 2022.

[82] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev, "Reproducible scaling laws for contrastive language-image learning," in *CVPR*, 2023.

[83] J. Cho, J. Lei, H. Tan, and M. Bansal, "Unifying vision-and-language tasks via text generation," in *ICML*, 2021.

[84] J. Cho, L. Li, Z. Yang, Z. Gan, L. Wang, and M. Bansal, "Diagnostic benchmark and iterative inpainting for layout-guided image generation," *arXiv preprint arXiv:2304.06671*, 2023.

[85] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, *et al.*, "Palm: Scaling language modeling with pathways," *arXiv preprint arXiv:2204.02311*, 2022.

[86] T. Computer, *Redpajama-data: An open source recipe to reproduce llama training dataset*, 2023. URL: https://github.com/togethercomputer/RedPajama-Data.

[87] M. Crawshaw, "Multi-task learning with deep neural networks: A survey," *arXiv preprint arXiv:2009.09796*, 2020.

[88] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE signal processing magazine*, vol. 35, no. 1, 2018, pp. 53–65.

[89] H. Dai, C. Ma, Z. Liu, Y. Li, P. Shu, X. Wei, L. Zhao, Z. Wu, D. Zhu, W. Liu, *et al.*, "Samaug: Point prompt augmentation for segment anything model," *arXiv preprint arXiv:2307.01187*, 2023.

[90] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," *arXiv preprint arXiv:2305.06500*, 2023.

[91] X. Dai, Y. Chen, B. Xiao, D. Chen, M. Liu, L. Yuan, and L. Zhang, "Dynamic head: Unifying object detection heads with attentions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7373–7382, 2021.

[92] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys (Csur)*, vol. 40, no. 2, 2008, pp. 1–60.

[93] C. Deng, Q. Wu, Q. Wu, F. Hu, F. Lyu, and M. Tan, "Visual grounding via accumulated attention," in *CVPR*, 2018.

[94] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[95] K. Desai and J. Johnson, "Virtex: Learning visual representations from textual annotations," in *CVPR*, 2021.

[96] K. Desai, G. Kaul, Z. Aysola, and J. Johnson, "Redcaps: Web-curated image-text data created by the people, for the people," in *NeurIPS, Track on Datasets and Benchmarks*, 2021.

[97] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," *arXiv preprint arXiv:2305.14314*, 2023.

[98] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019.

[99] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *NeurIPS*, 2021.

[100] J. Ding, N. Xue, G.-S. Xia, and D. Dai, *Decoupling zero-shot semantic segmentation*, 2022. arXiv: 2112.07910 [cs.CV].

[101] Z. Ding, J. Wang, and Z. Tu, "Open-vocabulary panoptic segmentation with maskclip," *arXiv preprint arXiv:2208.08984*, 2022.

[102] B. Dong, F. Zeng, T. Wang, X. Zhang, and Y. Wei, "Solq: Segmenting objects by learning queries," *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 21 898–21 909.

[103] X. Dong, J. Bao, T. Zhang, D. Chen, W. Zhang, L. Yuan, D. Chen, F. Wen, and N. Yu, "Bootstrapped masked autoencoders for vision bert pretraining," in *ECCV*, 2022.

[104] X. Dong, J. Bao, T. Zhang, D. Chen, W. Zhang, L. Yuan, D. Chen, F. Wen, N. Yu, and B. Guo, "Peco: Perceptual codebook for bert pre-training of vision transformers," in *AAAI*, 2023.

[105] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[106] Z.-Y. Dou, A. Kamath, Z. Gan, P. Zhang, J. Wang, L. Li, Z. Liu, C. Liu, Y. LeCun, N. Peng, J. Gao, and L. Wang, "Coarse-to-fine vision-language pre-training with fusion in the backbone," in *NeurIPS*, 2022.

[107] Z.-Y. Dou, Y. Xu, Z. Gan, J. Wang, S. Wang, L. Wang, C. Zhu, Z. Liu, M. Zeng, *et al.*, "An empirical study of training end-to-end vision-and-language transformers," in *CVPR*, 2022.

[108] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, *et al.*, "PaLM-E: An embodied multimodal language model," *arXiv preprint arXiv:2303.03378*, 2023.

[109] Y. Du, Z. Liu, J. Li, and W. X. Zhao, "A survey of vision-language pre-trained models," in *IJCAI survey track*, 2022.

[110] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, and Y. Akbari, "Image inpainting: A review," *Neural Processing Letters*, vol. 51, 2020, pp. 2007–2028.

[111] A. Ermolov, A. Siarohin, E. Sangineto, and N. Sebe, "Whitening for self-supervised representation learning," in *ICML*, 2021.

[112] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *CVPR*, 2021.

[113] M. Everingham and J. Winn, "The pascal visual object classes challenge 2012 (voc2012) development kit," *Pattern Analysis, Statistical Modelling and Computational Learning, Tech. Rep*, vol. 8, no. 5, 2011.

[114] L. Fan, D. Krishnan, P. Isola, D. Katabi, and Y. Tian, "Improving clip training with language rewrites," *arXiv preprint arXiv:2305.20088*, 2023.

[115] Y. Fan, O. Watkins, Y. Du, H. Liu, M. Ryu, C. Boutilier, P. Abbeel, M. Ghavamzadeh, K. Lee, and K. Lee, "Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models," *arXiv preprint arXiv:2305.16381*, 2023.

[116] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, and Y. Cao, "Eva: Exploring the limits of masked visual representation learning at scale," in *CVPR*, 2023.
[117] Z. Fang, J. Wang, X. Hu, L. Liang, Z. Gan, L. Wang, Y. Yang, and Z. Liu, "Injecting semantic concepts into end-to-end image captioning," in *CVPR*, 2022.
[118] C. Feichtenhofer, Y. Li, K. He, *et al.*, "Masked autoencoders as spatiotemporal learners," *NeurIPS*, 2022.
[119] C. Feng, Y. Zhong, Z. Jie, X. Chu, H. Ren, X. Wei, W. Xie, and L. Ma, "Promptdet: Towards open-vocabulary detection using uncurated images," in *European Conference on Computer Vision*, Springer, pp. 701–717, 2022.
[120] W. Feng, X. He, T.-J. Fu, V. Jampani, A. R. Akula, P. Narayana, S. Basu, X. E. Wang, and W. Y. Wang, "Training-free structured diffusion guidance for compositional text-to-image synthesis," in *The Eleventh International Conference on Learning Representations*, 2022.
[121] W. Feng, W. Zhu, T.-j. Fu, V. Jampani, A. Akula, X. He, S. Basu, X. E. Wang, and W. Y. Wang, "Layoutgpt: Compositional visual planning and generation with large language models," *arXiv preprint arXiv:2305.15393*, 2023.
[122] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in *NeurIPS*, 2013.
[123] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, Z. Qiu, W. Lin, J. Yang, X. Zheng, *et al.*, "Mme: A comprehensive evaluation benchmark for multimodal large language models," *arXiv preprint arXiv:2306.13394*, 2023.
[124] T.-J. Fu, W. Hu, X. Du, W. Y. Wang, Y. Yang, and Z. Gan, "Guiding instruction-based image editing via multimodal large language models," Sep. 2023.
[125] S. Y. Gadre, G. Ilharco, A. Fang, J. Hayase, G. Smyrnis, T. Nguyen, R. Marten, M. Wortsman, D. Ghosh, J. Zhang, *et al.*, "Datacomp: In search of the next generation of multimodal datasets," *arXiv preprint arXiv:2304.14108*, 2023.

[126] O. Gafni, A. Polyak, O. Ashual, S. Sheynin, D. Parikh, and Y. Taigman, “Make-a-scene: Scene-based text-to-image generation with human priors,” *arXiv preprint arXiv:2203.13131*, 2022.
[127] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, “An image is worth one word: Personalizing text-to-image generation using textual inversion,” *arXiv preprint arXiv:2208.01618*, 2022.
[128] Z. Gan, Y.-C. Chen, L. Li, C. Zhu, Y. Cheng, and J. Liu, “Large-scale adversarial training for vision-and-language representation learning,” in *NeurIPS*, 2020.
[129] Z. Gan, L. Li, C. Li, L. Wang, Z. Liu, J. Gao, *et al.*, “Vision-language pre-training: Basics, recent advances, and future trends,” *Foundations and Trends® in Computer Graphics and Vision*, 2022.
[130] D. Gao, L. Ji, L. Zhou, K. Q. Lin, J. Chen, Z. Fan, and M. Z. Shou, “Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn,” *arXiv preprint arXiv:2306.08640*, 2023.
[131] P. Gao, J. Han, R. Zhang, Z. Lin, S. Geng, A. Zhou, W. Zhang, P. Lu, C. He, X. Yue, *et al.*, “Llama-adapter v2: Parameter-efficient visual instruction model,” *arXiv preprint arXiv:2304.15010*, 2023.
[132] P. Gao, T. Ma, H. Li, Z. Lin, J. Dai, and Y. Qiao, “Convmae: Masked convolution meets masked autoencoders,” *arXiv preprint arXiv:2205.03892*, 2022.
[133] Y. Ge, Y. Ge, Z. Zeng, X. Wang, and Y. Shan, “Planting a seed of vision in large language model,” *arXiv preprint arXiv:2307.08041*, 2023.
[134] *Gen-2*, https://research.runwayml.com/gen2.
[135] X. Geng and H. Liu, *Openllama: An open reproduction of llama*, May 2023. URL: https://github.com/openlm-research/open_llama.
[136] Z. Geng, B. Yang, T. Hang, C. Li, S. Gu, T. Zhang, J. Bao, Z. Zhang, H. Hu, D. Chen, *et al.*, “Instructdiffusion: A generalist modeling interface for vision tasks,” *arXiv preprint arXiv:2309.03895*, 2023.

[137] G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin, "Open-vocabulary image segmentation," in *ECCV*, 2022.

[138] G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin, "Scaling open-vocabulary image segmentation with image-level labels," in *European Conference on Computer Vision*, Springer, pp. 540–557, 2022.

[139] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind: One embedding space to bind them all," in *CVPR*, 2023.

[140] R. Girshick, "Fast r-cnn," in *ICCV*, 2015.

[141] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, 2015, pp. 142–158.

[142] T. Gong, C. Lyu, S. Zhang, Y. Wang, M. Zheng, Q. Zhao, K. Liu, W. Zhang, P. Luo, and K. Chen, "Multimodal-gpt: A vision and language model for dialogue with humans," *arXiv preprint arXiv:2305.04790*, 2023.

[143] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, 2020.

[144] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *NeurIPS*, 2020.

[145] X. Gu, Y. Cui, J. Huang, A. Rashwan, X. Yang, X. Zhou, G. Ghiasi, W. Kuo, H. Chen, L.-C. Chen, *et al.*, "Dataseg: Taming a universal multi-dataset multi-task segmentation model," *arXiv preprint arXiv:2306.01736*, 2023.

[146] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," *arXiv preprint arXiv:2104.13921*, 2021.

[147] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," in *ICLR*, 2022.

[148] A. Gudibande, E. Wallace, C. Snell, X. Geng, H. Liu, P. Abbeel, S. Levine, and D. Song, "The false promise of imitating proprietary llms," *arXiv preprint arXiv:2305.15717*, 2023.

[149] L. Gui, B. Wang, Q. Huang, A. Hauptmann, Y. Bisk, and J. Gao, "Kat: A knowledge augmented transformer for vision-and-language," in *NAACL*, 2022.

[150] A. Gunjal, J. Yin, and E. Bas, "Detecting and preventing hallucinations in large vision language models," *arXiv preprint arXiv:2308.06394*, 2023.

[151] T. Gupta, A. Kamath, A. Kembhavi, and D. Hoiem, "Towards general purpose vision systems: An end-to-end task-agnostic vision-language architecture," in *CVPR*, 2022.

[152] T. Gupta, A. Kamath, A. Kembhavi, and D. Hoiem, "Towards general purpose vision systems: An end-to-end task-agnostic vision-language architecture," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16 399–16 409, Jun. 2022.

[153] T. Gupta and A. Kembhavi, "Visual programming: Compositional visual reasoning without training," *arXiv preprint arXiv:2211.11559*, 2022.

[154] T. Gupta and A. Kembhavi, "Visual programming: Compositional visual reasoning without training," *ArXiv*, vol. abs/2211.11559, 2022.

[155] T. Gupta and A. Kembhavi, "Visual programming: Compositional visual reasoning without training," in *CVPR*, 2023.

[156] T. Gupta, R. Marten, A. Kembhavi, and D. Hoiem, "Grit: General robust image task benchmark," *arXiv preprint arXiv:2204.13653*, 2022.

[157] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham, "Vizwiz grand challenge: Answering visual questions from blind people," in *CVPR*, 2018.

[158] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *AISTATS*, 2010.

[159] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, "Realm: Retrieval-augmented language model pre-training," *arXiv preprint arXiv:2002.08909*, 2020.
[160] A. M. Hafiz and G. M. Bhat, "A survey on instance segmentation: State of the art," *International journal of multimedia information retrieval*, 2020.
[161] A. W. Harley, Z. Fang, and K. Fragkiadaki, "Particle video revisited: Tracking through occlusions using point trajectories," in *European Conference on Computer Vision*, Springer, pp. 59–75, 2022.
[162] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *CVPR*, 2022.
[163] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.
[164] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *ICCV*, 2017.
[165] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 12, 2010, pp. 2341–2353.
[166] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
[167] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced bert with disentangled attention," in *ICLR*, 2021.
[168] R. He, S. Sun, X. Yu, C. Xue, W. Zhang, P. Torr, S. Bai, and X. Qi, "Is synthetic data from generative models ready for image recognition?" *arXiv preprint arXiv:2210.07574*, 2022.
[169] O. Henaff, "Data-efficient image recognition with contrastive predictive coding," in *ICML*, 2020.
[170] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-or, "Prompt-to-prompt image editing with cross-attention control," in *The Eleventh International Conference on Learning Representations*, 2022.

[171] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," *arXiv preprint arXiv:1808.06670*, 2018.

[172] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, *et al.*, "Imagen video: High definition video generation with diffusion models," *arXiv preprint arXiv:2210.02303*, 2022.

[173] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *NeurIPS*, 2020.

[174] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, *et al.*, "Training compute-optimal large language models," *arXiv preprint arXiv:2203.15556*, 2022.

[175] Y. Hong, H. Zhen, P. Chen, S. Zheng, Y. Du, Z. Chen, and C. Gan, "3d-llm: Injecting the 3d world into large language models," *arXiv preprint arXiv:2307.12981*, 2023.

[176] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[177] R. Hu, M. Rohrbach, and T. Darrell, "Segmentation from natural language expressions," in *European Conference on Computer Vision*, Springer, pp. 108–124, 2016.

[178] R. Hu and A. Singh, "Unit: Multimodal multitask learning with a unified transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1439–1449, Oct. 2021.

[179] R. Hu and A. Singh, "Unit: Multimodal multitask learning with a unified transformer," in *ICCV*, 2021.

[180] W. Hu, Y. Xu, Y. Li, W. Li, Z. Chen, and Z. Tu, "Bliva: A simple multimodal llm for better handling of text-rich visual questions," *arXiv preprint arXiv:2308.09936*, 2023.

[181] L. Huang, S. You, M. Zheng, F. Wang, C. Qian, and T. Yamasaki, "Green hierarchical vision transformer for masked image modeling," *NeurIPS*, 2022.

[182] R. Huang, M. Li, D. Yang, J. Shi, X. Chang, Z. Ye, Y. Wu, Z. Hong, J. Huang, J. Liu, *et al.*, "Audiogpt: Understanding and generating speech, music, sound, and talking head," *arXiv preprint arXiv:2304.12995*, 2023.

[183] S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. K. Mohammed, Q. Liu, *et al.*, "Language is not all you need: Aligning perception with language models," *arXiv preprint arXiv:2302.14045*, 2023.

[184] S. Huang, Z. Jiang, H. Dong, Y. Qiao, P. Gao, and H. Li, "Instruct2act: Mapping multi-modality instructions to robotic actions with large language model," *arXiv preprint arXiv:2305.11176*, 2023.

[185] W. Huang, P. Abbeel, D. Pathak, and I. Mordatch, "Language models as zero-shot planners: Extracting actionable knowledge for embodied agents," in *International Conference on Machine Learning*, PMLR, pp. 9118–9147, 2022.

[186] Y. Huang, Z. Meng, F. Liu, Y. Su, N. Collier, and Y. Lu, "Sparkles: Unlocking chats across multiple images for multimodal instruction-following models," *arXiv preprint arXiv:2308.16463*, 2023.

[187] Z. Huang, Z. Zeng, Y. Huang, B. Liu, D. Fu, and J. Fu, "Seeing out of the box: End-to-end pre-training for vision-language representation learning," in *CVPR*, 2021.

[188] Z. Huang, Z. Zeng, B. Liu, D. Fu, and J. Fu, "Pixel-BERT: Aligning image pixels with text by deep multi-modal transformers," *arXiv preprint arXiv:2004.00849*, 2020.

[189] D. Huynh, J. Kuen, Z. Lin, J. Gu, and E. Elhamifar, "Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7020–7031, 2022.

[190] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt, *Openclip*, version 0.1, Jul. 2021. DOI: 10.5281/zenodo.5143773.

[191] A. Jain, B. Mildenhall, J. T. Barron, P. Abbeel, and B. Poole, "Zero-shot text-guided object generation with dream fields," 2022.

[192] J. Jain, J. Li, M. T. Chiu, A. Hassani, N. Orlov, and H. Shi, "One-former: One transformer to rule universal image segmentation," in *CVPR*, 2023.

[193] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, 2020.

[194] K. R. Jerripothula, J. Cai, and J. Yuan, "Image co-segmentation via saliency co-fusion," *IEEE Transactions on Multimedia*, vol. 18, no. 9, 2016, pp. 1896–1909.

[195] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *ICML*, 2021.

[196] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[197] A. Joulin, F. Bach, and J. Ponce, "Discriminative clustering for image co-segmentation," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, pp. 1943–1950, 2010.

[198] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion, "Mdetr-modulated detection for end-to-end multi-modal understanding," in *ICCV*, 2021.

[199] M. Kang, J.-Y. Zhu, R. Zhang, J. Park, E. Shechtman, S. Paris, and T. Park, "Scaling up gans for text-to-image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10 124–10 134, 2023.

[200] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.

[201] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani, "Imagic: Text-based real image editing with diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6007–6017, 2023.

[202] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, "Referitgame: Referring to objects in photographs of natural scenes," in *EMNLP*, 2014.

[203] W. Kim, B. Son, and I. Kim, "ViLT: Vision-and-language transformer without convolution or region supervision," in *ICML*, 2021.

[204] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[205] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *CVPR*, 2019.

[206] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[207] J. Y. Koh, D. Fried, and R. Salakhutdinov, "Generating images with multimodal language models," *arXiv preprint arXiv:2305.17216*, 2023.

[208] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *arXiv preprint arXiv:2205.11916*, 2022.

[209] I. Kokkinos, "Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6129–6138, 2017.

[210] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, "Big transfer (bit): General visual representation learning," in *ECCV*, 2020.

[211] A. Kolesnikov, A. S. Pinto, L. Beyer, X. Zhai, J. Harmsen, and N. Houlsby, "Uvim: A unified modeling approach for vision with learned guiding codes," *arXiv preprint arXiv:2205.10337*, 2022.

[212] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NeurIPS*, 2012.

[213] N. Kumari, B. Zhang, R. Zhang, E. Shechtman, and J.-Y. Zhu, "Multi-concept customization of text-to-image diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1931–1941, 2023.

[214] W. Kuo, F. Bertsch, W. Li, A. Piergiovanni, M. Saffar, and A. Angelova, "Findit: Generalized localization with natural language queries," in *ECCV*, 2022.

[215] X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, and J. Jia, "Lisa: Reasoning segmentation via large language model," *arXiv preprint arXiv:2308.00692*, 2023.

[216] A. Lamb, V. Dumoulin, and A. Courville, "Discriminative regularization for generative models," *arXiv preprint arXiv:1602.03220*, 2016.

[217] J. Lambert, Z. Liu, O. Sener, J. Hays, and V. Koltun, "Mseg: A composite dataset for multi-domain semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2879–2888, 2020.

[218] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *International conference on machine learning*, PMLR, pp. 1558–1566, 2016.

[219] H. Laurençon, L. Saulnier, L. Tronchon, S. Bekman, A. Singh, A. Lozhkov, T. Wang, S. Karamcheti, A. M. Rush, D. Kiela, *et al.*, "Obelisc: An open web-scale filtered dataset of interleaved image-text documents," *arXiv preprint arXiv:2306.16527*, 2023.

[220] D. Lee, C. Kim, S. Kim, M. Cho, and W.-S. Han, "Autoregressive image generation using residual quantization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11 523–11 532, 2022.

[221] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *NeurIPS*, 2020.

[222] B. Li, H. Liu, L. Chen, Y. J. Lee, C. Li, and Z. Liu, "Benchmarking and analyzing generative data for visual recognition," *arXiv preprint arXiv:2307.13697*, 2023.

[223] B. Li, Y. Zhang, L. Chen, J. Wang, F. Pu, J. Yang, C. Li, and Z. Liu, "Mimic-it: Multi-modal in-context instruction tuning," *arXiv preprint arXiv:2306.05425*, 2023.

[224] B. Li, Y. Zhang, L. Chen, J. Wang, J. Yang, and Z. Liu, "Otter: A multi-modal model with in-context instruction tuning," *arXiv preprint arXiv:2305.03726*, 2023.

[225] B. Li, R. Wang, G. Wang, Y. Ge, Y. Ge, and Y. Shan, "Seed-bench: Benchmarking multimodal llms with generative comprehension," *arXiv preprint arXiv:2307.16125*, 2023.

[226] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, "Language-driven semantic segmentation," in *ICLR*, 2022.

[227] C. Li, H. Liu, L. H. Li, P. Zhang, J. Aneja, J. Yang, P. Jin, Y. J. Lee, H. Hu, Z. Liu, *et al.*, "Elevater: A benchmark and toolkit for evaluating language-augmented visual models," in *NeurIPS, Track on Datasets and Benchmarks*, 2022.

[228] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, "Llava-med: Training a large language-and-vision assistant for biomedicine in one day," *arXiv preprint arXiv:2306.00890*, 2023.

[229] C. Li, J. Yang, P. Zhang, M. Gao, B. Xiao, X. Dai, L. Yuan, and J. Gao, "Efficient self-supervised vision transformers for representation learning," *arXiv preprint arXiv:2106.09785*, 2021.

[230] F. Li, H. Zhang, P. Sun, X. Zou, S. Liu, J. Yang, C. Li, L. Zhang, and J. Gao, "Semantic-sam: Segment and recognize anything at any granularity," *arXiv preprint arXiv:2307.04767*, 2023.

[231] F. Li, H. Zhang, Y.-F. Zhang, S. Liu, J. Guo, L. M. Ni, P. Zhang, and L. Zhang, "Vision-language intelligence: Tasks, representation learning, and large models," *arXiv preprint arXiv:2203.01922*, 2022.

[232] G. Li, N. Duan, Y. Fang, M. Gong, and D. Jiang, "Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training," in *AAAI*, 2020.

[233] H. Li, J. Zhu, X. Jiang, X. Zhu, H. Li, C. Yuan, X. Wang, Y. Qiao, X. Wang, W. Wang, *et al.*, "Uni-perceiver v2: A generalist model for large-scale vision and vision-language tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2691–2700, 2023.

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

[235] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICML*, 2022.

[236] J. Li, R. R. Selvaraju, A. D. Gotmare, S. Joty, C. Xiong, and S. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *NeurIPS*, 2021.

[237] K. Li, Y. He, Y. Wang, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, and Y. Qiao, "Videochat: Chat-centric video understanding," *arXiv preprint arXiv:2305.06355*, 2023.

[238] L. Li, Y. Yin, S. Li, L. Chen, P. Wang, S. Ren, M. Li, Y. Yang, J. Xu, X. Sun, *et al.*, "M3it: A large-scale dataset towards multi-modal multilingual instruction tuning," *arXiv preprint arXiv:2306.04387*, 2023.

[239] L. Li, Z. Gan, Y. Cheng, and J. Liu, "Relation-aware graph attention network for visual question answering," in *ICCV*, 2019.

[240] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019.

[241] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, K.-W. Chang, and J. Gao, "Grounded language-image pre-training," *CVPR*, 2022.

[242] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, K.-W. Chang, and J. Gao, "Grounded language-image pre-training," in *CVPR*, 2022.

[243] M. Li, F. Song, B. Yu, H. Yu, Z. Li, F. Huang, and Y. Li, "Api-bank: A benchmark for tool-augmented llms," *arXiv preprint arXiv:2304.08244*, 2023.

[244] S. Li and N. Tajbakhsh, "Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs," *arXiv preprint arXiv:2308.03349*, 2023.
[245] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," *arXiv preprint arXiv:2101.00190*, 2021.
[246] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, and J. Gao, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *ECCV*, 2020.
[247] Y. Li, C. Zhang, G. Yu, Z. Wang, B. Fu, G. Lin, C. Shen, L. Chen, and Y. Wei, "Stablellava: Enhanced visual instruction tuning with synthesized image-dialogue data," *arXiv preprint arXiv:2308.10253*, 2023.
[248] Y. Li, F. Liang, L. Zhao, Y. Cui, W. Ouyang, J. Shao, F. Yu, and J. Yan, "Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm," in *ICLR*, 2022.
[249] Y. Li, H. Fan, R. Hu, C. Feichtenhofer, and K. He, "Scaling language-image pre-training via masking," in *CVPR*, 2023.
[250] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J.-R. Wen, "Evaluating object hallucination in large vision-language models," *arXiv preprint arXiv:2305.10355*, 2023.
[251] Y. Li, S. Bubeck, R. Eldan, A. D. Giorno, S. Gunasekar, and Y. T. Lee, *Textbooks are all you need ii: Phi-1.5 technical report*, 2023. arXiv: 2309.05463 [cs.CL].
[252] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. J. Lee, "Gligen: Open-set grounded text-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22 511–22 521, 2023.
[253] F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, P. Zhang, P. Vajda, and D. Marculescu, "Open-vocabulary semantic segmentation with mask-adapted clip," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7061–7070, 2023.

[254] Y. Liang, C. Wu, T. Song, W. Wu, Y. Xia, Y. Liu, Y. Ou, S. Lu, L. Ji, S. Mao, *et al.*, "Taskmatrix. ai: Completing tasks by connecting foundation models with millions of apis," *arXiv preprint arXiv:2303.16434*, 2023.

[255] B. Lin, B. Zhu, Y. Ye, M. Ning, P. Jin, and L. Yuan, "Video-llava: Learning united visual representation by alignment before projection," *arXiv preprint arXiv:2311.10122*, 2023.

[256] C.-H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M.-Y. Liu, and T.-Y. Lin, "Magic3d: High-resolution text-to-3d content creation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[257] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

[258] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.

[259] Y. Lin, H. Wu, R. Wang, H. Lu, X. Lin, H. Xiong, and L. Wang, "Towards language-guided interactive 3d generation: Llms as layout interpreter with generative feedback," *arXiv preprint arXiv:2305.15808*, 2023.

[260] C. Liu, Z. Lin, X. Shen, J. Yang, X. Lu, and A. Yuille, "Recurrent multimodal interaction for referring image segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1271–1280, 2017.

[261] F. Liu, K. Lin, L. Li, J. Wang, Y. Yacoob, and L. Wang, "Aligning large multi-modal model with robust instruction tuning," *arXiv preprint arXiv:2306.14565*, 2023.

[262] H. Liu, X. Jiang, X. Li, A. Guo, Y. Hu, D. Jiang, and B. Ren, "The devil is in the frequency: Geminated gestalt autoencoder for self-supervised visual pre-training," in *AAAI*, 2023.

[263] H. Liu, C. Li, Y. Li, and Y. J. Lee, *Improved baselines with visual instruction tuning*, 2023.

[264] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *arXiv preprint arXiv:2304.08485*, 2023.

[265] H. Liu, K. Son, J. Yang, C. Liu, J. Gao, Y. J. Lee, and C. Li, "Learning customized visual models with retrieval-augmented knowledge," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[266] J. Liu, H. Ding, Z. Cai, Y. Zhang, R. K. Satzoda, V. Mahadevan, and R. Manmatha, *Polyformer: Referring image segmentation as sequential polygon generation*, 2023. arXiv: 2302.07387 [cs.CV].

[267] N. Liu, S. Li, Y. Du, A. Torralba, and J. B. Tenenbaum, "Compositional visual generation with composable diffusion models," in *European Conference on Computer Vision*, Springer, pp. 423–439, 2022.

[268] R. Liu, R. Wu, B. V. Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick, *Zero-1-to-3: Zero-shot one image to 3d object*, 2023. arXiv: 2303.11328 [cs.CV].

[269] 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 [cs.LG].

[270] S. Liu, H. Cheng, H. Liu, H. Zhang, F. Li, T. Ren, X. Zou, J. Yang, H. Su, J. Zhu, L. Zhang, J. Gao, and C. Li, *Llava-plus: Learning to use tools for creating multimodal agents*, 2023.

[271] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.

[272] S. Liu, J. Ye, and X. Wang, "Any-to-any style transfer," *arXiv preprint arXiv:2304.09728*, 2023.

[273] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, Springer, pp. 21–37, 2016.

[274] W. Liu and Y. Zuo, "Stone needle: A general multimodal large-scale model framework towards healthcare," *arXiv preprint arXiv:2306.16034*, 2023.

[275] X. Liu, C. Gong, and Q. Liu, *Flow straight and fast: Learning to generate and transfer data with rectified flow*, 2022. arXiv: 2209.03003 [cs.LG].

[276] X. Liu, Z. Zhu, H. Liu, Y. Yuan, M. Cui, Q. Huang, J. Liang, Y. Cao, Q. Kong, M. D. Plumbley, *et al.*, "Wavjourney: Compositional audio creation with large language models," *arXiv preprint arXiv:2307.14335*, 2023.

[277] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[278] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, *et al.*, "Mmbench: Is your multi-modal model an all-around player?" *arXiv preprint arXiv:2307.06281*, 2023.

[279] Y. Liu, Z. Li, H. Li, W. Yu, M. Huang, D. Peng, M. Liu, M. Chen, C. Li, L. Jin, *et al.*, "On the hidden mystery of ocr in large multimodal models," *arXiv preprint arXiv:2305.07895*, 2023.

[280] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021.

[281] Z. Liu, Y. He, W. Wang, W. Wang, Y. Wang, S. Chen, Q. Zhang, Y. Yang, Q. Li, J. Yu, *et al.*, "Internchat: Solving vision-centric tasks by interacting with chatbots beyond language," *arXiv preprint arXiv:2305.05662*, 2023.

[282] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11 976–11 986, 2022.

[283] A. Long, W. Yin, T. Ajanthan, V. Nguyen, P. Purkait, R. Garg, A. Blair, C. Shen, and A. van den Hengel, "Retrieval augmented classification for long-tail visual recognition," in *CVPR*, 2022.

[284] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.

[285] C.-Z. Lu, X. Jin, Q. Hou, J. H. Liew, M.-M. Cheng, and J. Feng, "Delving deeper into data scaling in masked image modeling," *arXiv preprint arXiv:2305.15248*, 2023.

[286] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *NeurIPS*, 2019.

[287] J. Lu, C. Clark, R. Zellers, R. Mottaghi, and A. Kembhavi, "Unified-io: A unified model for vision, language, and multi-modal tasks," *arXiv preprint arXiv:2206.08916*, 2022.

[288] J. Lu, V. Goswami, M. Rohrbach, D. Parikh, and S. Lee, "12-in-1: Multi-task vision and language representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10 437–10 446, 2020.

[289] P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K.-W. Chang, M. Galley, and J. Gao, "Mathvista: Evaluating mathematical reason-ing of foundation models in visual contexts," *arXiv preprint arXiv:2310.02255*, 2023.

[290] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan, "Learn to explain: Multimodal reasoning via thought chains for science question answering," *Advances in Neural Information Processing Systems*, 2022.

[291] P. Lu, B. Peng, H. Cheng, M. Galley, K.-W. Chang, Y. N. Wu, S.-C. Zhu, and J. Gao, "Chameleon: Plug-and-play compositional reasoning with large language models," *arXiv preprint arXiv:2304.09842*, 2023.

[292] Q. Lu, J. Kuen, S. Tiancheng, G. Jiuxiang, G. Weidong, J. Jiaya, L. Zhe, and Y. Ming-Hsuan, "High-quality entity segmentation," in *ICCV*, 2023.

[293] Y. Lu, C. Li, H. Liu, J. Yang, J. Gao, and Y. Shen, "An empirical study of scaling instruction-tuned large multimodal models," *arXiv preprint*, 2023.

[294] T. Lüddecke and A. Ecker, "Image segmentation using text and image prompts," in *CVPR*, 2022.

[295] G. Luo, Y. Zhou, T. Ren, S. Chen, X. Sun, and R. Ji, "Cheap and quick: Efficient vision-language instruction tuning for large language models," *arXiv preprint arXiv:2305.15023*, 2023.

[296] H. Luo, J. Bao, Y. Wu, X. He, and T. Li, "Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation," in *International Conference on Machine Learning*, PMLR, pp. 23 033–23 044, 2023.

[297] R. Luo, Z. Zhao, M. Yang, J. Dong, M. Qiu, P. Lu, T. Wang, and Z. Wei, "Valley: Video assistant with large language model enhanced ability," *arXiv preprint arXiv:2306.07207*, 2023.

[298] W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu, and T.-K. Kim, "Multiple object tracking: A literature review," *Artificial intelligence*, vol. 293, 2021, p. 103 448.

[299] J. Ma and B. Wang, "Segment anything in medical images," *arXiv preprint arXiv:2304.12306*, 2023.

[300] Z. Ma, X. Hong, and Q. Shangguan, "Can sam count anything? an empirical study on sam counting," *arXiv preprint arXiv:2304.10817*, 2023.

[301] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," in *CVPR*, 2016.

[302] E. Margffoy-Tuay, J. C. Pérez, E. Botero, and P. Arbeláez, "Dynamic multimodal instance segmentation guided by natural language queries," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 630–645, 2018.

[303] K. Marino, X. Chen, D. Parikh, A. Gupta, and M. Rohrbach, "Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa," in *CVPR*, 2021.

[304] M. Mazumder, C. Banbury, X. Yao, B. Karlaš, W. G. Rojas, S. Diamos, G. Diamos, L. He, D. Kiela, D. Jurado, *et al.*, "Dataperf: Benchmarks for data-centric ai development," *arXiv preprint arXiv:2207.10062*, 2022.

[305] K. McGuinness and N. E. O'connor, "A comparative evaluation of interactive segmentation algorithms," *Pattern Recognition*, vol. 43, no. 2, 2010, pp. 434–444.

[306] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, "Sdedit: Guided image synthesis and editing with stochastic differential equations," in *International Conference on Learning Representations*, 2021.

[307] A. Mertan, D. J. Duff, and G. Unal, "Single image depth estimation: An overview," *Digital Signal Processing*, vol. 123, 2022, p. 103 441.

[308] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," in *ICCV*, 2019.

[309] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[310] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen, X. Wang, X. Zhai, T. Kipf, and N. Houlsby, *Simple open-vocabulary object detection with vision transformers*, 2022. arXiv: 2205.06230 [cs.CV].

[311] I. Misra and L. v. d. Maaten, "Self-supervised learning of pretext-invariant representations," in *CVPR*, 2020.

[312] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert, "Cross-stitch networks for multi-task learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3994–4003, 2016.

[313] M. Monajatipoor, L. H. Li, M. Rouhsedaghat, L. F. Yang, and K.-W. Chang, "Metavl: Transferring in-context learning ability from language models to vision-language models," *arXiv preprint arXiv:2306.01311*, 2023.

[314] *Moonvalley*, 2023. URL: https://moonvalley.ai/.

[315] M. Moor, Q. Huang, S. Wu, M. Yasunaga, C. Zakka, Y. Dalmia, E. P. Reis, P. Rajpurkar, and J. Leskovec, "Med-flamingo: A multimodal medical few-shot learner," *arXiv preprint arXiv:2307.15189*, 2023.

[316] *Morph*, 2023. URL: https://www.morphstudio.com/.

[317] E. N. Mortensen and W. A. Barrett, "Interactive segmentation with intelligent scissors," *Graphical models and image processing*, vol. 60, no. 5, 1998, pp. 349–384.

[318] MosaicML NLP Team, *Introducing mpt-7b: A new standard for open-source, ly usable llms*, 2023. URL: www.mosaicml.com/blog/mpt-7b.

[319] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The role of context for object detection and semantic segmentation in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 891–898, 2014.

[320] C. Mou, X. Wang, L. Xie, J. Zhang, Z. Qi, Y. Shan, and X. Qie, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," *arXiv preprint arXiv:2302.08453*, 2023.

[321] N. Mu, A. Kirillov, D. Wagner, and S. Xie, "Slip: Self-supervision meets language-image pre-training," *arXiv preprint arXiv:2112.12750*, 2021.

[322] Y. Mu, Q. Zhang, M. Hu, W. Wang, M. Ding, J. Jin, B. Wang, J. Dai, Y. Qiao, and P. Luo, "Embodiedgpt: Vision-language pre-training via embodied chain of thought," *arXiv preprint arXiv:2305.15021*, 2023.

[323] S. Munasinghe, R. Thushara, M. Maaz, H. A. Rasheed, S. Khan, M. Shah, and F. Khan, *Pg-video-llava: Pixel grounding large video-language models*, 2023. arXiv: 2311.13435 [cs.CV].

[324] K. Musgrave, S. Belongie, and S.-N. Lim, "A metric learning reality check," in *ECCV*, 2020.

[325] V. K. Nagaraja, V. I. Morariu, and L. S. Davis, "Modeling context between objects for referring expression understanding," in *ECCV*, 2016.

[326] R. Nakano, J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju, W. Saunders, *et al.*, "Webgpt: Browser-assisted question-answering with human feedback," *arXiv preprint arXiv:2112.09332*, 2021.

[327] T. Nguyen, G. Ilharco, M. Wortsman, S. Oh, and L. Schmidt, "Quality not quantity: On the interaction between dataset design and robustness of clip," *NeurIPS*, 2022.

[328] J. Ning, C. Li, Z. Zhang, Z. Geng, Q. Dai, K. He, and H. Hu, "All in tokens: Unifying output space of visual tasks via soft token," *arXiv preprint arXiv:2301.02229*, 2023.

[329] A. El-Nouby, G. Izacard, H. Touvron, I. Laptev, H. Jegou, and E. Grave, "Are large-scale datasets necessary for self-supervised pre-training?" *arXiv preprint arXiv:2112.10740*, 2021.

[330] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *NeurIPS*, 2017.

[331] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[332] A. v. d. Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," *arXiv preprint arXiv:1711.00937*, 2017.

[333] OpenAI, *ChatGPT*, 2022. URL: https://openai.com/blog/chatgpt/.

[334] OpenAI, *GPT-4 technical report*, 2023. URL: https://arxiv.org/abs/2303.08774.

[335] OpenAI, *Gpt-4 technical report*, 2023. arXiv: 2303.08774 [cs.CL].

[336] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[337] V. Ordonez, G. Kulkarni, and T. Berg, "Im2text: Describing images using 1 million captioned photographs," in *NeurIPS*, 2011.

[338] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, *et al.*, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 27 730–27 744.

[339] U. Ozbulak, H. J. Lee, B. Boga, E. T. Anzaku, H. Park, A. Van Messem, W. De Neve, and J. Vankerschaver, "Know your self-supervised learning: A survey on image-based generative and discriminative training," *arXiv preprint arXiv:2305.13689*, 2023.

[340] B. Paranjape, S. Lundberg, S. Singh, H. Hajishirzi, L. Zettlemoyer, and M. T. Ribeiro, "Art: Automatic multi-step reasoning and tool-use for large language models," *arXiv preprint arXiv:2303.09014*, 2023.

[341] S. G. Patil, T. Zhang, X. Wang, and J. E. Gonzalez, "Gorilla: Large language model connected with massive apis," *arXiv preprint arXiv:2305.15334*, 2023.

[342] W. Peebles and S. Xie, "Scalable diffusion models with transformers," *arXiv preprint arXiv:2212.09748*, 2022.

[343] G. Penedo, Q. Malartic, D. Hesslow, R. Cojocaru, A. Cappelli, H. Alobeidli, B. Pannier, E. Almazrouei, and J. Launay, "The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data, and web data only," *arXiv preprint arXiv:2306.01116*, 2023. URL: https://arxiv.org/abs/2306.01116.

[344] B. Peng, M. Galley, P. He, H. Cheng, Y. Xie, Y. Hu, Q. Huang, L. Liden, Z. Yu, W. Chen, and J. Gao, "Check your facts and try again: Improving large language models with external knowledge and automated feedback," *arXiv preprint arXiv:2302.12813*, 2023.

[345] B. Peng, C. Li, P. He, M. Galley, and J. Gao, "Instruction tuning with GPT-4," *arXiv preprint arXiv:2304.03277*, 2023.

[346] Z. Peng, L. Dong, H. Bao, Q. Ye, and F. Wei, "A unified view of masked image modeling," *arXiv preprint arXiv:2210.10615*, 2022.

[347] Z. Peng, L. Dong, H. Bao, Q. Ye, and F. Wei, "Beit v2: Masked image modeling with vector-quantized visual tokenizers," *arXiv preprint arXiv:2208.06366*, 2022.

[348] Z. Peng, W. Wang, L. Dong, Y. Hao, S. Huang, S. Ma, and F. Wei, "Kosmos-2: Grounding multimodal large language models to the world," *arXiv preprint arXiv:2306.14824*, 2023.

[349] M. E. Peters, M. Neumann, R. L. Logan IV, R. Schwartz, V. Joshi, S. Singh, and N. A. Smith, "Knowledge enhanced contextual word representations," *arXiv preprint arXiv:1909.04164*, 2019.

[350] H. Pham, Z. Dai, G. Ghiasi, H. Liu, A. W. Yu, M.-T. Luong, M. Tan, and Q. V. Le, "Combined scaling for zero-shot transfer learning," *arXiv preprint arXiv:2111.10050*, 2021.

[351] R. Pi, J. Gao, S. Diao, R. Pan, H. Dong, J. Zhang, L. Yao, J. Han, H. Xu, and L. K. T. Zhang, "Detgpt: Detect what you need via reasoning," *arXiv preprint arXiv:2305.14167*, 2023.

[352] *Pika 1.0*, 2023. URL: https://pika.art/.

[353] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *ICCV*, 2015.

[354] J. Pont-Tuset, J. Uijlings, S. Changpinyo, R. Soricut, and V. Ferrari, "Connecting vision and language with localized narratives," in *ECCV*, 2020.

[355] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," *arXiv*, 2022.

[356] C. Qian, C. Han, Y. R. Fung, Y. Qin, Z. Liu, and H. Ji, "Creator: Disentangling abstract and concrete reasonings of large language models through tool creation," *arXiv preprint arXiv:2305.14318*, 2023.

[357] R. Qian, Y. Li, Z. Xu, M.-H. Yang, S. Belongie, and Y. Cui, "Multimodal open-vocabulary video classification via pre-trained vision and language models," *arXiv preprint arXiv:2207.07646*, 2022.

[358] C. Qin, S. Zhang, N. Yu, Y. Feng, X. Yang, Y. Zhou, H. Wang, J. C. Niebles, C. Xiong, S. Savarese, *et al.*, "Unicontrol: A unified diffusion model for controllable visual generation in the wild," *arXiv preprint arXiv:2305.11147*, 2023.

[359] J. Qin, J. Wu, P. Yan, M. Li, R. Yuxi, X. Xiao, Y. Wang, R. Wang, S. Wen, X. Pan, and X. Wang, *Freeseg: Unified, universal and open-vocabulary image segmentation*, 2023. arXiv: 2303.17225 [cs.CV].

[360] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[361] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, *Robust speech recognition via large-scale weak supervision*, 2022. arXiv: 2212.04356 [eess.AS].

[362] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, 2019.

[363] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *JMLR*, 2020.

[364] S. Rahman, S. Khan, and N. Barnes, "Improved visual-semantic alignment for zero-shot object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 11 932–11 939, 2020.

[365] F. Rajič, L. Ke, Y.-W. Tai, C.-K. Tang, M. Danelljan, and F. Yu, "Segment anything meets point tracking," *arXiv preprint arXiv:2307.01197*, 2023.

[366] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

[367] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-Shot Text-to-Image Generation," in *ICML*, 2021.

[368] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International Conference on Machine Learning*, PMLR, pp. 8821–8831, 2021.

[369] Y. Rao, W. Zhao, G. Chen, Y. Tang, Z. Zhu, G. Huang, J. Zhou, and J. Lu, "Denseclip: Language-guided dense prediction with context-aware prompting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18 082–18 091, 2022.

[370] A. Razavi, A. Van den Oord, and O. Vinyals, "Generating diverse high-fidelity images with vq-vae-2," in *NeurIPS*, 2019.

[371] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

[372] S. Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barth-Maron, M. Gimenez, Y. Sulsky, J. Kay, J. T. Springenberg, *et al.*, "A generalist agent," *arXiv preprint arXiv:2205.06175*, 2022.

[373] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NeurIPS*, 2015.

[374] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor, "Imagenet-21k pretraining for the masses," *arXiv preprint arXiv:2104.10972*, 2021.

[375] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, *High-resolution image synthesis with latent diffusion models*, 2021. arXiv: 2112.10752 [cs.CV].

[376] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

[377] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, Springer, pp. 234–241, 2015.

[378] S. Roy, T. Wald, G. Koehler, M. R. Rokuss, N. Disch, J. Holzschuh, D. Zimmerer, and K. H. Maier-Hein, "Sam. md: Zero-shot medical image segmentation capabilities of the segment anything model," *arXiv preprint arXiv:2304.05396*, 2023.

[379] L. Ruan and Q. Jin, "Survey: Transformer based video-language pre-training," *AI Open*, 2022.

[380] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22 500–22 510, 2023.

[381] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *IJCV*, 2015.

[382] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *arXiv preprint arXiv:2205.11487*, 2022.

[383] M. B. Sariyildiz, J. Perez, and D. Larlus, "Learning visual representations with caption annotations," in *ECCV*, 2020.
[384] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda, and T. Scialom, "Toolformer: Language models can teach themselves to use tools," *arXiv preprint arXiv:2302.04761*, 2023.
[385] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *NeurIPS*, 2022.
[386] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki, "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs," *arXiv preprint arXiv:2111.02114*, 2021.
[387] D. Schwenk, A. Khandelwal, C. Clark, K. Marino, and R. Mottaghi, "A-okvqa: A benchmark for visual question answering using world knowledge," *arXiv preprint arXiv:2206.01718*, 2022.
[388] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *ACL*, 2016.
[389] S. Shao, Z. Li, T. Zhang, C. Peng, G. Yu, X. Zhang, J. Li, and J. Sun, "Objects365: A large-scale, high-quality dataset for object detection," in *ICCV*, 2019.
[390] W. Shao, Y. Hu, P. Gao, M. Lei, K. Zhang, F. Meng, P. Xu, S. Huang, H. Li, Y. Qiao, *et al.*, "Tiny lvlm-ehub: Early multimodal experiments with bard," *arXiv preprint arXiv:2308.03729*, 2023.
[391] ShareGPT, 2023. URL: https://sharegpt.com/.
[392] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *ACL*, 2018.
[393] Q. Shen, X. Yang, and X. Wang, "Anything-3d: Towards single-view anything reconstruction in the wild," *arXiv preprint arXiv:2304.10261*, 2023.
[394] S. Shen, C. Li, X. Hu, Y. Xie, J. Yang, P. Zhang, A. Rohrbach, Z. Gan, L. Wang, L. Yuan, C. Liu, K. Keutzer, T. Darrell, A. Rohrbach, and J. Gao, "K-lite: Learning transferable visual models with external knowledge," in *NeurIPS*, 2022.

[395] S. Shen, L. H. Li, H. Tan, M. Bansal, A. Rohrbach, K.-W. Chang, Z. Yao, and K. Keutzer, "How much can clip benefit vision-and-language tasks?" In *ICLR*, 2022.

[396] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang, "Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface," *arXiv preprint arXiv:2303.17580*, 2023.

[397] S. Sheynin, O. Ashual, A. Polyak, U. Singer, O. Gafni, E. Nachmani, and Y. Taigman, "Knn-diffusion: Image generation via large-scale retrieval," *arXiv preprint arXiv:2204.02849*, 2022.

[398] J. Shi, W. Xiong, Z. Lin, and H. J. Jung, "Instantbooth: Personalized text-to-image generation without test-time finetuning," *arXiv preprint arXiv:2304.03411*, 2023.

[399] P. Shi, J. Qiu, S. M. D. Abaxi, H. Wei, F. P.-W. Lo, and W. Yuan, "Generalist vision foundation models for medical imaging: A case study of segment anything model on zero-shot medical segmentation," *Diagnostics*, 2023.

[400] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, and S. Singh, "Autoprompt: Eliciting knowledge from language models with automatically generated prompts," *arXiv preprint arXiv:2010.15980*, 2020.

[401] O. Sidorov, R. Hu, M. Rohrbach, and A. Singh, "Textcaps: A dataset for image captioning with reading comprehension," in *ECCV*, 2020.

[402] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, Springer, pp. 746–760, 2012.

[403] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, *et al.*, "Make-a-video: Text-to-video generation without text-video data," *arXiv preprint arXiv:2209.14792*, 2022.

[404] A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, and D. Kiela, "Flava: A foundational language and vision alignment model," in *CVPR*, 2022.

[405] M. Singh, Q. Duval, K. V. Alwala, H. Fan, V. Aggarwal, A. Adcock, A. Joulin, P. Dollár, C. Feichtenhofer, R. Girshick, *et al.*, "The effectiveness of mae pre-pretraining for billion-scale pretraining," *arXiv preprint arXiv:2303.13496*, 2023.

[406] M. Singh, L. Gustafson, A. Adcock, V. de Freitas Reis, B. Gedik, R. P. Kosaraju, D. Mahajan, R. Girshick, P. Dollár, and L. Van Der Maaten, "Revisiting weakly supervised pre-training of visual perception models," in *CVPR*, 2022.

[407] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*, PMLR, pp. 2256–2265, 2015.

[408] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, *Consistency models*, 2023. arXiv: 2303.01469 [cs.LG].

[409] Y. Song and S. Ermon, "Improved techniques for training score-based generative models," *Advances in neural information processing systems*, vol. 33, 2020, pp. 12 438–12 448.

[410] Y. Song, W. Xiong, D. Zhu, C. Li, K. Wang, Y. Tian, and S. Li, "Restgpt: Connecting large language models with real-world applications via restful apis," *arXiv preprint arXiv:2306.06624*, 2023.

[411] K. Srinivasan, K. Raman, J. Chen, M. Bendersky, and M. Najork, "Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning," *arXiv preprint arXiv:2103.01913*, 2021.

[412] *Stable diffusion*, 2022. URL: https://github.com/CompVis/stable-diffusion.

[413] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "VL-BERT: Pre-training of generic visual-linguistic representations," in *ICLR*, 2019.

[414] Y. Su, T. Lan, H. Li, J. Xu, Y. Wang, and D. Cai, "Pandagpt: One model to instruction-follow them all," *arXiv preprint arXiv:2305.16355*, 2023.

[415] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *ICCV*, 2017.

[416] Q. Sun, Q. Yu, Y. Cui, F. Zhang, X. Zhang, Y. Wang, H. Gao, J. Liu, T. Huang, and X. Wang, "Generative pretraining in multimodality," *arXiv preprint arXiv:2307.05222*, 2023.

[417] Y. Sun, Y. Yang, H. Peng, Y. Shen, Y. Yang, H. Hu, L. Qiu, and H. Koike, "Imagebrush: Learning visual in-context instructions for exemplar-based image manipulation," *arXiv preprint arXiv:2308.00906*, 2023.

[418] Y. Sun, C. Zhu, S. Zheng, K. Zhang, Z. Shui, X. Yu, Y. Zhao, H. Li, Y. Zhang, R. Zhao, *et al.*, "Pathasst: Redefining pathology through generative foundation ai assistant for pathology," *arXiv preprint arXiv:2305.15072*, 2023.

[419] D. Surís, S. Menon, and C. Vondrick, "Vipergpt: Visual inference via python execution for reasoning," *arXiv preprint arXiv:2303.08128*, 2023.

[420] *Svd-xt*, 2023. URL: https://huggingface.co/stabilityai/stable-video-diffusion-img2vid-xt.

[421] H. Tan and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," in *EMNLP*, 2019.

[422] L. Tang, H. Xiao, and B. Li, "Can sam segment anything? when sam meets camouflaged object detection," *arXiv preprint arXiv:2304.04709*, 2023.

[423] Z. Tang, Z. Yang, C. Zhu, M. Zeng, and M. Bansal, "Any-to-any generation via composable diffusion," 2023. arXiv: 2305.11846 [cs.CV].

[424] C. Tao, X. Zhu, W. Su, G. Huang, B. Li, J. Zhou, Y. Qiao, X. Wang, and J. Dai, "Siamese image modeling for self-supervised vision representation learning," in *CVPR*, 2023.

[425] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto, *Stanford alpaca: An instruction-following llama model*, 2023. URL: https://github.com/tatsu-lab/stanford_alpaca.

[426] G. G. Team, *Gemini: A family of highly capable multimodal models*, 2023. arXiv: 2312.11805 [cs.CL].

[427] The Vicuna Team, *Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality*, 2023. URL: https://vicuna.lmsys.org/.

[428] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "Yfcc100m: The new data in multimedia research," *Communications of the ACM*, 2016.

[429] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *ECCV*, 2020.

[430] Z. Tian, C. Shen, and H. Chen, "Conditional convolutions for instance segmentation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, Springer, pp. 282–298, 2020.

[431] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *NeurIPS*, 2022.

[432] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *ICML*, 2021.

[433] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[434] M. Tschannen, M. Kumar, A. Steiner, X. Zhai, N. Houlsby, and L. Beyer, "Image captioners are scalable vision learners too," *arXiv preprint arXiv:2306.07915*, 2023.

[435] T. Tu, S. Azizi, D. Driess, M. Schaekermann, M. Amin, P.-C. Chang, A. Carroll, C. Lau, R. Tanno, I. Ktena, *et al.*, "Towards generalist biomedical ai," *arXiv preprint arXiv:2307.14334*, 2023.

[436] A. Vahdat and J. Kautz, "Nvae: A deep hierarchical variational autoencoder," 2020.

[437] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017.

[438] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *CVPR*, 2015.

[439] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 mscoco image captioning challenge," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 4, 2016, pp. 652–663.

[440] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, Ieee, vol. 1, pp. I–I, 2001.

[441] B. Wang and A. Komatsuzaki, *GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model*, May 2021. URL: https://github.com/kingoflolz/mesh-transformer-jax.

[442] B. Wang, F. Wu, X. Han, J. Peng, H. Zhong, P. Zhang, X. Dong, W. Li, W. Li, J. Wang, *et al.*, "Vigc: Visual instruction generation and correction," *arXiv preprint arXiv:2308.12714*, 2023.

[443] F. Wang, T. Kong, R. Zhang, H. Liu, and H. Li, "Self-supervised learning by estimating twin class distribution," *TIP*, 2023.

[444] H. Wang, Y. Wang, Z. Zhou, X. Ji, Z. Li, D. Gong, J. Zhou, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," *CVPR*, 2018.

[445] H. Wang, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen, "Max-deeplab: End-to-end panoptic segmentation with mask transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5463–5474, 2021.

[446] J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, and L. Wang, "Git: A generative image-to-text transformer for vision and language," *arXiv preprint arXiv:2205.14100*, 2022.

[447] J. Wang, D. Chen, C. Luo, X. Dai, L. Yuan, Z. Wu, and Y.-G. Jiang, "Chatvideo: A tracklet-centric multimodal and versatile video understanding system," *arXiv preprint arXiv:2304.14407*, 2023.

[448] J. Wang, L. Meng, Z. Weng, B. He, Z. Wu, and Y.-G. Jiang, *To see is to believe: Prompting gpt-4v for better visual instruction tuning*, 2023. arXiv: 2311.07574 [cs.CV].

[449] J. Wang, Y. Zhou, G. Xu, P. Shi, C. Zhao, H. Xu, Q. Ye, M. Yan, J. Zhang, J. Zhu, *et al.*, "Evaluation and analysis of hallucination in large vision-language models," *arXiv preprint arXiv:2308.15126*, 2023.

[450] L. Wang, B. Huang, Z. Zhao, Z. Tong, Y. He, Y. Wang, Y. Wang, and Y. Qiao, "Videomae v2: Scaling video masked autoencoders with dual masking," in *CVPR*, 2023.

[451] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang, "Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," in *ICML*, 2022.

[452] R. Wang, D. Chen, Z. Wu, Y. Chen, X. Dai, M. Liu, Y.-G. Jiang, L. Zhou, and L. Yuan, "Bevt: Bert pretraining of video transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14 733–14 743, 2022.

[453] T. Wang, L. Li, K. Lin, C.-C. Lin, Z. Yang, H. Zhang, Z. Liu, and L. Wang, "Disco: Disentangled control for referring human dance generation in real world," *arXiv preprint arXiv:2307.00040*, 2023.

[454] T. Wang, J. Zhang, J. Fei, Y. Ge, H. Zheng, Y. Tang, Z. Li, M. Gao, S. Zhao, Y. Shan, *et al.*, "Caption anything: Interactive image description with diverse multimodal controls," *arXiv preprint arXiv:2305.02677*, 2023.

[455] W. Wang, J. Liu, Z. Lin, J. Yan, S. Chen, C. Low, T. Hoang, J. Wu, J. H. Liew, H. Yan, D. Zhou, and J. Feng, *Magicvideo-v2: Multi-stage high-aesthetic video generation*, 2024. arXiv: 2401.04468 [cs.CV].

[456] W. Wang, Z. Chen, X. Chen, J. Wu, X. Zhu, G. Zeng, P. Luo, T. Lu, J. Zhou, Y. Qiao, *et al.*, "VisionLLM: Large language model is also an open-ended decoder for vision-centric tasks," *arXiv preprint arXiv:2305.11175*, 2023.

[457] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som, and F. Wei, "Image as a foreign language: Beit pretraining for all vision and vision-language tasks," *arXiv preprint arXiv:2208.10442*, 2022.

[458] W. Wang, H. Bao, L. Dong, and F. Wei, "Vlmo: Unified vision-language pre-training with mixture-of-modality-experts," *arXiv preprint arXiv:2111.02358*, 2021.

[459] X. Wang, W. Wang, Y. Cao, C. Shen, and T. Huang, "Images speak in images: A generalist painter for in-context visual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6830–6839, 2023.

[460] X. Wang, X. Zhang, Y. Cao, W. Wang, C. Shen, and T. Huang, "Seggpt: Segmenting everything in context," *arXiv preprint arXiv:2304.03284*, 2023.
[461] Y. Wang, H. Ivison, P. Dasigi, J. Hessel, T. Khot, K. R. Chandu, D. Wadden, K. MacMillan, N. A. Smith, I. Beltagy, *et al.*, "How far can camels go? exploring the state of instruction tuning on open resources," *arXiv preprint arXiv:2306.04751*, 2023.
[462] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi, "Self-instruct: Aligning language model with self generated instructions," *arXiv preprint arXiv:2212.10560*, 2022.
[463] Y. Wang, S. Mishra, P. Alipoormolabashi, Y. Kordi, A. Mirzaei, A. Arunkumar, A. Ashok, A. S. Dhanasekaran, A. Naik, D. Stap, *et al.*, "Benchmarking generalization via in-context instructions on 1,600+ language tasks," *arXiv preprint arXiv:2204.07705*, 2022.
[464] Z. Wang, H. Huang, Y. Zhao, Z. Zhang, and Z. Zhao, "Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes," *arXiv preprint arXiv:2308.08769*, 2023.
[465] Z. Wang, Y. Jiang, Y. Lu, Y. Shen, P. He, W. Chen, Z. Wang, and M. Zhou, *In-context learning unlocked for diffusion models*, 2023. arXiv: 2305.01115 [cs.CV].
[466] Z. Wang, C. Lu, Y. Wang, F. Bao, C. Li, H. Su, and J. Zhu, "Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation," *arXiv preprint arXiv:2305.16213*, 2023.
[467] Z. Wang, J. Chen, and S. C. Hoi, "Deep learning for image super-resolution: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, 2020, pp. 3365–3387.
[468] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "Simvlm: Simple visual language model pretraining with weak supervision," in *ICLR*, 2022.
[469] F. Weers, V. Shankar, A. Katharopoulos, Y. Yang, and T. Gunter, "Masked autoencoding does not help natural language supervision at scale," in *CVPR*, 2023.

[470] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer, "Masked feature prediction for self-supervised visual pre-training," in *CVPR*, 2021.

[471] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. Chi, Q. Le, and D. Zhou, "Chain of thought prompting elicits reasoning in large language models," *arXiv preprint arXiv:2201.11903*, 2022.

[472] L. Wei, Z. Jiang, W. Huang, and L. Sun, "Instructiongpt-4: A 200-instruction paradigm for fine-tuning minigpt-4," *arXiv preprint arXiv:2308.12067*, 2023.

[473] L. Wei, L. Xie, W. Zhou, H. Li, and Q. Tian, "Mvp: Multimodality-guided visual pre-training," in *ECCV*, 2022.

[474] Y. Wei, Y. Zhang, Z. Ji, J. Bai, L. Zhang, and W. Zuo, "Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation," *arXiv preprint arXiv:2302.13848*, 2023.

[475] L. Weng, "Llm-powered autonomous agents," *lilianweng.github.io*, Jun. 2023. URL: https://lilianweng.github.io/posts/2023-06-23-agent/.

[476] C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "Towards generalist foundation model for radiology," *arXiv preprint arXiv:2308.02463*, 2023.

[477] C. Wu, S. Yin, W. Qi, X. Wang, Z. Tang, and N. Duan, "Visual chatgpt: Talking, drawing and editing with visual foundation models," *arXiv preprint arXiv:2303.04671*, 2023.

[478] J. Wu, J. Wang, Z. Yang, Z. Gan, Z. Liu, J. Yuan, and L. Wang, "Grit: A generative region-to-text transformer for object understanding," *arXiv preprint arXiv:2212.00280*, 2022.

[479] J. Wu, J. Lu, A. Sabharwal, and R. Mottaghi, "Multi-modal answer validation for knowledge-based VQA," *arXiv preprint arXiv:2103.12248*, 2021.

[480] J. Wu, Y. Jiang, P. Sun, Z. Yuan, and P. Luo, "Language as queries for referring video object segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4974–4984, 2022.

[481] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua, "Next-gpt: Any-to-any multimodal llm," *CoRR*, vol. abs/2309.05519, 2023.

[482] W. Wu, A. Timofeev, C. Chen, B. Zhang, K. Duan, S. Liu, Y. Zheng, J. Shlens, X. Du, Z. Gan, *et al.*, "Mofi: Learning image representations from noisy entity annotated images," *arXiv preprint arXiv:2306.07952*, 2023.

[483] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2411–2418, 2013.

[484] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *CVPR*, 2018.

[485] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly," *TPAMI*, 2018.

[486] Z. Xiao, Y. Chen, L. Zhang, J. Yao, Z. Wu, X. Yu, Y. Pan, L. Zhao, C. Ma, X. Liu, *et al.*, "Instruction-vit: Multi-modal prompts for instruction learning in vit," *arXiv preprint arXiv:2305.00201*, 2023.

[487] D. Xie, R. Wang, J. Ma, C. Chen, H. Lu, D. Yang, F. Shi, and X. Lin, "Edit everything: A text-guided generative system for images editing," *arXiv preprint arXiv:2304.14006*, 2023.

[488] X. Xie, L. Fu, Z. Zhang, Z. Wang, and X. Bai, *Toward understanding wordart: Corner-guided transformer for scene text recognition*, 2022. arXiv: 2208.00438 [cs.CV].

[489] Z. Xie, Y. Lin, Z. Yao, Z. Zhang, Q. Dai, Y. Cao, and H. Hu, "Self-supervised learning with swin transformers," *arXiv preprint arXiv:2105.04553*, 2021.

[490] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "Simmim: A simple framework for masked image modeling," in *CVPR*, 2022.

[491] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, Y. Wei, Q. Dai, and H. Hu, "On data scaling in masked image modeling," in *CVPR*, 2023.

[492] D. Xu, W. Ouyang, X. Wang, and N. Sebe, "Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 675–684, 2018.

[493] H. Xu, M. Yan, C. Li, B. Bi, S. Huang, W. Xiao, and F. Huang, *E2e-vlp: End-to-end vision-language pre-training enhanced by visual learning*, 2021. arXiv: 2106.01804 [cs.CV].

[494] J. Xu, S. De Mello, S. Liu, W. Byeon, T. Breuel, J. Kautz, and X. Wang, "Groupvit: Semantic segmentation emerges from text supervision," in *CVPR*, 2022.

[495] J. Xu, S. Liu, A. Vahdat, W. Byeon, X. Wang, and S. De Mello, "Open-vocabulary panoptic segmentation with text-to-image diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2955–2966, 2023.

[496] J. Xu, L. Xu, Y. Yang, X. Li, Y. Xie, Y.-J. Huang, and Y. Li, "U-llava: Unifying multi-modal tasks via large language model," *arXiv preprint arXiv:2311.05348*, 2023.

[497] P. Xu, W. Shao, K. Zhang, P. Gao, S. Liu, M. Lei, F. Meng, S. Huang, Y. Qiao, and P. Luo, "Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models," *arXiv preprint arXiv:2306.09265*, 2023.

[498] R. Xu, X. Wang, T. Wang, Y. Chen, J. Pang, and D. Lin, "Pointllm: Empowering large language models to understand point clouds," *arXiv preprint arXiv:2308.16911*, 2023.

[499] Y. Xu, Y. Zhao, Z. Xiao, and T. Hou, *Ufogen: You forward once large scale text-to-image generation via diffusion gans*, 2023. arXiv: 2311.09257 [cs.CV].

[500] Z. Xu, Y. Shen, and L. Huang, "Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning," *arXiv preprint arXiv:2212.10773*, 2022.

[501] B. Yan, Y. Jiang, J. Wu, D. Wang, P. Luo, Z. Yuan, and H. Lu, "Universal instance perception as object discovery and retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15 325–15 336, 2023.

[502] B. Yang, S. Gu, B. Zhang, T. Zhang, X. Chen, X. Sun, D. Chen, and F. Wen, "Paint by example: Exemplar-based image editing with diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18 381–18 391, 2023.

[503] J. Yang, C. Li, P. Zhang, B. Xiao, C. Liu, L. Yuan, and J. Gao, "Unified contrastive learning in image-text-label space," in *CVPR*, 2022.

[504] J. Yang, C. Li, P. Zhang, B. Xiao, L. Yuan, C. Liu, and J. Gao, "Unicl: Unified contrastive learning in image-text-label space," in *CVPR*, 2022.

[505] J. Yang, M. Gao, Z. Li, S. Gao, F. Wang, and F. Zheng, "Track anything: Segment anything meets videos," *arXiv preprint arXiv:2304.11968*, 2023.

[506] R. Yang, L. Song, Y. Li, S. Zhao, Y. Ge, X. Li, and Y. Shan, "Gpt4tools: Teaching large language model to use tools via self-instruction," *arXiv preprint arXiv:2305.18752*, 2023.

[507] Z. Yang, J. Wang, Y. Tang, K. Chen, H. Zhao, and P. H. Torr, "Lavt: Language-aware vision transformer for referring image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18 155–18 165, 2022.

[508] Z. Yang, Z. Gan, J. Wang, X. Hu, F. Ahmed, Z. Liu, Y. Lu, and L. Wang, "Crossing the format boundary of text and boxes: Towards unified vision-language modeling," *arXiv preprint arXiv:2111.12085*, 2021.

[509] Z. Yang, Z. Gan, J. Wang, X. Hu, F. Ahmed, Z. Liu, Y. Lu, and L. Wang, "Unitab: Unifying text and box outputs for grounded vision-language modeling," in *European Conference on Computer Vision*, Springer, pp. 521–539, 2022.

[510] Z. Yang, Z. Gan, J. Wang, X. Hu, Y. Lu, Z. Liu, and L. Wang, "An empirical study of gpt-3 for few-shot knowledge-based vqa," in *AAAI*, 2022.

[511] Z. Yang, J. Wang, Z. Gan, L. Li, K. Lin, C. Wu, N. Duan, Z. Liu, C. Liu, M. Zeng, *et al.*, "Reco: Region-controlled text-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14 246–14 255, 2023.

[512] Z. Yang, W. Ping, Z. Liu, V. Korthikanti, W. Nie, D.-A. Huang, L. Fan, Z. Yu, S. Lan, B. Li, *et al.*, "Re-vilm: Retrieval-augmented visual language model for zero and few-shot image captioning," *arXiv preprint arXiv:2302.04858*, 2023.

[513] Z. Yang*, L. Li*, J. Wang*, K. Lin*, E. Azarnasab*, F. Ahmed*, Z. Liu, C. Liu, M. Zeng, and L. Wang, "Mm-react: Prompting chatgpt for multimodal reasoning and action," 2023.

[514] L. Yao, J. Han, X. Liang, D. Xu, W. Zhang, Z. Li, and H. Xu, "Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23 497–23 506, 2023.

[515] L. Yao, J. Han, Y. Wen, X. Liang, D. Xu, W. Zhang, Z. Li, C. Xu, and H. Xu, "Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection," *arXiv preprint arXiv:2209.09407*, 2022.

[516] L. Yao, R. Huang, L. Hou, G. Lu, M. Niu, H. Xu, X. Liang, Z. Li, X. Jiang, and C. Xu, "Filip: Fine-grained interactive language-image pre-training," in *ICLR*, 2022.

[517] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," *arXiv preprint arXiv:2210.03629*, 2022.

[518] M. Yasunaga, A. Aghajanyan, W. Shi, R. James, J. Leskovec, P. Liang, M. Lewis, L. Zettlemoyer, and W.-t. Yih, "Retrieval-augmented multimodal language modeling," *arXiv preprint arXiv:2211.12561*, 2022.

[519] J. Ye, A. Hu, H. Xu, Q. Ye, M. Yan, Y. Dan, C. Zhao, G. Xu, C. Li, J. Tian, *et al.*, "Mplug-docowl: Modularized multimodal large language model for document understanding," *arXiv preprint arXiv:2307.02499*, 2023.

[520] L. Ye, M. Rochan, Z. Liu, and Y. Wang, "Cross-modal self-attention network for referring image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10 502–10 511, 2019.

[521] M. Ye, X. Zhang, P. C. Yuen, and S.-F. Chang, "Unsupervised embedding learning via invariant and spreading instance feature," in *CVPR*, 2019.

[522] Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, Y. Shi, *et al.*, "Mplug-owl: Modularization empowers large language models with multimodality," *arXiv preprint arXiv:2304.14178*, 2023.

[523] K. Yi, Y. Ge, X. Li, S. Yang, D. Li, J. Wu, Y. Shan, and X. Qie, "Masked image modeling with denoising contrast," *arXiv preprint arXiv:2205.09616*, 2022.

[524] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *Acm computing surveys (CSUR)*, vol. 38, no. 4, 2006, 13–es.

[525] D. Yin, L. Dong, H. Cheng, X. Liu, K.-W. Chang, F. Wei, and J. Gao, "A survey of knowledge-intensive nlp with pre-trained language models," *arXiv preprint arXiv:2202.08772*, 2022.

[526] Z. Yin, J. Wang, J. Cao, Z. Shi, D. Liu, M. Li, L. Sheng, L. Bai, X. Huang, Z. Wang, *et al.*, "Lamm: Language-assisted multimodal instruction-tuning dataset, framework, and benchmark," *arXiv preprint arXiv:2306.06687*, 2023.

[527] S. Yoon, W. Y. Kang, S. Jeon, S. Lee, C. Han, J. Park, and E.-S. Kim, "Image-to-image retrieval by learning similarity between scene graphs," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 10 718–10 726, 2021.

[528] H. You, H. Zhang, Z. Gan, X. Du, B. Zhang, Z. Wang, L. Cao, S.-F. Chang, and Y. Yang, "Ferret: Refer and ground anything anywhere at any granularity," *arXiv preprint arXiv:2310.07704*, 2023.

[529] J. Yu, X. Li, J. Y. Koh, H. Zhang, R. Pang, J. Qin, A. Ku, Y. Xu, J. Baldridge, and Y. Wu, "Vector-quantized image modeling with improved vqgan," *arXiv preprint arXiv:2110.04627*, 2021.

[530] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "Coca: Contrastive captioners are image-text foundation models," *TMLR*, 2022.

[531] J. Yu, Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan, *et al.*, "Scaling autoregressive models for content-rich text-to-image generation," *Transactions on Machine Learning Research*, 2022.

[532] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *ECCV*, 2016.

[533] L. Yu and et al, "Scaling autoregressive multi-modal models: Pretraining and instruction tuning," 2023.

[534] Q. Yu, J. Li, W. Ye, S. Tang, and Y. Zhuang, "Interactive data synthesis for systematic vision adaptation via llms-aigcs collaboration," *arXiv preprint arXiv:2305.12799*, 2023.

[535] Q. Yu, J. He, X. Deng, X. Shen, and L.-C. Chen, "Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip," *arXiv preprint arXiv:2308.02487*, 2023.

[536] T. Yu, R. Feng, R. Feng, J. Liu, X. Jin, W. Zeng, and Z. Chen, "Inpaint anything: Segment anything meets image inpainting," *arXiv preprint arXiv:2304.06790*, 2023.

[537] W. Yu, Z. Yang, L. Li, J. Wang, K. Lin, Z. Liu, X. Wang, and L. Wang, "Mm-vet: Evaluating large multimodal models for integrated capabilities," *arXiv preprint arXiv:2308.02490*, 2023.

[538] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *CVPR*, 2019.

[539] L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li, C. Liu, M. Liu, Z. Liu, Y. Lu, Y. Shi, L. Wang, J. Wang, B. Xiao, Z. Xiao, J. Yang, M. Zeng, L. Zhou, and P. Zhang, "Florence: A new foundation model for computer vision," *arXiv preprint arXiv:2111.11432*, 2021.

[540] A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese, "Taskonomy: Disentangling task transfer learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3712–3722, 2018.

[541] Y. Zang, W. Li, J. Han, K. Zhou, and C. C. Loy, "Contextual object detection with multimodal large language models," *arXiv preprint arXiv:2305.18279*, 2023.

[542] Y. Zang, W. Li, K. Zhou, C. Huang, and C. C. Loy, "Open-vocabulary detr with conditional matching," *arXiv preprint arXiv:2203.11876*, 2022.

[543] A. Zareian, K. D. Rosa, D. H. Hu, and S.-F. Chang, "Open-vocabulary object detection using captions," in *CVPR*, 2021.

[544] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *ICML*, 2021.

[545] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi, "From recognition to cognition: Visual commonsense reasoning," in *CVPR*, 2019.

[546] Y. Zeng, X. Zhang, and H. Li, "Multi-grained vision language pre-training: Aligning texts with visual concepts," in *ICML*, 2022.

[547] Y. Zeng, Z. Lin, J. Zhang, Q. Liu, J. Collomosse, J. Kuen, and V. M. Patel, "Scenecomposer: Any-level semantic image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22 468–22 478, 2023.

[548] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," in *CVPR*, 2022.

[549] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," *arXiv preprint arXiv:2303.15343*, 2023.

[550] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, and L. Beyer, "Lit: Zero-shot transfer with locked-image text tuning," in *CVPR*, 2022.

[551] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal intelligence: Representation learning, information fusion, and applications," *JSTSP*, 2020.

[552] C. Zhang, S. Zheng, C. Li, Y. Qiao, T. Kang, X. Shan, C. Zhang, C. Qin, F. Rameau, S.-H. Bae, *et al.*, "A survey on segment anything model (sam): Vision foundation model meets prompt engineering," *arXiv preprint arXiv:2306.06211*, 2023.

[553] C. Zhang, C. Zhang, M. Zhang, and I. S. Kweon, "Text-to-image diffusion model in generative ai: A survey," *arXiv preprint arXiv:2303.07909*, 2023.

[554] C. Zhang, L. Liu, Y. Cui, G. Huang, W. Lin, Y. Yang, and Y. Hu, "A comprehensive survey on segment anything model for vision and beyond," *arXiv preprint arXiv:2305.08196*, 2023.

[555] H. Zhang, X. Li, and L. Bing, "Video-llama: An instruction-tuned audio-visual language model for video understanding," *arXiv preprint arXiv:2306.02858*, 2023.

[556] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," *arXiv preprint arXiv:2203.03605*, 2022.

[557] H. Zhang, F. Li, X. Zou, S. Liu, C. Li, J. Gao, J. Yang, and L. Zhang, "A simple framework for open-vocabulary segmentation and detection," *arXiv preprint arXiv:2303.08131*, 2023.

[558] H. Zhang, H. Li, F. Li, T. Ren, X. Zou, S. Liu, S. Huang, J. Gao, L. Zhang, C. Li, and J. Yang, *Llava-grounding: Grounded visual chat with large multimodal models*, 2023.

[559] H. Zhang, P. Zhang, X. Hu, Y.-C. Chen, L. H. Li, X. Dai, L. Wang, L. Yuan, J.-N. Hwang, and J. Gao, "Glipv2: Unifying localization and vision-language understanding," in *ECCV*, 2022.

[560] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *arXiv preprint arXiv:2304.00685*, 2023.

[561] J. Zhang, F. Shen, L. Liu, F. Zhu, M. Yu, L. Shao, H. T. Shen, and L. Van Gool, "Generative domain-migration hashing for sketch-to-image retrieval," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 297–314, 2018.

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

[563] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, "VinVL: Revisiting visual representations in vision-language models," in *CVPR*, 2021.

[564] R. Zhang, Z. Jiang, Z. Guo, S. Yan, J. Pan, H. Dong, P. Gao, and H. Li, "Personalize segment anything model with one shot," *arXiv preprint arXiv:2305.03048*, 2023.

[565] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, *et al.*, "Instruction tuning for large language models: A survey," *arXiv preprint arXiv:2308.10792*, 2023.

[566] S. Zhang, P. Sun, S. Chen, M. Xiao, W. Shao, W. Zhang, K. Chen, and P. Luo, "Gpt4roi: Instruction tuning large language model on region-of-interest," *arXiv preprint arXiv:2307.03601*, 2023.

[567] S. Zhang, C. Gong, L. Wu, X. Liu, and M. Zhou, "Automl-gpt: Automatic machine learning with gpt," *arXiv preprint arXiv:2305.02499*, 2023.

[568] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[569] W. Zhang, S. M. Aljunied, C. Gao, Y. K. Chia, and L. Bing, "M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models," *arXiv preprint arXiv:2306.05179*, 2023.

[570] X. Zhang, C. Wu, Z. Zhao, W. Lin, Y. Zhang, Y. Wang, and W. Xie, "Pmc-vqa: Visual instruction tuning for medical visual question answering," *arXiv preprint arXiv:2305.10415*, 2023.

[571] X. Zhang, Y. Tian, W. Huang, Q. Ye, Q. Dai, L. Xie, and Q. Tian, "Hivit: Hierarchical vision transformer meets masked image modeling," *arXiv preprint arXiv:2205.14949*, 2022.

[572] X. Zhang, J. Chen, J. Yuan, Q. Chen, J. Wang, X. Wang, S. Han, X. Chen, J. Pi, K. Yao, *et al.*, "Cae v2: Context autoencoder with clip target," *arXiv preprint arXiv:2211.09799*, 2022.

[573] Y. Zhang, R. Zhang, J. Gu, Y. Zhou, N. Lipka, D. Yang, and T. Sun, "Llavar: Enhanced visual instruction tuning for text-rich image understanding," *arXiv preprint arXiv:2306.17107*, 2023.

[574] Y. Zhang and R. Jiao, "How segment anything model (sam) boost medical image segmentation?" *arXiv preprint arXiv:2305.03678*, 2023.

[575] Y. Zhang, X. Huang, J. Ma, Z. Li, Z. Luo, Y. Xie, Y. Qin, T. Luo, Y. Li, S. Liu, *et al.*, "Recognize anything: A strong image tagging model," *arXiv preprint arXiv:2306.03514*, 2023.

[576] B. Zhao, B. Wu, and T. Huang, "Svit: Scaling up visual instruction tuning," *arXiv preprint arXiv:2307.04087*, 2023.
[577] S. Zhao, D. Chen, Y.-C. Chen, J. Bao, S. Hao, L. Yuan, and K.-Y. K. Wong, "Uni-controlnet: All-in-one control to text-to-image diffusion models," *arXiv preprint arXiv:2305.16322*, 2023.
[578] Y. Zhao, Z. Lin, D. Zhou, Z. Huang, J. Feng, and B. Kang, "Bubogpt: Enabling visual grounding in multi-modal llms," *arXiv preprint arXiv:2307.08581*, 2023.
[579] Y. Zhao, T. Pang, C. Du, X. Yang, C. Li, N.-M. Cheung, and M. Lin, "On evaluating adversarial robustness of large vision-language models," *arXiv preprint arXiv:2305.16934*, 2023.
[580] Z. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh, "Calibrate before use: Improving few-shot performance of language models," in *International Conference on Machine Learning*, PMLR, pp. 12 697–12 706, 2021.
[581] Y. Zhong, J. Yang, P. Zhang, C. Li, N. Codella, L. H. Li, L. Zhou, X. Dai, L. Yuan, Y. Li, *et al.*, "Regionclip: Region-based language-image pretraining," in *CVPR*, 2022.
[582] Y. Zhong, J. Yang, P. Zhang, C. Li, N. Codella, L. H. Li, L. Zhou, X. Dai, L. Yuan, Y. Li, *et al.*, "Regionclip: Region-based language-image pretraining," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16 793–16 803, 2022.
[583] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 633–641, 2017.
[584] C. Zhou, C. C. Loy, and B. Dai, "Extract free dense labels from clip," in *ECCV*, 2022.
[585] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, *et al.*, "Lima: Less is more for alignment," *arXiv preprint arXiv:2305.11206*, 2023.
[586] G. Zhou, Y. Hong, and Q. Wu, "Navgpt: Explicit reasoning in vision-and-language navigation with large language models," *arXiv preprint arXiv:2305.16986*, 2023.

[587] J. Zhou, L. Dong, Z. Gan, L. Wang, and F. Wei, "Non-contrastive learning meets language-image pre-training," in *CVPR*, 2023.

[588] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong, "Ibot: Image bert pre-training with online tokenizer," *arXiv preprint arXiv:2111.07832*, 2021.

[589] T. Zhou, Y. Zhang, Y. Zhou, Y. Wu, and C. Gong, "Can sam segment polyps?" *arXiv preprint arXiv:2304.07583*, 2023.

[590] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra, "Detecting twenty-thousand classes using image-level supervision," in *European Conference on Computer Vision*, Springer, pp. 350–368, 2022.

[591] Y. Zhou, C. Li, C. Chen, J. Gao, and J. Xu, "Lafite2: Few-shot text-to-image generation," *arXiv preprint arXiv:2210.14124*, 2022.

[592] Y. Zhou and N. Shimada, "Vision + language applications: A survey," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 826–842, 2023.

[593] 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 [cs.CV].

[594] P. Zhu, H. Wang, and V. Saligrama, "Zero shot detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 4, 2019, pp. 998–1010.

[595] P. Zhu, H. Wang, and V. Saligrama, "Don't even look once: Synthesizing features for zero-shot detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11 693–11 702, 2020.

[596] W. Zhu, J. Hessel, A. Awadalla, S. Y. Gadre, J. Dodge, A. Fang, Y. Yu, L. Schmidt, W. Y. Wang, and Y. Choi, "Multimodal c4: An open, billion-scale corpus of images interleaved with text," *arXiv preprint arXiv:2304.06939*, 2023.

[597] Y. Zhu, M. Zhu, N. Liu, Z. Ou, X. Mou, and J. Tang, "Llava-phi: Efficient multi-modal assistant with small language model," *arXiv preprint arXiv:2401.02330*, 2024.

[598] Z. Zong, G. Song, and Y. Liu, *Detrs with collaborative hybrid assignments training*, 2023. arXiv: 2211.12860 [cs.CV].

[599] X. Zou, Z.-Y. Dou, J. Yang, Z. Gan, L. Li, C. Li, X. Dai, H. Behl, J. Wang, L. Yuan, *et al.*, "Generalized decoding for pixel, image, and language," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[600] X. Zou, H. Liu, and Y. J. Lee, "End-to-end instance edge detection," *arXiv preprint arXiv:2204.02898*, 2022.

[601] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Gao, and Y. J. Lee, "Segment everything everywhere all at once," *arXiv preprint arXiv:2304.06718*, 2023.