Vision-Language Pre-Training: Basics, Recent Advances, and Future Trends

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

Video Summarization Overview
Mayu Otani, Yale Song and Yang Wang
ISBN: 978-1-63828-078-1

A Comprehensive Review of Modern Object Segmentation Approaches
Yuanbo Wang, Unaiza Ahsan, Hanyan Li and Matthew Hagen
ISBN: 978-1-63828-070-5

Deep Learning for Image/Video Restoration and Super-resolution
A. Murat Tekalp
ISBN: 978-1-68083-972-2

Deep Learning for Multimedia Forensics
Irene Amerini, Aris Anagnostopoulos, Luca Maiano and Lorenzo Ricciardi Celsi
ISBN: 978-1-68083-854-1

Computer Vision for Autonomous Vehicles: Problems, Datasets and State of the Art
Joel Janai, Fatma Güney, Aseem Behl and Andreas Geiger
ISBN: 978-1-68083-688-2

Discrete Graphical Models - An Optimization Perspective
Bogdan Savchynskyy
ISBN: 978-1-68083-638-7

Vision-Language Pre-Training: Basics, Recent Advances, and Future Trends

Zhe Gan
Microsoft Corporation
pkuganzhe@gmail.com

Linjie Li
Microsoft Corporation

Chunyuan Li
Microsoft Corporation

Lijuan Wang
Microsoft Corporation

Zicheng Liu
Microsoft Corporation

Jianfeng Gao
Microsoft Corporation

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

Z. Gan *et al.*. *Vision-Language Pre-Training: Basics, Recent Advances, and Future Trends*. Foundations and Trends® in Computer Graphics and Vision, vol. 14, no. 3–4, pp. 163–352, 2022.

ISBN: 978-1-63828-133-7

Foundations and Trends® in Computer Graphics and Vision
Volume 14, Issue 3–4, 2022

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, 2022, Volume 14, 4 issues. ISSN paper version 1572-2740. ISSN online version 1572-2759. Also available as a combined paper and online subscription.

Contents

Vision-Language Pre-Training: Basics, Recent Advances, and Future Trends

Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu and Jianfeng Gao

Microsoft Corporation, USA; pkuganzhe@gmail.com

ABSTRACT

This monograph surveys vision-language pre-training (VLP) methods for multimodal intelligence that have been developed in the last few years. We group these approaches into three categories: (i) VLP for image-text tasks, such as image captioning, image-text retrieval, visual question answering, and visual grounding; (ii) VLP for core computer vision tasks, such as (open-set) image classification, object detection, and segmentation; and (iii) VLP for video-text tasks, such as video captioning, video-text retrieval, and video question answering. For each category, we present a comprehensive review of state-of-the-art methods, and discuss the progress that has been made and challenges still being faced, using specific systems and models as case studies. In

Zhe Gan and Jianfeng Gao initiated the project. Zhe Gan and Linjie Li took lead in the writing of Section 1. Linjie Li and Jianfeng Gao took lead in the writing of Section 2. Zhe Gan further took lead in the writing of Sections 3 and 7. Chunyuan Li took lead in the writing of Section 4. Linjie Li further took lead in the writing of Section 5. Lijuan Wang and Zicheng Liu took lead in the writing of Section 6. All the authors provided project advice, and contributed to editing and proofreading.

Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu and Jianfeng Gao (2022), "Vision-Language Pre-Training: Basics, Recent Advances, and Future Trends", Foundations and Trends® in Computer Graphics and Vision: Vol. 14, No. 3–4, pp 163–352. DOI: 10.1561/0600000105.

addition, for each category, we discuss advanced topics being actively explored in the research community, such as big foundation models, unified modeling, in-context few-shot learning, knowledge, robustness, and computer vision in the wild, to name a few.

1

Introduction

Humans perceive the world through many channels, such as images viewed by the eyes, or voices heard by the ears. Though any individual channel might be incomplete or noisy, humans can naturally align and fuse information collected from multiple channels in order to grasp the key concepts needed for a better understanding of the world.

One of the core aspirations in AI is to develop algorithms that endow computers with an ability to effectively learn from multimodal (or, multi-channel) data. This data is similar to sights and sounds attained from *vision* and *language* that help humans make sense of the world around us. For example, computers could mimic this ability by searching the most relevant images to a text query (or vice versa), and by describing the content of an image using natural language.

Vision-and-Language (VL), a popular research area that sits at the nexus of Computer Vision and Natural Language Processing (NLP), aims to achieve this goal. Inspired by the great success of language model pre-training in NLP (*e.g.*, BERT [74], RoBERTa [262], T5 [327], and GPT-3 [33]), Vision-Language Pre-training (VLP) has recently attracted rapidly growing attention from both communities. With the promise to learn universal transferable visual and vision-language representations, VLP has become an increasingly central training paradigm for modern VL research.

Recently, there are some related papers on VLP. For example, [501] focused on task-specific VL methods before the era of pre-training, and provided a concise discussion of VLP models. [85] and [231] focused on VLP, but mainly on image-text tasks, without touch on video-text tasks. [343] focused on VLP for video-text tasks. In [47], the authors reviewed VLP methods for image-text and video-text tasks. However, the discussion is not in depth. The contributions of this monograph are summarized as follows.

- We provide a comprehensive survey on modern VLP, not only covering its successful applications to traditional image-text and video-text tasks (*e.g.*, image/video captioning, retrieval, and question answering), but also showing its great potential for core computer vision tasks (*e.g.*, image classification, object detection and segmentation).

- We provide in-depth discussions on advanced topics at the frontier of VLP, ranging from big foundation models, unified modeling, in-context few-shot learning, knowledge-enhanced VLP, multilingual VLP, model robustness, model compression, to computer vision in the wild.

- We picture the landscape of VL systems developed in research communities and released to the public, demonstrating via case studies the progress we have made and the challenges we are facing.

1.1 Who Should Read this Monograph?

This monograph is based on our CVPR 2022 tutorial,[1] with researchers in the computer vision and NLP communities as our primary target audience. It provides a detailed presentation of the important ideas and insights needed to understand modern VLP methods, and serves as a valuable resource for students, researchers, engineers, and practitioners

[1] https://vlp-tutorial.github.io/.

that are interested in large-scale pre-training for VL representation learning and its applications in computer vision and multimodal tasks. The monograph is structured as follows.

- Section 1 introduces the landscape of VL research, and presents a historical view on the transition of VL research from task-specific methods to large-scale pre-training.
- Section 2 introduces early task-specific VL methods for visual question answering, image captioning, and image-text retrieval, which serve as the foundation to understand modern VLP methods.
- Section 3 describes VLP methods for image-text tasks, such as image captioning, image-text retrieval, visual question answering, and visual grounding.
- Section 4 describes VLP methods for core computer vision tasks, including (open-vocabulary) image classification, object detection and segmentation.
- Section 5 describes VLP methods for video-text tasks, such as video captioning, video-text retrieval, and video question answering.
- Section 6 briefly reviews VL systems developed in industry and the challenges to deploy these VL systems in real-world settings.
- Section 7 concludes the monograph and discusses research trends.

Relations between core sections. Sections 2–5 are the core sections of this monograph. An overview of these sections is provided in Figure 1.1. As the wave of VLP starts with image-text tasks, we first provide a comprehensive review on the transition from early task-specific methods (Section 2) to the most recent VLP methods (Section 3) with image-text inputs. In Section 4, we discuss how core computer vision tasks can be viewed as image-text tasks with open-vocabulary predictions, when powered by contrastively pre-trained image-text models (such as CLIP [326]), and further enable computer vision in the wild [229].

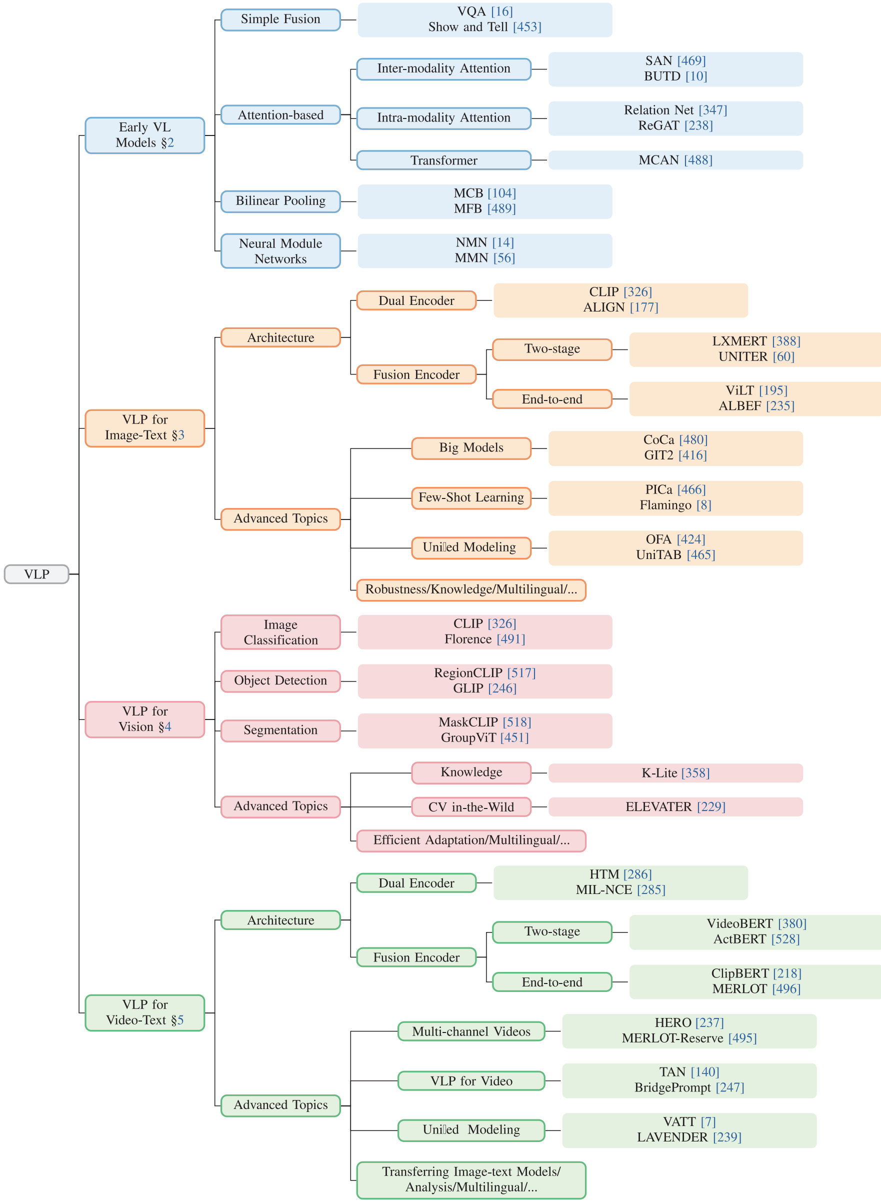

Figure 1.1: Overview of the monograph structure, detailing Sections 2–5.

Extending image-text tasks to more modalities, we present how VLP methods can serve more applications with video-text inputs in Section 5.

How to read the monograph. Readers with different backgrounds may have different purposes for reading this monograph. Below, 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 video-language pre-training, then you can directly jump to Section 5.
- If you are a beginner in the VLP field, and are interested in getting a glimpse of the cutting-edge research of VLP, it is also highly suggested to read the whole monograph section by section, as it provides a comprehensive literature review that helps you understand the VLP landscape.
- If you already have rich experience in VLP and are very familiar with the literature, feel free to jump to specific sections you want to read. In particular, we include in each section a dedicated part in which we discuss advanced topics. For example, in Section 3.5, we discuss big foundation models, unified image-text modeling, in-context few-shot learning, knowledge, robustness and probing analysis, etc.

1.2 Vision-and-Language: What Kinds of Problems?

We live in a multimodal world, and our brains naturally learn to process multi-sense signals received from the environment to help us make sense of the world around us. More specifically, *vision* is a large portion of how humans perceive, while *language* is a large portion of how humans communicate. A multimodal AI system, by its definition, should have the ability to process such multimodal signals effectively and efficiently. Among the ever-growing literature on VL research, in this monograph, we group VL problems into three categories, as detailed below.

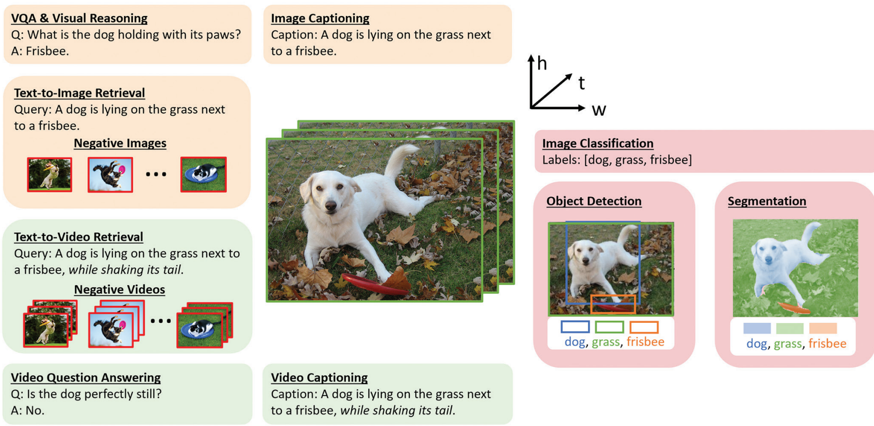

Figure 1.2: Illustration of representative tasks from three categories of VL problems covered in this monograph: image-text tasks , vision tasks as VL problems , and video-text tasks .

- **Image-Text Tasks.** Arguably, the most important and well-studied tasks in VL research are image-text retrieval, image captioning [408], and visual question answering (VQA) [16] (highlighted with orange in Figure 1.2). Centered around these tasks, many related tasks have been proposed and studied.

  – **VQA and visual reasoning.** As extensions to visual question answering, researchers have developed datasets for visual reasoning [170], [378], visual commonsense reasoning [493], visual dialog [70], knowledge-based VQA [283], scene-text-based VQA [370], *etc.* The answers required in these these tasks can be open-ended free-form texts, or selected from multiple choices.

  – **Image captioning.** In addition to the setting where short single-sentence generation is required [254], researchers have also developed datasets for image paragraph captioning [200], scene-text-based image captioning [366], visual storytelling [164], and so on.

  – **Image-text retrieval.** Popular image-text retrieval datasets are based on image captioning datasets [58], [321]. AI models are required to retrieve the most relevant text (or image) from a large corpus, given the image (or text) query.

  – **Visual grounding.** Instead of text outputs, referring expression comprehension and phrase grounding [321], [483] requires bounding box outputs, where the model needs to predict the bounding box corresponding to the input text query.
  – **Text-to-image generation.** It can be considered as the dual task of image captioning, where the system is required to create a high-fidelity image based on the text input. A brief discussion on this task is provided in Section 3.6.

- **Computer Vision Tasks as VL Problems.** Image classification, object detection, and segmentation (highlighted with pink in Figure 1.2) are core visual recognition tasks in computer vision. Traditionally, these tasks are considered as pure vision problems. With the advent of CLIP [326] and ALIGN [177], researchers have realized that language supervision can play an important role in computer vision tasks. First, the use of noisy image-text data crawled from the web allows large-scale pre-training of vision encoders from scratch. Second, instead of treating the supervision signals (*e.g.*, class labels) as one-hot vectors, we take the semantic meaning behind the labels into consideration and cast these computer vision tasks as VL problems. This perspective generalizes the traditional close-set classification or detection models to recognizing unseen concepts in real-world applications, such as open-vocabulary object detection.
- **Video-Text Tasks.** Besides static images, videos are another important type of visual modality. Naturally, all aforementioned image-text tasks have their video-text counterparts, such as video captioning, retrieval, and question answering (highlighted with green in Figure 1.2). The uniqueness of video inputs, in comparison to images, requires an AI system to not only capture spatial information within a single video frame, but also capture the inherent temporal dependencies among video frames.

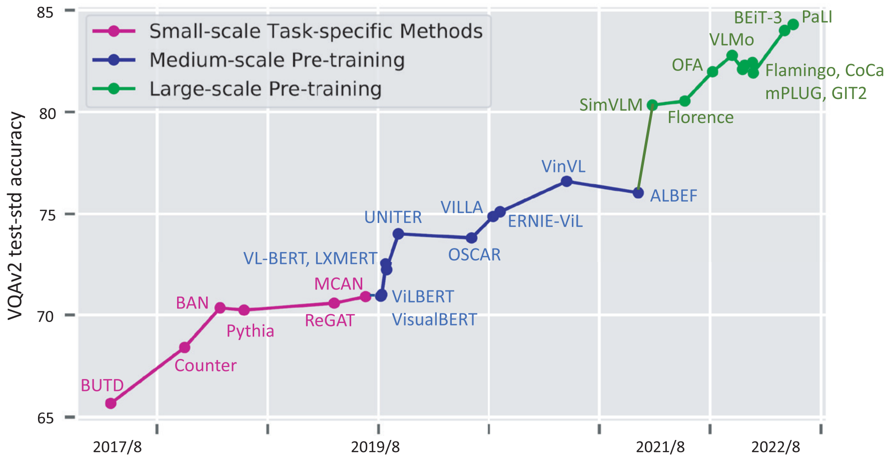

Figure 1.3: The transition from task-specific methods to large-scale pre-training, using the VQA task as a case study. Every time when there was a transition, we observe a big performance lift, *e.g.*, from MCAN [487] to UNITER [60], and from ALBEF [235] to SimVLM [433]. Methods before August 2017 were not drawn; only some representative VLP works are shown to avoid the figure to be too crowded.

While this monograph provides a comprehensive survey of VLP, some of the important VL topics are not discussed. For example, Vision-Language Navigation (VLN) [12], another emerging topic at the intersection of VL research and embodied AI, is not covered in this monograph.

1.3 The Transition From Task-Specific Methods to Large-Scale Pre-training

From a historical perspective, the progress of VL research can be divided into three stages. In Figure 1.3, we use the performance of the popular VQA task to illustrate the research transition from task-specific methods to medium-scale and large-scale pre-training.

- **Small-scale task-specific method design (2014/11–2019/8).** At this stage, many task-specific methods have been developed for image captioning and VQA. For example, an important line of work is to design various attention mechanisms based on pre-extracted visual features (*e.g.*, ResNet [143], Faster RCNN [338], C3D [402]), pre-trained word embeddings (*e.g.*, GLoVe [316], word2vec [288]), and LSTM [152], as we will review in Section 2. These attention method designs have been used to capture multi-modal alignment, perform object relational reasoning, and model multi-step reasoning.

- **Medium-scale pre-training (2019/8–2021/8).** Inspired by the great success of BERT [74] in NLP, the VL field has gradually shifted to using Transformer-based multimodal fusion models that are pre-trained in medium-scale settings, *e.g.*, using image-text datasets up to 4 M images (roughly 10 M image-text pairs in total), with model sizes ranging from 110 M (BERT-base) to 340 M (BERT-large). Typical examples of medium-scale VLP models include UNITER [60] and OSCAR [250], as will be described in Section 3.
- **Large-scale pre-training (2021/8-now).** With the advent of CLIP [326] and ALIGN [177] that aim to train image-text dual encoders from noisy image-text pairs crawled from the web, large-scale VLP shows great promise and is becoming the foundation of VL research. We have witnessed a boom of big multimodal foundation models, *e.g.*, SimVLM [433], Florence [490], Flamingo [8], CoCa [479] and GIT [415]. The high computational cost of VLP can be amortized via adapting the pre-trained models to a wide range of downstream tasks. The number of image-text pairs used for pre-training has increased to over 12B, with model sizes growing to 5B, as in GIT [415]. We provide some detailed discussion on big models in Section 3.5.1.

1.4 What is a Good VLP Model From an Overall Perspective?

While VLP is an emerging field with many new exciting papers appearing, it remains less clear what is the north star we are pursuing as a community. We provide our perspective on the direction. We believe a good VLP model should:

- **Achieve good performance on a wide range of downstream tasks.** The task coverage can be considered in a two-level granularity. First, the problem types are broad, for example, one model can perform on image-text tasks such as VQA, image captioning and text-to-image generation in Section 3, core computer vision tasks such as image classification, object detection and segmentation in Section 4, video-text tasks such as video QA and captioning in Section 5. Second, for each problem type, there is a broad coverage of datasets that represent different use scenarios. For example, in [229] the authors present 20 image classification datasets and 35 object detection datasets to illustrate various scenarios in the wild.

- **Adapt to new tasks with minimal cost.** The adaptation cost needs to be low when deploying a VLP model to a new task. Various efficiency metrics can be considered to measure the adaptation cost, including inference speed, GPU usage for further model weight update, the number of training samples, and the number of trainable parameters. This is an area not well defined yet, and there has been some early effort. For example, in [229] the authors provide a definition by decomposing the adaptation cost into sample-efficiency and parameter-efficiency.

To summarize, the north star of a good VLP model is a single unified model with fixed model weights (or, with inexpensive finetuning) that performs well on all the tasks above. This is an ambitious goal that the community is collectively working towards. Developing a central benchmark is itself an open research problem. We advocate for considering the following factors when benchmarking VLP models: the coverage of tasks, the performance on these tasks, and the cost of adaptation.

1.5 Related Materials: Slide Decks and Pre-recorded Talks

This monograph extends what we present in CVPR tutorials by covering the most recent advances in the field. Below, we provide a list of slide decks and pre-recorded talks that relate to the topics in each section, for references.

- **Section 2**:
  - CVPR 2020 Tutorial: VQA and visual reasoning (Youtube, Bilibili)
  - CVPR 2020 Tutorial: Image captioning (Youtube, Bilibili)
- **Section 3**:
  - CVPR 2022 Tutorial: Overview of Image-Text Pre-training (YouTube, Bilibili)
  - CVPR 2022 Tutorial: Unified Image-Text Modeling (YouTube, Bilibili)
  - CVPR 2022 Tutorial: Advanced Topics in Image-Text Pre-training (YouTube, Bilibili)
  - CVPR 2021 Tutorial: Representations and Training Strategies for VLP (YouTube)
  - CVPR 2021 Tutorial: Robustness, Efficiency and Extensions for VLP (YouTube)
  - CVPR 2020 Tutorial: Self-supervised Image-Text Learning (YouTube, Bilibili)
- **Section 4**:
  - CVPR 2022 Tutorial: VLP for Image Classification (Youtube, Bilibili)
  - CVPR 2022 Tutorial: VLP for Object Detection (Youtube, Bilibili)
  - CVPR 2022 Tutorial: Benchmarks for Computer Vision in the Wild (YouTube, Bilibili)
- **Section 5**:
  - CVPR 2022 Tutorial: Overview of Video-Text Pre-training (YouTube, Bilibili)
  - CVPR 2022 Tutorial: Learning from Multi-channel Videos: Methods and Benchmarks (YouTube, Bilibili)

- CVPR 2022 Tutorial: Advanced Topics in Video-Text Pre-training (YouTube, Bilibili)
- CVPR 2021 Tutorial: Video-and-Language Pre-training (Youtube)

References

[1] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, “Youtube-8m: A large-scale video classification benchmark,” *arXiv preprint arXiv:1609.08675*, 2016.

[2] V. Agarwal, R. Shetty, and M. Fritz, “Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing,” in *CVPR*, 2020.

[3] A. Aghajanyan, B. Huang, C. Ross, V. Karpukhin, H. Xu, N. Goyal, D. Okhonko, M. Joshi, G. Ghosh, M. Lewis, *et al.*, “Cm3: A causal masked multimodal model of the internet,” *arXiv preprint arXiv:2201.07520*, 2022.

[4] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi, “Don’t just assume; look and answer: Overcoming priors for visual question answering,” in *CVPR*, 2018.

[5] A. Agrawal, I. Kajić, E. Bugliarello, E. Davoodi, A. Gergely, P. Blunsom, and A. Nematzadeh, “Rethinking evaluation practices in visual question answering: A case study on out-of-distribution generalization,” *arXiv preprint arXiv:2205.12191*, 2022.

[6] 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.

[7] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong, "Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text," in *NeurIPS*, 2021.

[8] 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.

[9] J.-B. Alayrac, A. Recasens, R. Schneider, R. Arandjelović, J. Ramapuram, J. De Fauw, L. Smaira, S. Dieleman, and A. Zisserman, "Self-supervised multimodal versatile networks," in *NeurIPS*, 2020.

[10] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *ECCV*, 2016.

[11] 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.

[12] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. Van Den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *CVPR*, 2018.

[13] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Learning to compose neural networks for question answering," in *NAACL*, 2016.

[14] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Neural module networks," in *CVPR*, 2016.

[15] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional image captioning," in *CVPR*, 2018.

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

[17] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *ICCV*, 2021.

[18] R. Bachmann, D. Mizrahi, A. Atanov, and A. Zamir, "Multimae: Multi-modal multi-task masked autoencoders," *arXiv preprint arXiv:2204.01678*, 2022.

[19] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “Wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *NeurIPS*, 2020.

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

[21] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, “Frozen in time: A joint video and image encoder for end-to-end retrieval,” in *ICCV*, 2021.

[22] S. Banerjee and A. Lavie, “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments,” in *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005.

[23] H. Bao, L. Dong, and F. Wei, “BEiT: BERT pre-training of image transformers,” in *ICLR*, 2022.

[24] H. Bao, W. Wang, L. Dong, and F. Wei, “VL-BEiT: Generative vision-language pretraining,” *arXiv preprint arXiv:2206.01127*, 2022.

[25] H. Ben-Younes, R. Cadene, M. Cord, and N. Thome, “Mutan: Multimodal tucker fusion for visual question answering,” in *ICCV*, 2017.

[26] H. Ben-Younes, R. Cadene, N. Thome, and M. Cord, “Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection,” in *AAAI*, 2019.

[27] F. Bianchi, G. Attanasio, R. Pisoni, S. Terragni, G. Sarti, and S. Lakshmi, “Contrastive language-image pre-training for the italian language,” *arXiv preprint arXiv:2108.08688*, 2021.

[28] A. Birhane, V. U. Prabhu, and E. Kahembwe, “Multimodal datasets: Misogyny, pornography, and malignant stereotypes,” *arXiv preprint arXiv:2110.01963*, 2021.

[29] A. F. Biten, R. Litman, Y. Xie, S. Appalaraju, and R. Manmatha, “Latr: Layout-aware transformer for scene-text vqa,” in *CVPR*, 2022.

[30] A. F. Biten, R. Tito, A. Mafla, L. Gomez, M. Rusinol, E. Valveny, C. Jawahar, and D. Karatzas, “Scene text visual question answering,” in *CVPR*, 2019.

[31] Y. Bitton, N. B. Guetta, R. Yosef, Y. Elovici, M. Bansal, G. Stanovsky, and R. Schwartz, "WinoGAViL: Gamified association benchmark to challenge vision-and-language models," *arXiv preprint arXiv:2207.12576*, 2022.

[32] L. Breitfeller, E. Ahn, D. Jurgens, and Y. Tsvetkov, "Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts," in *EMNLP*, 2019.

[33] 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.

[34] S. Buch, C. Eyzaguirre, A. Gaidon, J. Wu, L. Fei-Fei, and J. C. Niebles, "Revisiting the 'Video' in video-language understanding," in *CVPR*, 2022.

[35] E. Bugliarello, R. Cotterell, N. Okazaki, and D. Elliott, "Multimodal pretraining unmasked: Unifying the vision and language BERTs," *TACL*, 2021.

[36] R. Cadene, H. Ben-Younes, M. Cord, and N. Thome, "Murel: Multimodal relational reasoning for visual question answering," in *CVPR*, 2019.

[37] R. Cadene, C. Dancette, H. Ben-younes, M. Cord, and D. Parikh, "RUBi: Reducing unimodal biases for visual question answering," in *NeurIPS*, 2019.

[38] 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.

[39] J. Cao, Z. Gan, Y. Cheng, L. Yu, Y.-C. Chen, and J. Liu, "Behind the scene: Revealing the secrets of pre-trained vision-and-language models," in *ECCV*, 2020.

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

[41] F. Carlsson, P. Eisen, F. Rekathati, and M. Sahlgren, "Cross-lingual and multilingual CLIP," in *Proceedings of the Language Resources and Evaluation Conference*, 2022.

[42] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017.

[43] Y. Chang, M. Narang, H. Suzuki, G. Cao, J. Gao, and Y. Bisk, "Webqa: Multihop and multimodal qa," in *CVPR*, 2022.
[44] 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.
[45] B. Chen, A. Rouditchenko, K. Duarte, H. Kuehne, S. Thomas, A. Boggust, R. Panda, B. Kingsbury, R. Feris, D. Harwath, *et al.*, "Multimodal clustering networks for self-supervised learning from unlabeled videos," in *ICCV*, 2021.
[46] D. Chen and W. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *ACL*, 2011.
[47] 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.
[48] L. Chen, Z. Gan, Y. Cheng, L. Li, L. Carin, and J. Liu, "Graph optimal transport for cross-domain alignment," in *ICML*, 2020.
[49] L. Chen, Y. Zhang, R. Zhang, C. Tao, Z. Gan, H. Zhang, B. Li, D. Shen, C. Chen, and L. Carin, "Improving sequence-to-sequence learning via optimal transport," in *ICLR*, 2019.
[50] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever, "Generative pretraining from pixels," in *ICML*, 2020.
[51] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *JSTSP*, 2022.
[52] T. Chen and J. Luo, "Expressing objects just like words: Recurrent visual embedding for image-text matching," in *AAAI*, 2020.
[53] T. Chen, J. Frankle, S. Chang, S. Liu, Y. Zhang, Z. Wang, and M. Carbin, "The lottery ticket hypothesis for pre-trained bert networks," in *NeurIPS*, 2020.
[54] T. Chen, S. Saxena, L. Li, D. J. Fleet, and G. Hinton, "Pix2seq: A language modeling framework for object detection," in *ICLR*, 2022.

[55] 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.

[56] W. Chen, Z. Gan, L. Li, Y. Cheng, W. Wang, and J. Liu, "Meta module network for compositional visual reasoning," in *WACV*, 2021.

[57] 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.

[58] 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.

[59] X. Chen, L. Ma, J. Chen, Z. Jie, W. Liu, and J. Luo, "Real-time referring expression comprehension by single-stage grounding network," *arXiv preprint arXiv:1812.03426*, 2018.

[60] 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.

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

[62] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *EMNLP*, 2014.

[63] S. Choi, K.-W. On, Y.-J. Heo, A. Seo, Y. Jang, M. Lee, and B.-T. Zhang, "DramaQA: Character-centered video story understanding with hierarchical qa," in *AAAI*, 2021.

[64] 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.

[65] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," in *ICLR*, 2020.

[66] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *CVPR*, 2020.

[67] Y. Cui, Z. Yu, C. Wang, Z. Zhao, J. Zhang, M. Wang, and J. Yu, "ROSITA: Enhancing vision-and-language semantic alignments via cross-and intra-modal knowledge integration," in *ACMMM*, 2021.

[68] Y. Dai, D. Tang, L. Liu, M. Tan, C. Zhou, J. Wang, Z. Feng, F. Zhang, X. Hu, and S. Shi, "One model, multiple modalities: A sparsely activated approach for text, sound, image, video and code," *arXiv preprint arXiv:2205.06126*, 2022.

[69] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, *et al.*, "Scaling egocentric vision: The epic-kitchens dataset," in *ECCV*, 2018.

[70] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra, "Visual dialog," in *CVPR*, 2017.

[71] 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.

[72] J. Deng, Z. Yang, T. Chen, W. Zhou, and H. Li, "Transvg: End-to-end visual grounding with transformers," in *ICCV*, 2021.

[73] 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.

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

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

[76] H. Diao, Y. Zhang, L. Ma, and H. Lu, "Similarity reasoning and filtration for image-text matching," in *AAAI*, 2021.

[77] S. Diao, W. Zhou, X. Zhang, and J. Wang, "Prefix language models are unified modal learners," *arXiv preprint arXiv:2206.07699*, 2022.

[78] M. Ding, Z. Yang, W. Hong, W. Zheng, C. Zhou, D. Yin, J. Lin, X. Zou, Z. Shao, H. Yang, *et al.*, "Cogview: Mastering text-to-image generation via transformers," in *NeurIPS*, 2021.
[79] M. Ding, W. Zheng, W. Hong, and J. Tang, "Cogview2: Faster and better text-to-image generation via hierarchical transformers," *arXiv preprint arXiv:2204.14217*, 2022.
[80] Z. Ding, J. Wang, and Z. Tu, "Open-vocabulary panoptic segmentation with MaskCLIP," *arXiv preprint arXiv:2208.08984*, 2022.
[81] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *CVPR*, 2015.
[82] 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 16×16 words: Transformers for image recognition at scale," in *ICLR*, 2021.
[83] Z.-Y. Dou, A. Kamath, Z. Gan, P. Zhang, J. Wang, L. Li, Z. Liu, C. Liu, Y. LeCun, N. Peng, *et al.*, "Coarse-to-fine vision-language pre-training with fusion in the backbone," in *NeurIPS*, 2022.
[84] 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.
[85] Y. Du, Z. Liu, J. Li, and W. X. Zhao, "A survey of vision-language pre-trained models," in *IJCAI Survey Track*, 2022.
[86] J. Duan, L. Chen, S. Tran, J. Yang, Y. Xu, B. Zeng, and T. Chilimbi, "Multi-modal alignment using representation codebook," in *CVPR*, 2022.
[87] P. Esser, R. Rombach, A. Blattmann, and B. Ommer, "Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis," in *NeurIPS*, 2021.
[88] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *CVPR*, 2021.
[89] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: Improving visual-semantic embeddings with hard negatives," *arXiv preprint arXiv:1707.05612*, 2017.

[90] A. Fang, G. Ilharco, M. Wortsman, Y. Wan, V. Shankar, A. Dave, and L. Schmidt, "Data determines distributional robustness in contrastive language image pre-training (clip)," *arXiv preprint arXiv:2205.01397*, 2022.

[91] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, *et al.*, "From captions to visual concepts and back," in *CVPR*, 2015.

[92] 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.

[93] Z. Fang, J. Wang, X. Hu, L. Wang, Y. Yang, and Z. Liu, "Compressing visual-linguistic model via knowledge distillation," in *ICCV*, 2021.

[94] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *CVPR*, 2009.

[95] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in *ECCV*, 2010.

[96] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *ICCV*, 2019.

[97] C. Feng, Y. Zhong, Z. Jie, X. Chu, H. Ren, X. Wei, W. Xie, and L. Ma, "Promptdet: Expand your detector vocabulary with uncurated images," in *ECCV*, 2022.

[98] S. Frank, E. Bugliarello, and D. Elliott, "Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers," *arXiv preprint arXiv:2109.04448*, 2021.

[99] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," in *ICML*, 2019.

[100] 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.

[101] T.-J. Fu, L. Li, Z. Gan, K. Lin, W. Y. Wang, L. Wang, and Z. Liu, "VIOLET: End-to-end video-language transformers with masked visual-token modeling," *arXiv preprint arXiv:2111.12681*, 2021.

[102] T.-J. Fu, L. Li, Z. Gan, K. Lin, W. Y. Wang, L. Wang, and Z. Liu, "An empirical study of end-to-end video-language transformers with masked visual modeling," *arXiv preprint arXiv:2209.01540*, 2022.

[103] Y. Fu and L. Sigal, "Semi-supervised vocabulary-informed learning," in *CVPR*, 2016.

[104] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *EMNLP*, 2016.

[105] V. Gabeur, C. Sun, K. Alahari, and C. Schmid, "Multi-modal transformer for video retrieval," in *ECCV*, 2020.

[106] 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.

[107] Z. Gan, Y.-C. Chen, L. Li, T. Chen, Y. Cheng, S. Wang, and J. Liu, "Playing lottery tickets with vision and language," in *AAAI*, 2022.

[108] 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.

[109] Z. Gan, Y. Cheng, A. E. Kholy, L. Li, J. Liu, and J. Gao, "Multi-step reasoning via recurrent dual attention for visual dialog," in *ACL*, 2019.

[110] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu, "Are you talking to a machine? dataset and methods for multilingual image question," in *NeurIPS*, 2015.

[111] J. Gao, M. Galley, L. Li, *et al.*, "Neural approaches to conversational AI," in *Foundations and Trends® in Information Retrieval*, 2019.

[112] J. Gao, C. Xiong, P. Bennett, and N. Craswell, "Neural approaches to conversational information retrieval," *arXiv preprint arXiv:2201.05176*, 2022.

[113] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "CLIP-adapter: Better vision-language models with feature adapters," *arXiv preprint arXiv:2110.04544*, 2021.

[114] P. Gao, Z. Jiang, H. You, P. Lu, S. C. Hoi, X. Wang, and H. Li, "Dynamic fusion with intra-and inter-modality attention flow for visual question answering," in *CVPR*, 2019.
[115] Y. Gao, J. Liu, Z. Xu, J. Zhang, K. Li, and C. Shen, "Pyramidclip: Hierarchical feature alignment for vision-language model pretraining," *arXiv preprint arXiv:2204.14095*, 2022.
[116] Y. Ge, Y. Ge, X. Liu, D. Li, Y. Shan, X. Qie, and P. Luo, "Bridging video-text retrieval with multiple choice questions," in *CVPR*, 2022.
[117] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford, "Datasheets for datasets," *CACM*, 2021.
[118] X. Geng, H. Liu, L. Lee, D. Schuurams, S. Levine, and P. Abbeel, "Multimodal masked autoencoders learn transferable representations," *arXiv preprint arXiv:2205.14204*, 2022.
[119] G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin, "Open-vocabulary image segmentation," in *ECCV*, 2022.
[120] R. Girshick, "Fast r-cnn," in *ICCV*, 2015.
[121] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014.
[122] S. Goel, H. Bansal, S. Bhatia, R. A. Rossi, V. Vinay, and A. Grover, "Cyclip: Cyclic contrastive language-image pretraining," *arXiv preprint arXiv:2205.14459*, 2022.
[123] S. Goenka, Z. Zheng, A. Jaiswal, R. Chada, Y. Wu, V. Hedau, and P. Natarajan, "FashionVLP: Vision language transformer for fashion retrieval with feedback," in *CVPR*, 2022.
[124] T. Gokhale, P. Banerjee, C. Baral, and Y. Yang, "Vqa-lol: Visual question answering under the lens of logic," in *ECCV*, 2020.
[125] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, *et al.*, "The 'something something' video database for learning and evaluating visual common sense," in *ICCV*, 2017.
[126] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering," in *CVPR*, 2017.

[127] T. Grigoryev, A. Voynov, and A. Babenko, "When, why, and which pretrained GANs are useful?" In *ICLR*, 2022.
[128] J. Gu, X. Meng, G. Lu, L. Hou, M. Niu, H. Xu, X. Liang, W. Zhang, X. Jiang, and C. Xu, "Wukong: 100 million large-scale chinese cross-modal pre-training dataset and a foundation framework," *arXiv preprint arXiv:2202.06767*, 2022.
[129] J. Gu, E. Stefani, Q. Wu, J. Thomason, and X. E. Wang, "Vision-and-language navigation: A survey of tasks, methods, and future directions," *arXiv preprint arXiv:2203.12667*, 2022.
[130] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, and B. Guo, "Vector quantized diffusion model for text-to-image synthesis," in *CVPR*, 2022.
[131] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," in *ICLR*, 2022.
[132] 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.
[133] A. Gupta, P. Dollar, and R. Girshick, "Lvis: A dataset for large vocabulary instance segmentation," in *CVPR*, 2019.
[134] 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.
[135] T. Gupta, R. Marten, A. Kembhavi, and D. Hoiem, "GRIT: General robust image task benchmark," *arXiv preprint arXiv:2204.13653*, 2022.
[136] 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.
[137] D. Gurari, Y. Zhao, M. Zhang, and N. Bhattacharya, "Captioning images taken by people who are blind," in *ECCV*, 2020.
[138] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *AISTATS*, 2010.
[139] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "Retrieval augmented language model pre-training," in *ICML*, 2020.

[140] T. Han, W. Xie, and A. Zisserman, "Temporal alignment networks for long-term video," in *CVPR*, 2022.
[141] W. Hao, C. Li, X. Li, L. Carin, and J. Gao, "Towards learning a generic agent for vision-and-language navigation via pre-training," in *CVPR*, 2020.
[142] Y. Hao, H. Song, L. Dong, S. Huang, Z. Chi, W. Wang, S. Ma, and F. Wei, "Language models are general-purpose interfaces," *arXiv preprint arXiv:2206.06336*, 2022.
[143] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
[144] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced bert with disentangled attention," in *ICLR*, 2021.
[145] X. He, C. Li, P. Zhang, J. Yang, and X. E. Wang, "Parameter-efficient fine-tuning for vision transformers," *arXiv preprint arXiv:2203.16329*, 2022.
[146] L. A. Hendricks, J. Mellor, R. Schneider, J.-B. Alayrac, and A. Nematzadeh, "Decoupling the role of data, attention, and losses in multimodal transformers," *TACL*, 2021.
[147] L. A. Hendricks and A. Nematzadeh, "Probing image-language transformers for verb understanding," *arXiv preprint arXiv:2106.09141*, 2021.
[148] L. A. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell, "Localizing moments in video with natural language," in *ICCV*, 2017.
[149] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," in *NeurIPS*, 2019.
[150] 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.
[151] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *NeurIPS*, 2020.
[152] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, 1997.

[153] 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.

[154] Y. Hong, Q. Wu, Y. Qi, C. Rodriguez-Opazo, and S. Gould, "Vln bert: A recurrent vision-and-language bert for navigation," in *CVPR*, 2021.

[155] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for NLP," in *ICML*, 2019.

[156] 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.

[157] R. Hu, J. Andreas, T. Darrell, and K. Saenko, "Explainable neural computation via stack neural module networks," in *ECCV*, 2018.

[158] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko, "Learning to reason: End-to-end module networks for visual question answering," in *ICCV*, 2017.

[159] R. Hu, A. Rohrbach, T. Darrell, and K. Saenko, "Language-conditioned graph networks for relational reasoning," in *ICCV*, 2019.

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

[161] X. Hu, Z. Gan, J. Wang, Z. Yang, Z. Liu, Y. Lu, and L. Wang, "Scaling up vision-language pre-training for image captioning," in *CVPR*, 2022.

[162] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *ICCV*, 2019.

[163] L. Huang, G. Niu, J. Liu, X. Xiao, and H. Wu, "DU-VLG: Unifying vision-and-language generation via dual sequence-to-sequence pre-training," in *Findings of ACL*, 2022.

[164] T.-H. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. Girshick, X. He, P. Kohli, D. Batra, *et al.*, "Visual storytelling," in *NAACL*, 2016.

[165] Y. Huang, W. Wang, and L. Wang, "Instance-aware image and sentence matching with selective multimodal lstm," in *CVPR*, 2017.

[166] P.-Y. Huang, M. Patrick, J. Hu, G. Neubig, F. Metze, and A. Hauptmann, "Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models," in *NAACL*, 2021.

[167] 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.

[168] 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.

[169] D. A. Hudson and C. D. Manning, "Compositional attention networks for machine reasoning," in *ICLR*, 2018.

[170] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *CVPR*, 2019.

[171] D. Hudson and C. D. Manning, "Learning by abstraction: The neural state machine," in *NeurIPS*, 2019.

[172] Y. Huo, M. Zhang, G. Liu, H. Lu, Y. Gao, G. Yang, J. Wen, H. Zhang, B. Xu, W. Zheng, *et al.*, "Wenlan: Bridging vision and language by large-scale multi-modal pre-training," *arXiv preprint arXiv:2103.06561*, 2021.

[173] 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*, 2021.

[174] A. Jabri, A. Joulin, and L. v. d. Maaten, "Revisiting visual question answering baselines," in *ECCV*, 2016.

[175] A. Jain, M. Guo, K. Srinivasan, T. Chen, S. Kudugunta, C. Jia, Y. Yang, and J. Baldridge, "Mural: Multimodal, multitask retrieval across languages," *arXiv preprint arXiv:2109.05125*, 2021.

[176] Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. Kim, "TGIF-QA: Toward spatio-temporal reasoning in visual question answering," in *CVPR*, 2017.

[177] 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.

[178] H. Jiang, I. Misra, M. Rohrbach, E. Learned-Miller, and X. Chen, "In defense of grid features for visual question answering," in *CVPR*, 2020.

[179] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "Tinybert: Distilling bert for natural language understanding," in *EMNLP*, 2020.

[180] C. E. Jimenez, O. Russakovsky, and K. Narasimhan, "CARETS: A consistency and robustness evaluative test suite for VQA," *arXiv preprint arXiv:2203.07613*, 2022.

[181] W. Jin, Y. Cheng, Y. Shen, W. Chen, and X. Ren, "A good prompt is worth millions of parameters? low-resource prompt-based learning for vision-language models," in *ACL*, 2022.

[182] J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," in *CVPR*, 2017.

[183] J. Johnson, B. Hariharan, L. Van Der Maaten, J. Hoffman, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "Inferring and executing programs for visual reasoning," in *ICCV*, 2017.

[184] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu, "Exploring the limits of language modeling," *arXiv preprint arXiv:1602.02410*, 2016.

[185] C. Ju, T. Han, K. Zheng, Y. Zhang, and W. Xie, "Prompting visual-language models for efficient video understanding," in *ECCV*, 2022.

[186] 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.

[187] A. Kamath, C. Clark, T. Gupta, E. Kolve, D. Hoiem, and A. Kembhavi, "Webly supervised concept expansion for general purpose vision models," *arXiv preprint arXiv:2202.02317*, 2022.

[188] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015.

[189] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[190] C. Kervadec, G. Antipov, M. Baccouche, and C. Wolf, "Roses are red, violets are blue... but should vqa expect them to?" In *CVPR*, 2021.

[191] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," in *NeurIPS*, 2018.

[192] J.-H. Kim, S.-W. Lee, D. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang, "Multimodal residual learning for visual qa," in *NeurIPS*, 2016.

[193] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," in *ICLR*, 2017.

[194] T. Kim, G. Song, S. Lee, S. Kim, Y. Seo, S. Lee, S. H. Kim, H. Lee, and K. Bae, "L-verse: Bidirectional generation between image and text," in *CVPR*, 2022.

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

[196] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," in *NeurIPS Deep Learning Workshop*, 2014.

[197] B. Klein, G. Lev, G. Sadeh, and L. Wolf, "Associating neural word embeddings with deep image representations using fisher vectors," in *CVPR*, 2015.

[198] B. Ko and G. Gu, "Large-scale bilingual language-image contrastive learning," *arXiv preprint arXiv:2203.14463*, 2022.

[199] 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.

[200] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei, "A hierarchical approach for generating descriptive image paragraphs," in *CVPR*, 2017.

[201] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles, "Dense-captioning events in videos," in *ICCV*, 2017.

[202] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-captioning events in videos," in *ICCV*, 2017.

[203] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *IJCV*, 2017.

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

[205] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *ICCV*, 2011.

[206] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Babytalk: Understanding and generating simple image descriptions," *TPAMI*, 2013.

[207] A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang, "Fine-tuning can distort pretrained features and underperform out-of-distribution," *arXiv preprint arXiv:2202.10054*, 2022.

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

[209] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, *et al.*, "The open images dataset v4," *IJCV*, 2020.

[210] G. KV and A. Mittal, "Reducing language biases in visual question answering with visually-grounded question encoder," in *ECCV*, 2020.

[211] G. Kwon, Z. Cai, A. Ravichandran, E. Bas, R. Bhotika, and S. Soatto, "Masked vision and language modeling for multi-modal representation learning," *arXiv preprint arXiv:2208.02131*, 2022.

[212] C. H. Lampert, H. Nickisch, and S. Harmeling, “Attribute-based classification for zero-shot visual object categorization,” *TPAMI*, 2013.

[213] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “Albert: A lite bert for self-supervised learning of language representations,” in *ICLR*, 2020.

[214] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, “Stacked cross attention for image-text matching,” in *ECCV*, 2018.

[215] C. Lei, S. Luo, Y. Liu, W. He, J. Wang, G. Wang, H. Tang, C. Miao, and H. Li, “Understanding chinese video and language via contrastive multimodal pre-training,” in *ACMMM*, 2021.

[216] J. Lei, T. L. Berg, and M. Bansal, “Revealing single frame bias for video-and-language learning,” *arXiv preprint arXiv:2206.03428*, 2022.

[217] J. Lei, X. Chen, N. Zhang, M. Wang, M. Bansal, T. L. Berg, and L. Yu, “Loopitr: Combining dual and cross encoder architectures for image-text retrieval,” *arXiv preprint arXiv:2203.05465*, 2022.

[218] J. Lei, L. Li, L. Zhou, Z. Gan, T. L. Berg, M. Bansal, and J. Liu, “Less is more: Clipbert for video-and-language learning via sparse sampling,” in *CVPR*, 2021.

[219] J. Lei, L. Yu, M. Bansal, and T. L. Berg, “TVQA: Localized, compositional video question answering,” in *EMNLP*, 2018.

[220] J. Lei, L. Yu, T. L. Berg, and M. Bansal, “TVQA+: Spatio-temporal grounding for video question answering,” in *ACL*, 2020.

[221] J. Lei, L. Yu, T. L. Berg, and M. Bansal, “Tvr: A large-scale dataset for video-subtitle moment retrieval,” in *ECCV*, 2020.

[222] J. Lei, L. Yu, T. L. Berg, and M. Bansal, “What is more likely to happen next? video-and-language future event prediction,” in *EMNLP*, 2020.

[223] B. Lester, R. Al-Rfou, and N. Constant, “The power of scale for parameter-efficient prompt tuning,” in *EMNLP*, 2021.

[224] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, “Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” in *ACL*, 2020.

[225] 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.

[226] B. Li, X. Qi, T. Lukasiewicz, and P. Torr, "Controllable text-to-image generation," in *NeurIPS*, 2019.

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

[228] C. Li, H. Xu, J. Tian, W. Wang, M. Yan, B. Bi, J. Ye, H. Chen, G. Xu, Z. Cao, *et al.*, "mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections," *arXiv preprint arXiv:2205.12005*, 2022.

[229] 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.

[230] D. Li, J. Li, H. Li, J. C. Niebles, and S. C. Hoi, "Align and prompt: Video-and-language pre-training with entity prompts," in *CVPR*, 2022.

[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] G. Li, L. Zhu, P. Liu, and Y. Yang, "Entangled transformer for image captioning," in *ICCV*, 2019.

[234] 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.

[235] 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.

[236] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Visual semantic reasoning for image-text matching," in *ICCV*, 2019.

[237] L. Li, Y.-C. Chen, Y. Cheng, Z. Gan, L. Yu, and J. Liu, "Hero: Hierarchical encoder for video+ language omni-representation pre-training," in *EMNLP*, 2020.
[238] L. Li, Z. Gan, Y. Cheng, and J. Liu, "Relation-aware graph attention network for visual question answering," in *ICCV*, 2019.
[239] L. Li, Z. Gan, K. Lin, C.-C. Lin, Z. Liu, C. Liu, and L. Wang, "LAVENDER: Unifying video-language understanding as masked language modeling," *arXiv preprint arXiv:2206.07160*, 2022.
[240] L. Li, Z. Gan, and J. Liu, "A closer look at the robustness of vision-and-language pre-trained models," *arXiv preprint arXiv: 2012.08673*, 2020.
[241] L. Li, J. Lei, Z. Gan, and J. Liu, "Adversarial vqa: A new benchmark for evaluating the robustness of vqa models," in *ICCV*, 2021.
[242] L. Li, J. Lei, Z. Gan, L. Yu, Y.-C. Chen, R. Pillai, Y. Cheng, L. Zhou, X. E. Wang, W. Y. Wang, *et al.*, "Value: A multi-task benchmark for video-and-language understanding evaluation," in *NeurIPS, Track on Datasets and Benchmarks*, 2021.
[243] 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.
[244] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "What does BERT with vision look at?" In *ACL*, 2020.
[245] L. H. Li, H. You, Z. Wang, A. Zareian, S.-F. Chang, and K.-W. Chang, "Unsupervised vision-and-language pre-training without parallel images and captions," in *NAACL*, 2021.
[246] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, *et al.*, "Grounded language-image pre-training," in *CVPR*, 2022.
[247] M. Li, L. Chen, Y. Duan, Z. Hu, J. Feng, J. Zhou, and J. Lu, "Bridge-prompt: Towards ordinal action understanding in instructional videos," in *CVPR*, 2022.
[248] W. Li, C. Gao, G. Niu, X. Xiao, H. Liu, J. Liu, H. Wu, and H. Wang, "Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning," in *ACL*, 2021.

[249] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *ACL*, 2021.

[250] 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.

[251] 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.

[252] Y. Liao, S. Liu, G. Li, F. Wang, Y. Chen, C. Qian, and B. Li, "A real-time cross-modality correlation filtering method for referring expression comprehension," in *CVPR*, 2020.

[253] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, 2004.

[254] 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.

[255] B. Liu, Z. Huang, Z. Zeng, Z. Chen, and J. Fu, "VQA challenge 2019 runner-up talk," *VQA Challenge 2019*, 2019.

[256] C. Liu, Z. Mao, A.-A. Liu, T. Zhang, B. Wang, and Y. Zhang, "Focus your attention: A bidirectional focal attention network for image-text matching," in *ACMMM*, 2019.

[257] C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, and Y. Zhang, "Graph structured network for image-text matching," in *CVPR*, 2020.

[258] J. Liu, W. Chen, Y. Cheng, Z. Gan, L. Yu, Y. Yang, and J. Liu, "Violin: A large-scale dataset for video-and-language inference," in *CVPR*, 2020.

[259] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *arXiv preprint arXiv:2107.13586*, 2021.

[260] S. Liu, H. Fan, S. Qian, Y. Chen, W. Ding, and Z. Wang, "Hit: Hierarchical transformer with momentum contrast for video-text retrieval," in *ICCV*, 2021.

[261] X. Liu, P. He, W. Chen, and J. Gao, "Multi-task deep neural networks for natural language understanding," in *ACL*, 2019.
[262] 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.
[263] 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.
[264] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *CVPR*, 2022.
[265] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2018.
[266] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *NeurIPS*, 2019.
[267] 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.
[268] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *CVPR*, 2017.
[269] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *NeurIPS*, 2016.
[270] K. Lu, A. Grover, P. Abbeel, and I. Mordatch, "Frozen pretrained transformers as universal computation engines," in *AAAI*, 2022.
[271] Y. Lu, W. Zhu, X. E. Wang, M. Eckstein, and W. Y. Wang, "Imagination-augmented natural language understanding," in *NAACL*, 2022.
[272] T. Lüddecke and A. Ecker, "Image segmentation using text and image prompts," in *CVPR*, 2022.
[273] H. Luo, L. Ji, B. Shi, H. Huang, N. Duan, T. Li, J. Li, T. Bharti, and M. Zhou, "Univl: A unified video and language pre-training model for multimodal understanding and generation," *arXiv preprint arXiv:2002.06353*, 2020.

[274] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li, "Clip4clip: An empirical study of clip for end to end video clip retrieval," *arXiv preprint arXiv:2104.08860*, 2021.

[275] Y. Luo, J. Ji, X. Sun, L. Cao, Y. Wu, F. Huang, C.-W. Lin, and R. Ji, "Dual-level collaborative transformer for image captioning," in *AAAI*, 2021.

[276] L. Ma, Z. Lu, and H. Li, "Learning to answer questions from image using convolutional neural network," in *AAAI*, 2016.

[277] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A neural-based approach to answering questions about images," in *ICCV*, 2015.

[278] E. Mansimov, E. Parisotto, J. L. Ba, and R. Salakhutdinov, "Generating images from captions with attention," in *ICLR*, 2016.

[279] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu, "The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision," in *ICLR*, 2019.

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

[281] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," *arXiv preprint arXiv:1412.6632*, 2014.

[282] 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.

[283] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "Ok-vqa: A visual question answering benchmark requiring external knowledge," in *CVPR*, 2019.

[284] Microsoft, "Responsible bots: 10 guidelines for developers of conversational AI," 2018.

[285] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman, "End-to-end learning of visual representations from uncurated instructional videos," in *CVPR*, 2020.

[286] 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.

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

[288] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *NeurIPS*, 2013.

[289] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, “Deep learning–based text classification: A comprehensive review,” *ACM Computing Surveys (CSUR)*, 2021.

[290] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen, *et al.*, “Simple open-vocabulary object detection with vision transformers,” in *ECCV*, 2022.

[291] A. Mishra, S. Shekhar, A. K. Singh, and A. Chakraborty, “OCR-VQA: Visual question answering by reading text in images,” in *ICDAR*, 2019.

[292] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, “Model cards for model reporting,” in *ACM FAccT*, 2019.

[293] N. Mu, A. Kirillov, D. Wagner, and S. Xie, “SLIP: Self-supervision meets language-image pre-training,” *arXiv preprint arXiv:2112.12750*, 2021.

[294] V. Murahari, D. Batra, D. Parikh, and A. Das, “Large-scale pretraining for visual dialog: A simple state-of-the-art baseline,” in *ECCV*, 2020.

[295] B. Mustafa, C. Riquelme, J. Puigcerver, R. Jenatton, and N. Houlsby, “Multimodal contrastive learning with limoe: The language-image mixture of experts,” *arXiv preprint arXiv:2206.02770*, 2022.

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

[297] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *CVPR*, 2017.

[298] D.-K. Nguyen and T. Okatani, "Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering," in *CVPR*, 2018.

[299] B. Ni, H. Peng, M. Chen, S. Zhang, G. Meng, J. Fu, S. Xiang, and H. Ling, "Expanding language-image pretrained models for general video recognition," *arXiv preprint arXiv:2208.02816*, 2022.

[300] M. Ni, H. Huang, L. Su, E. Cui, T. Bharti, L. Wang, J. Gao, D. Zhang, and N. Duan, "M3p: Learning universal representations via multitask multilingual multimodal pre-training," in *CVPR*, 2021.

[301] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.

[302] Y. Nie, L. Li, Z. Gan, S. Wang, C. Zhu, M. Zeng, Z. Liu, M. Bansal, and L. Wang, "Mlp architectures for vision-and-language modeling: An empirical study," *arXiv preprint arXiv:2112.04453*, 2021.

[303] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela, "Adversarial NLI: A new benchmark for natural language understanding," in *ACL*, 2020.

[304] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Hierarchical multimodal lstm for dense visual-semantic embedding," in *ICCV*, 2017.

[305] W. Norcliffe-Brown, S. Vafeias, and S. Parisot, "Learning conditioned graph structures for interpretable visual question answering," in *NeurIPS*, 2018.

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

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

[308] Y. Pan, Y. Li, J. Luo, J. Xu, T. Yao, and T. Mei, "Auto-captions on GIF: A large-scale video-sentence dataset for vision-language pre-training," *arXiv preprint arXiv:2007.02375*, 2020.

[309] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *CVPR*, 2020.

[310] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *ACL*, 2002.

[311] L. Parcalabescu, M. Cafagna, L. Muradjan, A. Frank, I. Calixto, and A. Gatt, "VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena," *arXiv preprint arXiv:2112.07566*, 2021.

[312] L. Parcalabescu, A. Gatt, A. Frank, and I. Calixto, "Seeing past words: Testing the cross-modal capabilities of pretrained v and l models on counting tasks," *arXiv preprint arXiv:2012.12352*, 2020.

[313] J. S. Park, S. Shen, A. Farhadi, T. Darrell, Y. Choi, and A. Rohrbach, "Exposing the limits of video-text models through contrast sets," in *NAACL*, 2022.

[314] M. Patrick, P.-Y. Huang, Y. Asano, F. Metze, A. G. Hauptmann, J. F. Henriques, and A. Vedaldi, "Support-set bottlenecks for video-text representation learning," in *ICLR*, 2020.

[315] G. Patterson and J. Hays, "SUN attribute database: Discovering, annotating, and recognizing scene attributes," in *CVPR*, 2012.

[316] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014.

[317] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *CVPR*, 2016.

[318] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *AAAI*, 2018.

[319] 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.

[320] B. A. Plummer, P. Kordas, M. H. Kiapour, S. Zheng, R. Piramuthu, and S. Lazebnik, "Conditional image-text embedding networks," in *ECCV*, 2018.
[321] 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.
[322] J. Pont-Tuset, J. Uijlings, S. Changpinyo, R. Soricut, and V. Ferrari, "Connecting vision and language with localized narratives," in *ECCV*, 2020.
[323] M. Pushkarna, A. Zaldivar, and O. Kjartansson, "Data cards: Purposeful and transparent dataset documentation for responsible AI," in *ACM FAccT*, 2022.
[324] 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.
[325] T. Qiao, J. Zhang, D. Xu, and D. Tao, "Mirrorgan: Learning text-to-image generation by redescription," in *CVPR*, 2019.
[326] 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.
[327] 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.
[328] 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.
[329] 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.
[330] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *ICCV*, 2021.

[331] 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," *arXiv preprint arXiv:2112.01518*, 2021.

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

[333] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[334] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *ICML*, 2016.

[335] 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.

[336] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal, "Grounding action descriptions in videos," *TACL*, 2013.

[337] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," in *NeurIPS*, 2015.

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

[339] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele, "A dataset for movie description," in *CVPR*, 2015.

[340] M. Rohrbach, M. Stark, and B. Schiele, "Evaluating knowledge transfer and zero-shot learning in a large-scale setting," in *CVPR*, 2011.

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

[342] A. Rouditchenko, A. Boggust, D. Harwath, B. Chen, D. Joshi, S. Thomas, K. Audhkhasi, H. Kuehne, R. Panda, R. Feris, *et al.*, "Avlnet: Learning audio-visual language representations from instructional videos," in *InterSpeech*, 2021.

[343] L. Ruan and Q. Jin, “Survey: Transformer based video-language pre-training,” *AI Open*, 2022.
[344] 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.
[345] K. Saito, K. Sohn, X. Zhang, C.-L. Li, C.-Y. Lee, K. Saenko, and T. Pfister, “Prefix conditioning unifies language and label supervision,” *arXiv preprint arXiv:2206.01125*, 2022.
[346] E. Salin, B. Farah, S. Ayache, and B. Favre, “Are vision-language transformers learning multimodal representations? a probing perspective,” in *AAAI*, 2022.
[347] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap, “A simple neural network module for relational reasoning,” in *NeurIPS*, 2017.
[348] 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.
[349] 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.
[350] R. R. Selvaraju, P. Tendulkar, D. Parikh, E. Horvitz, M. T. Ribeiro, B. Nushi, and E. Kamar, “Squinting at vqa models: Introspecting vqa models with sub-questions,” in *CVPR*, 2020.
[351] R. Sennrich, B. Haddow, and A. Birch, “Neural machine translation of rare words with subword units,” in *ACL*, 2016.
[352] P. H. Seo, A. Nagrani, A. Arnab, and C. Schmid, “End-to-end generative pretraining for multimodal video captioning,” in *CVPR*, 2022.
[353] M. Shah, X. Chen, M. Rohrbach, and D. Parikh, “Cycle-consistency for robust visual question answering,” in *CVPR*, 2019.
[354] S. Shah, A. Mishra, N. Yadati, and P. P. Talukdar, “Kvqa: Knowledge-aware visual question answering,” in *AAAI*, 2019.

[355] 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.

[356] 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.

[357] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *ICLR*, 2017.

[358] S. Shen, C. Li, X. Hu, Y. Xie, J. Yang, P. Zhang, A. Rohrbach, Z. Gan, L. Wang, L. Yuan, *et al.*, "K-lite: Learning transferable visual models with external knowledge," in *NeurIPS*, 2022.

[359] 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.

[360] S. Sheng, A. Singh, V. Goswami, J. Magana, T. Thrush, W. Galuba, D. Parikh, and D. Kiela, "Human-adversarial visual question answering," in *NeurIPS*, 2021.

[361] V. Shevchenko, D. Teney, A. Dick, and A. v. d. Hengel, "Reasoning over vision and language: E xploring the benefits of supplemental knowledge," *arXiv preprint arXiv:2101.06013*, 2021.

[362] K. J. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," in *CVPR*, 2016.

[363] M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro, "Megatron-lm: Training multi-billion parameter language models using model parallelism," *arXiv preprint arXiv:1909.08053*, 2019.

[364] A. Shonenkov, A. Kuznetsov, D. Dimitrov, T. Shavrina, D. Chesakov, A. Maltseva, A. Fenogenova, I. Pavlov, A. Emelyanov, S. Markov, *et al.*, "RuCLIP–new models and experiments: A technical report," *arXiv preprint arXiv:2202.10784*, 2022.

[365] N. Shvetsova, B. Chen, A. Rouditchenko, S. Thomas, B. Kingsbury, R. S. Feris, D. Harwath, J. Glass, and H. Kuehne, "Everything at once-multi-modal fusion transformer for video retrieval," in *CVPR*, 2022.

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

[367] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv: 1409.1556*, 2014.

[368] 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.

[369] 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.

[370] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach, "Towards vqa models that can read," in *CVPR*, 2019.

[371] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, "Grounded compositional semantics for finding and describing images with sentences," *TACL*, 2014.

[372] H. Song, L. Dong, W.-N. Zhang, T. Liu, and F. Wei, "CLIP models are few-shot learners: Empirical studies on VQA and visual entailment," in *ACL*, 2022.

[373] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *ICLR*, 2021.

[374] 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.

[375] T. Srinivasan and Y. Bisk, "Worst of both worlds: Biases compound in pre-trained vision-and-language models," in *4th Workshop on Gender Bias in Natural Language Processing, NAACL*, 2022.

[376] 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.

[377] Y. Su, T. Lan, Y. Liu, F. Liu, D. Yogatama, Y. Wang, L. Kong, and N. Collier, "Language models can see: Plugging visual controls in text generation," *arXiv preprint arXiv:2205.02655*, 2022.

[378] A. Suhr, S. Zhou, A. Zhang, I. Zhang, H. Bai, and Y. Artzi, "A corpus for reasoning about natural language grounded in photographs," in *ACL*, 2019.

[379] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "VideoBERT: A joint model for video and language representation learning," in *ICCV*, 2019.

[380] S. Sun, Y.-C. Chen, L. Li, S. Wang, Y. Fang, and J. Liu, "Lightningdot: Pre-training visual-semantic embeddings for real-time image-text retrieval," in *NAACL*, 2021.

[381] S. Sun, Y. Cheng, Z. Gan, and J. Liu, "Patient knowledge distillation for bert model compression," in *EMNLP*, 2019.

[382] S. Sun, Z. Gan, Y. Cheng, Y. Fang, S. Wang, and J. Liu, "Contrastive distillation on intermediate representations for language model compression," in *EMNLP*, 2020.

[383] Y.-L. Sung, J. Cho, and M. Bansal, "Lst: Ladder side-tuning for parameter and memory efficient transfer learning," in *NeurIPS*, 2022.

[384] Y.-L. Sung, J. Cho, and M. Bansal, "Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks," in *CVPR*, 2022.

[385] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *NeurIPS*, 2014.

[386] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015.

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

[388] H. Tan and M. Bansal, "Vokenization: Improving language understanding with contextualized, visual-grounded supervision," in *EMNLP*, 2020.

[389] H. Tan, J. Lei, T. Wolf, and M. Bansal, "VIMPAC: Video pre-training via masked token prediction and contrastive learning," *arXiv preprint arXiv:2106.11250*, 2021.

[390] H. Tan, X. Liu, X. Li, Y. Zhang, and B. Yin, "Semantics-enhanced adversarial nets for text-to-image synthesis," in *ICCV*, 2019.

[391] M. Tang, Z. Wang, Z. Liu, F. Rao, D. Li, and X. Li, "Clip4caption: Clip for video caption," in *ACMMM*, 2021.

[392] Y. Tang, D. Ding, Y. Rao, Y. Zheng, D. Zhang, L. Zhao, J. Lu, and J. Zhou, "Coin: A large-scale dataset for comprehensive instructional video analysis," in *CVPR*, 2019.

[393] Z. Tang, J. Cho, H. Tan, and M. Bansal, "Vidlankd: Improving language understanding via video-distilled knowledge transfer," in *NeurIPS*, 2021.

[394] Z. Tang, J. Lei, and M. Bansal, "DeCEMBERT: Learning from noisy instructional videos via dense captions and entropy minimization," in *NAACL*, 2021.

[395] D. Teney, P. Anderson, X. He, and A. van den Hengel, "Tips and tricks for visual question answering: Learnings from the 2017 challenge," in *CVPR*, 2018.

[396] D. Teney, L. Liu, and A. van Den Hengel, "Graph-structured representations for visual question answering," in *CVPR*, 2017.

[397] Y. Tewel, Y. Shalev, I. Schwartz, and L. Wolf, "Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic," in *CVPR*, 2022.

[398] T. Thrush, R. Jiang, M. Bartolo, A. Singh, A. Williams, D. Kiela, and C. Ross, "Winoground: Probing vision and language models for visio-linguistic compositionality," in *CVPR*, 2022.

[399] C. Tian, W. Wang, X. Zhu, X. Wang, J. Dai, and Y. Qiao, "Vl-ltr: Learning class-wise visual-linguistic representation for long-tailed visual recognition," *arXiv preprint arXiv:2111.13579*, 2021.

[400] A. Torabi, N. Tandon, and L. Sigal, "Learning language-visual embedding for movie understanding with natural-language," *arXiv preprint arXiv:1609.08124*, 2016.

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

[402] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015.
[403] M. Tsimpoukelli, J. Menick, S. Cabi, S. Eslami, O. Vinyals, and F. Hill, "Multimodal few-shot learning with frozen language models," in *NeurIPS*, 2021.
[404] 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.
[405] R. Vedantam, K. Desai, S. Lee, M. Rohrbach, D. Batra, and D. Parikh, "Probabilistic neural symbolic models for interpretable visual question answering," in *ICML*, 2019.
[406] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *CVPR*, 2015.
[407] R. Villegas, M. Babaeizadeh, P.-J. Kindermans, H. Moraldo, H. Zhang, M. T. Saffar, S. Castro, J. Kunze, and D. Erhan, "Phenaki: Variable length video generation from open domain textual description," *arXiv preprint arXiv:2210.02399*, 2022.
[408] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *CVPR*, 2015.
[409] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200-2011 dataset," 2011.
[410] A. J. Wang, Y. Ge, R. Yan, Y. Ge, X. Lin, G. Cai, J. Wu, Y. Shan, X. Qie, and M. Z. Shou, "All in one: Exploring unified video-language pre-training," *arXiv preprint arXiv:2203.07303*, 2022.
[411] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "Superglue: A stickier benchmark for general-purpose language understanding systems," in *NeurIPS*, 2019.
[412] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *ICLR*, 2019.
[413] J. Wang, X. Hu, Z. Gan, Z. Yang, X. Dai, Z. Liu, Y. Lu, and L. Wang, "UFO: A unified transformer for vision-language representation learning," *arXiv preprint arXiv:2111.10023*, 2021.

[414] J. Wang, X. Hu, P. Zhang, X. Li, L. Wang, L. Zhang, J. Gao, and Z. Liu, "Minivlm: A smaller and faster vision-language model," *arXiv preprint arXiv:2012.06946*, 2020.

[415] 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.

[416] J. Wang, Y. Ge, G. Cai, R. Yan, X. Lin, Y. Shan, X. Qie, and M. Z. Shou, "Object-aware video-language pre-training for retrieval," in *CVPR*, 2022.

[417] J. Wang, D. Chen, Z. Wu, C. Luo, L. Zhou, Y. Zhao, Y. Xie, C. Liu, Y.-G. Jiang, and L. Yuan, "Omnivl: One foundation model for image-language and video-language tasks," in *NeurIPS*, 2022.

[418] L. Wang, Y. Li, J. Huang, and S. Lazebnik, "Learning two-branch neural networks for image-text matching tasks," *TPAMI*, 2018.

[419] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *CVPR*, 2016.

[420] M. Wang, J. Xing, and Y. Liu, "Actionclip: A new paradigm for video action recognition," *arXiv preprint arXiv:2109.08472*, 2021.

[421] P. Wang, Q. Wu, C. Shen, A. Dick, and A. Van Den Hengel, "Fvqa: Fact-based visual question answering," *TPAMI*, 2017.

[422] P. Wang, Q. Wu, C. Shen, A. v. d. Hengel, and A. Dick, "Explicit knowledge-based reasoning for visual question answering," in *IJCAI*, 2017.

[423] 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.

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

[425] W. Wang, L. Dong, H. Cheng, H. Song, X. Liu, X. Yan, J. Gao, and F. Wei, "Visually-augmented language modeling," *arXiv preprint arXiv:2205.10178*, 2022.

[426] 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.

[427] 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.

[428] X. Wang, J. Wu, J. Chen, L. Li, Y.-F. Wang, and W. Y. Wang, "VATEX: A large-scale, high-quality multilingual dataset for video-and-language research," in *ICCV*, 2019.

[429] Y. Wang, H. Yang, X. Qian, L. Ma, J. Lu, B. Li, and X. Fan, "Position focused attention network for image-text matching," in *IJCAI*, 2019.

[430] Y. Wang, J. Xu, and Y. Sun, "End-to-end transformer based model for image captioning," in *AAAI*, 2022.

[431] Y. Wang, S. Joty, M. R. Lyu, I. King, C. Xiong, and S. C. Hoi, "Vd-bert: A unified vision and dialog transformer with bert," in *EMNLP*, 2020.

[432] Z. Wang, M. Li, R. Xu, L. Zhou, J. Lei, X. Lin, S. Wang, Z. Yang, C. Zhu, D. Hoiem, *et al.*, "Language models with image descriptors are strong few-shot video-language learners," *arXiv preprint arXiv:2205.10747*, 2022.

[433] 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.

[434] M. Wortsman, G. Ilharco, J. W. Kim, M. Li, S. Kornblith, R. Roelofs, R. G. Lopes, H. Hajishirzi, A. Farhadi, H. Namkoong, *et al.*, "Robust fine-tuning of zero-shot models," in *CVPR*, 2022.

[435] M. Wray, H. Doughty, and D. Damen, "On semantic similarity in video retrieval," in *CVPR*, 2021.

[436] B. Wu, S. Yu, Z. Chen, J. B. Tenenbaum, and C. Gan, "STAR: A benchmark for situated reasoning in real-world videos," in *NeurIPS*, 2021.

[437] C. Wu, J. Liang, X. Hu, Z. Gan, J. Wang, L. Wang, Z. Liu, Y. Fang, and N. Duan, "NUWA-infinity: Autoregressive over autoregressive generation for infinite visual synthesis," *arXiv preprint arXiv:2207.09814*, 2022.

[438] C. Wu, J. Liang, L. Ji, F. Yang, Y. Fang, D. Jiang, and N. Duan, "N\ uwa: Visual synthesis pre-training for neural visual world creation," in *ECCV*, 2022.

[439] C. Wu, Z. Lin, S. Cohen, T. Bui, and S. Maji, "Phrasecut: Language-based image segmentation in the wild," in *CVPR*, 2020.

[440] J. Wu, J. Lu, A. Sabharwal, and R. Mottaghi, "Multi-modal answer validation for knowledge-based vqa," in *AAAI*, 2022.

[441] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[442] J. Xiao, X. Shang, A. Yao, and T.-S. Chua, "Next-qa: Next phase of question-answering to explaining temporal actions," in *CVPR*, 2021.

[443] N. Xie, F. Lai, D. Doran, and A. Kadav, "Visual entailment: A novel task for fine-grained image understanding," *arXiv preprint arXiv:1901.06706*, 2019.

[444] Y. Xie, X. Wang, R. Wang, and H. Zha, "A fast proximal point method for computing exact wasserstein distance," in *UAI*, 2020.

[445] Y. Xie, L. Zhou, X. Dai, L. Yuan, N. Bach, C. Liu, and M. Zeng, "Visual clues: Bridging vision and language foundations for image paragraph captioning," *arXiv preprint arXiv:2206.01843*, 2022.

[446] D. Xu, Z. Zhao, J. Xiao, F. Wu, H. Zhang, X. He, and Y. Zhuang, "Video question answering via gradually refined attention over appearance and motion," in *ACMMM*, 2017.

[447] 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," in *ACL*, 2021.

[448] H. Xu, G. Ghosh, P.-Y. Huang, P. Arora, M. Aminzadeh, C. Feichtenhofer, F. Metze, and L. Zettlemoyer, "VLM: Task-agnostic video-language model pre-training for video understanding," in *Findings of ACL*, 2021.

[449] H. Xu, G. Ghosh, P.-Y. Huang, D. Okhonko, A. Aghajanyan, F. Metze, L. Zettlemoyer, and C. Feichtenhofer, "VideoCLIP: Contrastive pre-training for zero-shot video-text understanding," in *EMNLP*, 2021.

[450] 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.

[451] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *CVPR*, 2016.

[452] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015.

[453] M. Xu, Z. Zhang, F. Wei, Y. Lin, Y. Cao, H. Hu, and X. Bai, "A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model," *arXiv preprint arXiv:2112.14757*, 2021.

[454] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," in *CVPR*, 2018.

[455] H. Xue, T. Hang, Y. Zeng, Y. Sun, B. Liu, H. Yang, J. Fu, and B. Guo, "Advancing high-resolution video-language representation with large-scale video transcriptions," in *CVPR*, 2022.

[456] H. Xue, Y. Huang, B. Liu, H. Peng, J. Fu, H. Li, and J. Luo, "Probing inter-modality: Visual parsing with self-attention for vision-language pre-training," in *NeurIPS*, 2021.

[457] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid, "Just ask: Learning to answer questions from millions of narrated videos," in *ICCV*, 2021.

[458] J. Yang, Y. Bisk, and J. Gao, "Taco: Token-aware cascade contrastive learning for video-text alignment," in *ICCV*, 2021.

[459] 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.

[460] J. Yang, J. Duan, S. Tran, Y. Xu, S. Chanda, L. Chen, B. Zeng, T. Chilimbi, and J. Huang, "Vision-language pre-training with triple contrastive learning," in *CVPR*, 2022.

[461] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, Y. Shao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *arXiv preprint arXiv:2209.00796*, 2022.

[462] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *CVPR*, 2019.

[463] Z. Yang, T. Chen, L. Wang, and J. Luo, "Improving one-stage visual grounding by recursive sub-query construction," in *ECCV*, 2020.

[464] 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.

[465] 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.

[466] Z. Yang, B. Gong, L. Wang, W. Huang, D. Yu, and J. Luo, "A fast and accurate one-stage approach to visual grounding," in *ICCV*, 2019.

[467] Z. Yang, Y. Lu, J. Wang, X. Yin, D. Florencio, L. Wang, C. Zhang, L. Zhang, and J. Luo, "Tap: Text-aware pre-training for text-vqa and text-caption," in *CVPR*, 2021.

[468] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *CVPR*, 2016.

[469] 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.

[470] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *ECCV*, 2018.

[471] T. Yao, Y. Pan, Y. Li, and T. Mei, "Hierarchy parsing for image captioning," in *ICCV*, 2019.

[472] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *ICCV*, 2017.

[473] Y. Yao, A. Zhang, Z. Zhang, Z. Liu, T.-S. Chua, and M. Sun, "CPT: Colorful prompt tuning for pre-trained vision-language models," *arXiv preprint arXiv:2109.11797*, 2021.

[474] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, and J. Tenenbaum, "Neural-symbolic vqa: Disentangling reasoning from vision and language understanding," in *NeurIPS*, 2018.

[475] 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.

[476] H. You, L. Zhou, B. Xiao, N. Codella, Y. Cheng, R. Xu, S.-F. Chang, and L. Yuan, "Learning visual representation from modality-shared contrastive language-image pre-training," in *ECCV*, 2022.

[477] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *CVPR*, 2016.

[478] F. Yu, J. Tang, W. Yin, Y. Sun, H. Tian, H. Wu, and H. Wang, "Ernie-vil: Knowledge enhanced vision-language representations through scene graphs," in *AAAI*, 2021.

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

[480] 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," *arXiv preprint arXiv:2206.10789*, 2022.

[481] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, and T. L. Berg, "Mattnet: Modular attention network for referring expression comprehension," in *CVPR*, 2018.

[482] L. Yu, E. Park, A. C. Berg, and T. L. Berg, "Visual madlibs: Fill in the blank image generation and question answering," *arXiv preprint arXiv:1506.00278*, 2015.

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

[484] Y. Yu, J. Kim, and G. Kim, "A joint Sequence fusion model for video question answering and retrieval," in *ECCV*, 2018.

[485] Z. Yu, D. Xu, J. Yu, T. Yu, Z. Zhao, Y. Zhuang, and D. Tao, "Activitynet-qa: A dataset for understanding complex web videos via question answering," in *AAAI*, 2019.

[486] Z. Yu, J. Yu, Y. Cui, and J. Li, "VQA challenge 2019 winner talk," *VQA Challenge 2019*, 2019.

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

[488] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *ICCV*, 2017.

[489] H. Yuan, J. Jiang, S. Albanie, T. Feng, Z. Huang, D. Ni, and M. Tang, "Rlip: Relational language-image pre-training for human-object interaction detection," *arXiv preprint arXiv:2209.01814*, 2022.

[490] 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.

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

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

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

[494] R. Zellers, J. Lu, X. Lu, Y. Yu, Y. Zhao, M. Salehi, A. Kusupati, J. Hessel, A. Farhadi, and Y. Choi, "MERLOT reserve: Neural script knowledge through vision and language and sound," in *CVPR*, 2022.

[495] R. Zellers, X. Lu, J. Hessel, Y. Yu, J. S. Park, J. Cao, A. Farhadi, and Y. Choi, "Merlot: Multimodal neural script knowledge models," in *NeurIPS*, 2021.

[496] A. Zeng, A. Wong, S. Welker, K. Choromanski, F. Tombari, A. Purohit, M. Ryoo, V. Sindhwani, J. Lee, V. Vanhoucke, *et al.*, "Socratic models: Composing zero-shot multimodal reasoning with language," *arXiv preprint arXiv:2204.00598*, 2022.

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

[498] Y. Zeng, W. Zhou, A. Luo, and X. Zhang, "Cross-view language modeling: Towards unified cross-lingual cross-modal pre-training," *arXiv preprint arXiv:2206.00621*, 2022.

[499] Z. Zeng, Y. Luo, Z. Liu, F. Rao, D. Li, W. Guo, and Z. Wen, "Tencent-MVSE: A large-scale benchmark dataset for multi-modal video similarity evaluation," in *CVPR*, 2022.

[500] 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.

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

[502] C. Zhang, B. Van Durme, Z. Li, and E. Stengel-Eskin, "Visual commonsense in pretrained unimodal and multimodal models," *arXiv preprint arXiv:2205.01850*, 2022.

[503] D. Zhang, X. Dai, X. Wang, and Y.-F. Wang, "S3D: Single shot multi-span detector via fully 3D convolutional networks," in *BMVC*, 2018.

[504] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *ICCV*, 2017.

[505] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan++: Realistic image synthesis with stacked generative adversarial networks," *TPAMI*, 2018.

[506] H. Zhang, W. Yin, Y. Fang, L. Li, B. Duan, Z. Wu, Y. Sun, H. Tian, H. Wu, and H. Wang, "ERNIE-ViLG: Unified generative pre-training for bidirectional vision-language generation," *arXiv preprint arXiv:2112.15283*, 2021.

[507] H. Zhang, Y. Niu, and S.-F. Chang, "Grounding referring expressions in images by variational context," in *CVPR*, 2018.

[508] 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.

[509] J. Zhang, J. P. Chang, C. Danescu-Niculescu-Mizil, L. Dixon, Y. Hua, N. Thain, and D. Taraborelli, "Conversations gone awry: Detecting early signs of conversational failure," in *ACL*, 2018.

[510] 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.

[511] Q. Zhang, Z. Lei, Z. Zhang, and S. Z. Li, "Context-aware attention network for image-text retrieval," in *CVPR*, 2020.

[512] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, *et al.*, "Opt: Open pre-trained transformer language models," *arXiv preprint arXiv:2205.01068*, 2022.

[513] D. Zhao, A. Wang, and O. Russakovsky, "Understanding and evaluating racial biases in image captioning," in *ECCV*, 2021.

[514] H. Zhao, I. Hadji, N. Dvornik, K. G. Derpanis, R. P. Wildes, and A. D. Jepson, "P3IV: Probabilistic procedure planning from instructional videos with weak supervision," in *CVPR*, 2022.

[515] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, and Y.-D. Shen, "Dual-path convolutional image-text embeddings with instance loss," *ACM TOMM*, 2020.

[516] 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.

[517] C. Zhou, C. C. Loy, and B. Dai, "Extract free dense labels from CLIP," in *ECCV*, 2022.

[518] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *CVPR*, 2022.

[519] L. Zhou, J. Gao, D. Li, and H.-Y. Shum, "The design and implementation of xiaoice, an empathetic social chatbot," *Computational Linguistics*, 2020.

[520] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and vqa," in *AAAI*, 2020.

[521] L. Zhou, C. Xu, and J. J. Corso, "Towards automatic learning of procedures from web instructional videos," in *AAAI*, 2018.

[522] M. Zhou, L. Yu, A. Singh, M. Wang, Z. Yu, and N. Zhang, "Unsupervised vision-and-language pre-training via retrieval-based multi-granular alignment," in *CVPR*, 2022.

[523] M. Zhou, L. Zhou, S. Wang, Y. Cheng, L. Li, Z. Yu, and J. Liu, "Uc2: Universal cross-lingual cross-modal vision-and-language pre-training," in *CVPR*, 2021.

[524] W. Zhou, Y. Zeng, S. Diao, and X. Zhang, "VLUE: A multi-task benchmark for evaluating vision-language models," in *ICML*, 2022.

[525] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra, "Detecting twenty-thousand classes using image-level supervision," *arXiv preprint arXiv:2201.02605*, 2022.

[526] Y. Zhou, R. Zhang, C. Chen, C. Li, C. Tensmeyer, T. Yu, J. Gu, J. Xu, and T. Sun, "LAFITE: Towards language-free training for text-to-image generation," in *CVPR*, 2022.

[527] L. Zhu and Y. Yang, "Actbert: Learning global-local video-text representations," in *CVPR*, 2020.

[528] X. Zhu, J. Zhu, H. Li, X. Wu, H. Li, X. Wang, and J. Dai, "Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks," in *CVPR*, 2022.

[529] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7w: Grounded question answering in images," in *CVPR*, 2016.

[530] M. Zhuge, D. Gao, D.-P. Fan, L. Jin, B. Chen, H. Zhou, M. Qiu, and L. Shao, "Kaleido-bert: Vision-language pre-training on fashion domain," in *CVPR*, 2021.

[531] D. Zhukov, J.-B. Alayrac, R. G. Cinbis, D. Fouhey, I. Laptev, and J. Sivic, "Cross-task weakly supervised learning from instructional videos," in *CVPR*, 2019.

[532] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *ECCV*, 2014.