Title
Learning to Describe Images via Natural Language

Author
Yan, An

Publication Date
2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

Learning to Describe Images via Natural Language

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Computer Science

by

An Yan

Committee in charge:

Professor Julian McAuley, Chair
Professor Sanjoy Dasgupta
Professor Chun-Nan Hsu
Professor Jingbo Shang

2024

The Dissertation of An Yan is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Professor Julian McAuley, for his support as the chair of my committee. It is a privilege to work under his guidance and have the freedom to explore research areas based on my personal interests.

My appreciation goes out to all the members of my dissertation committee members. Prof. Jingbo Shang, Prof. Sanjoy Dasgupta, and Dr. Chun-Nan Hsu, for offering me advice and encouragement which significantly shaped my PhD study at UC San Diego. I have learned a lot from them, through project discussions, taking their classes, and personal meetings.

Subsequently, my research would not have been made progress without help from many other mentors I had the privilege to work with: William Wang (UCSB), Xin Eric Wang (UCSC), Zhengyuan Yang, Jianfeng Wang and Jianwei Yang (Microsoft), Raghav Addanki, Zhao Song and Tong Yu (Adobe), Cem Akkaya, Licheng Yu, Charlie Zhu and Jian Jin (Meta), Chaosheng Dong, Jinmiao Fu and Yan Gao (Amazon), Craig Bennett and Nic Jedem (Amazon).

I would also like to thank my friends, collaborators and labmates: Wanrong Zhu, Tsu-Jui Fu, Amilcare Gentili, Ruida Zhou, Yiwu Zhong, Chengyu Dong, Jiacheng Li, Zhankui He, Zexue He, Yu Wang, Jessica Echterhoff, Canwen Xu, Noveen Sachdeva, Wang-Cheng Kang, Mengting Wan, Jianmo Ni, Petros Karypis, and many more.

Lastly, I am grateful for my family and my love, for the support and love they gave to me, as always. Thank you all for being there for me.

As for writing this thesis, I would like to thank all of my co-authors who kindly approved the following publications and material to be included in my dissertation:

Chapter 1, in part, is a reprint of the material as it appears in the following publications:

“Learning Concise and Descriptive Attributes for Visual Recognition” by An Yan, Yu Wang, Yiwu Zhong, Chengyu Dong, Zexue He, Yujie Lu, William Wang, Jingbo Shang, Julian McAuley, published at *International Conference on Computer Vision 2023*. The dissertation author was the primary investigator and author of this paper.

“Mitigating Spurious Correlations for Medical Image Classification via Natural Language

Concepts " by An Yan, Yu Wang, Petros Karypis, Zexue He, Amilcare Gentili, Chun-Nan Hsu, Julian McAuley, published at Medical Imaging Workshop in *Conference on Neural Information Processing Systems 2023*. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in part, is a reprint of the material as it appears in the following publications:

"Describing Visual Differences Needs Semantic Understanding of Individuals" by An Yan, Xin Wang, Tsu-Jui Fu, William Wang, published at *European Chapter of the Association for Computational Linguistics*,2021 The dissertation author was the primary investigator and author of this paper.

"Weakly Supervised Contrastive Learning for Chest X-Ray Report Generation" by An Yan, Zexue He, Xing Lu, Jiang Du, Eric Chang, Amilcare Gentili, Julian McAuley, Chun-Nan Hsu, published at *Empirical Methods in Natural Language Processing*, 2021 The dissertation author was the primary investigator and author of this paper.

"Personalized Showcases: Generating Multi-Modal Explanations for Recommendations" by An Yan, Zhankui He, Jiacheng Li, Tianyang Zhang, Julian McAuley, published at *The International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023. The dissertation author was the primary investigator and author of this paper.

"Visualize Before You Write: Imagination-Guided Open-Ended Text Generation" by Wanrong Zhu, An Yan, Yujie Lu, Wenda Xu, Xin Eric Wang, Miguel Eckstein, William Wang, published at *European Chapter of the Association for Computational Linguistics*, 2023. The dissertation author was one of the primary authors of this paper.

Chapter 3, in part, is currently being prepared for submission for the publication of the material:

"GPT-4V in Wonderland: Large Multimodal Models for Zero-Shot Smartphone GUI Navigation" by An Yan, Zhengyuan Yang, Wanrong Zhu, Kevin Lin, Linjie Li, Jianfeng Wang, Jianwei Yang, Yiwu Zhong, Julian McAuley, Jianfeng Gao, Zicheng Liu, Lijuan Wang. The dissertation author was the primary investigator and author of this paper.

VITA

2014–2018 Bachelor of Engineering, University of Science and Technology of China

2018–2020 Master of Science, University of California San Diego

2020–2024 Doctor of Philosophy, University of California San Diego

PUBLICATIONS

"Driving through the Concept Gridlock: Unraveling Explainability Bottlenecks in Automated Driving" Winter Conference on Applications of Computer Vision, 2024

"Mitigating Spurious Correlations for Medical Image Classification via Natural Language Concepts" Medical Imaging Workshop at Conference on Neural Information Processing Systems 2023

"MedEval: A Multi-Level, Multi-Task, and Multi-Domain Medical Benchmark for Language Model Evaluation" Empirical Methods in Natural Language Processing, 2023

"Learning Concise and Descriptive Attributes for Visual Recognition" International Conference on Computer Vision, 2023

"Personalized Showcases: Generating Multi-Modal Explanations for Recommendations" The International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023

"Visualize Before You Write: Imagination-Guided Open-Ended Text Generation" European Chapter of the Association for Computational Linguistics, 2023

"ImaginE: An Imagination-Based Automatic Evaluation Metric for Natural Language Generation" European Chapter of the Association for Computational Linguistics, 2023

"Disambiguating Medical Reports via Contrastive Knowledge Infusion" AAAI Conference on Artificial Intelligence, 2023

"RadBERT: Adapting Language Models to Radiology" Journal of Radiology: Artificial Intelligence, 2022

"Personalized Complementary Product Recommendation" The ACM Web Conference, 2022

"Weakly Supervised Contrastive Learning for Chest X-Ray Report Generation" Empirical Methods in Natural Language Processing, 2021

"Describing Visual Differences Needs Semantic Understanding of Individuals" European Chapter of the Association for Computational Linguistics, 2021

"2D Convolutional Neural Networks for Sequential Recommendation" ACM International Conference on Information and Knowledge Management, 2019

"PA3D: Pose-Action 3D Machine for Video Recognition" IEEE Conference on Computer Vision and Pattern Recognition, 2019

FIELDS OF STUDY

Major Field: Artificial Intelligence, Computer Vision and Natural Language Processing

ABSTRACT OF THE DISSERTATION

Learning to Describe Images via Natural Language

by

An Yan

Doctor of Philosophy in Computer Science

University of California San Diego, 2024

Professor Julian McAuley, Chair

Teaching machines to describe visual images is one of the most long-standing challenges in the field of Machine Learning. This thesis tackles the problem of describing images via natural language: how to build machine learning models to read visual images and describe their content as well as answer relevant questions. From the application perspective, strong image captioning systems can contribute to applications such as visual question answering, dialogue systems and visual-based robotics. For a long term goal, if we can build such systems (e.g., GPT-4V and beyond), they would be a crucial step towards building Artificial General Intelligence: computers that can perceive and explore the world as humans do.

This thesis focus on neural models: building vision understanding and language genera-

tion models with deep neural networks. It mainly consists of three parts.

First, we will introduce concept bottleneck models, a class of models that build concept layers for visual understanding. We will present our work on learning a concise concept space, and follow-up applications for medical imaging to gain robustness.

In the second part of this thesis, we investigate how we can build practical image captioning systems based on different neural text generation architectures, from LSTM to transformers and pre-trained language models. In particular, we will cover four different tasks: 1) how we can describe the visual difference of two images; 2) how we can write medical reports given Chest X-rays to assist doctors; 3) how to generate personalized explanations for recommender systems; 4) how to augment text generation with visual imagination generated from vision diffusion models.

In the third part, we will discuss recent advances, future directions and open questions in this field, focusing on aspects of datasets, models, and applications. We will also introduce some of our on-going attempts for these directions: for example, how to navigate phone screens and complete mobile tasks with GPT-4V.

In summary, my research contributes to the field of vision and language, specifically visual understanding via natural language, from the aspects of data curation, algorithm designing, model training, as well as various downstream applications.

Introduction

0.1 Motivation

Teaching machines to describe visual images via natural language is one of the most long-standing challenges in the field of Artificial Intelligence. Before we start, we need to discuss what it means to describe images via natural language? One of the most well-known example is the image captioning task from Microsoft COCO (MS-COCO) [56], as shown in Figure 0.1, where the task is to provide a generic caption for an input image. Overall, there are many tasks that fit into this setting, we provide some typical tasks here:

1. Image captioning: Given an image, describe the content in one single sentence or multiple sentences with detailed analysis.

2. Visual Question Answering: Given an image and a question in natural language, generate or select an answer for the question.

3. Visual Dialog: Given ore or more images, the model is asked to have a conversation and multi-turn dialog with a human user.

Before the era of transformers [95] and pretrained models (e.g., BERT [20], CLIP [79], GPT-4V [73]), these tasks are often explored individually, and different architectures or learning algorithms are designed. But most recently, the trend in the research community is to pretrain large-scale models that can unify those tasks in a single model. Perhaps the most representative and advanced example is GPT-4V, a pretrained Large Multimodal Models (LMMs) that serves as a chatbot which capable of taking visual inputs and chat with human users. These models

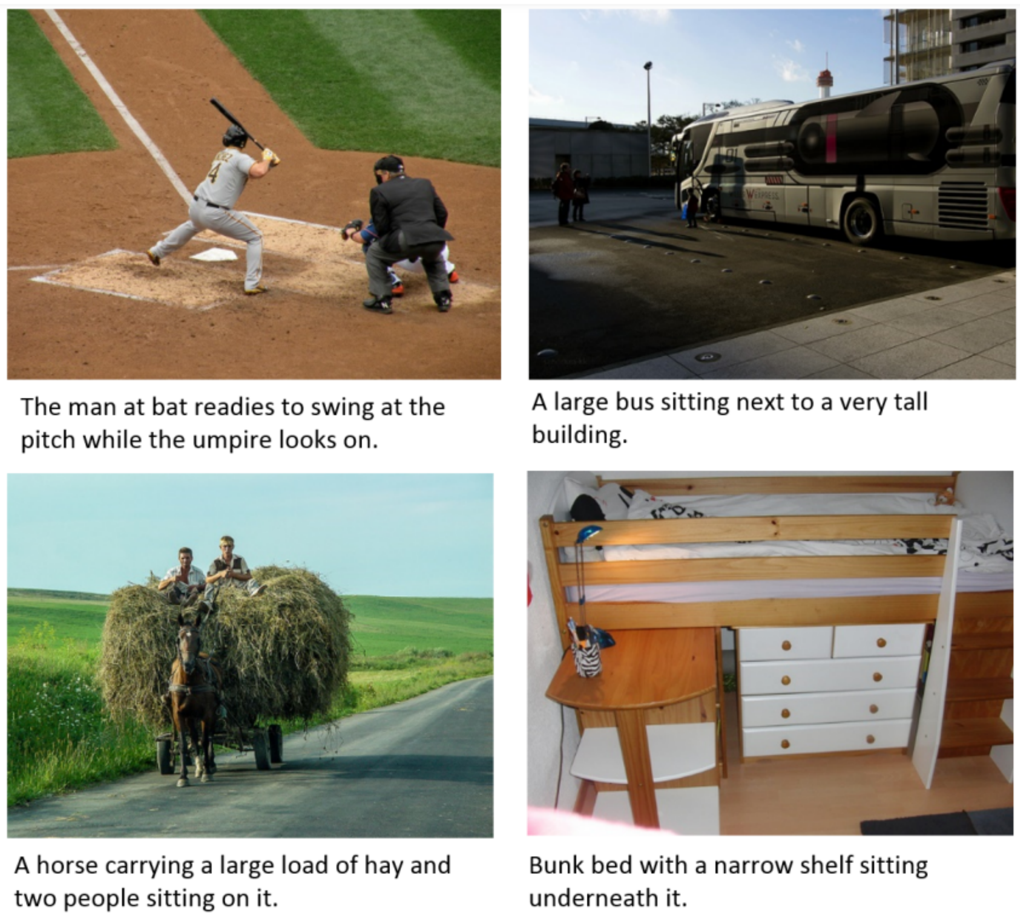


Figure 0.1. Example images and captions from the Microsoft COCO dataset [56].

have enabled various applications, such as chatbots [73, 59], image generation [5], video generation [74], mobile assistants [113], robot navigation [67], etc.

In this thesis, we are interested in how we can build practical neural models to describe images, leveraging "small models" such as ResNet [31], LSTM [34] and non-pretrained transformers [95], as well as recent "foundation models" such as CLIP, GPT-X, and image diffusion models. Our vision-language applications spans from interpretable and robust image classification, automated driving, medical report generation, personalization and recommendation, to mobile GUI navigation and more.

0.2 Thesis Outline

This thesis mainly consists of three parts:

In the first part, Chapter 1, we will introduce the idea of understanding visual features via natural language concepts, which builds concept bottleneck models for interpretable image classification. In our work [111], we find that there is great noise and redundancy in the current formulation of concept bottleneck models. We design a learning-to-search algorithm to find a concise subset of representative concepts. We then present an application following this direction, for medical imaging classification [112].

In the second part of this thesis, Chapter 2, we present some of our efforts on building practical image captioning models. Specifically, we will cover four different tasks: 1) how to describe the visual difference of two similar images [110]; 2) how to generate medical reports given Chest X-rays [108]; 3) how to generate personalized explanations for recommendation [109]; 4) how to augment text generation with visual knowledge from vision diffusion models [131].

The third part, Chapter 3, we will explore recent advances, and future directions for vision and language. Is GPT-4V all we need? If not, what we can improve on top of it? We will summarize some failure cases from interacting with GPT-4V, and discuss two main aspects for data-drive machine learning: data and model. Lastly, we show some promising applications using LMMs, including our effort on building phone screen navigation agents with GPT-4V [113].

0.3 Contributions

We make the following contributions in this thesis: 1) We made non-trivial contributions to the research topic of concept bottleneck models, which builds interpretable and robust image classifiers. 2) We made the effort of tackling different image captioning tasks, and proposed various learning paradigms for these tasks. 3) We discuss the future of vision-language. We are also among the first to explore multimodal agents with multimodal foundation models.

Chapter 1

Visual Understanding via Concepts

In this chapter, we will cover the essence of concept bottleneck models: a family of neural classification models that are built with intermediate concept layers to obtain interpretability. Before delving into the details of our work, we will give a brief introduction to the basics of concept bottleneck models in Section 1.1. We will cover basic ideas and recent advance of building concept bottleneck models with foundation models.

We then move to our work of learning concise concept bottleneck models in Section 1.2. We show the problem of recent LLM-based concept bottleneck models: there is great redundancy in these large-scale concepts generated by LLMs. In particular, we design a simple yet effective learning-to-search method, to efficiently find much smaller subsets of concepts that can maintain similar classification performance compared with these large-scale concepts.

Lastly, We present a follow-up work along this direction for medical imaging [112]. We show an additional benefit of using concepts: classification with concepts naturally brings robustness. This work has the potential to mitigate spurious correlations in neural networks, which is especially important in the medical domain, as there are many confounding factors that could impact the diagnostic decision of a machine learning model.

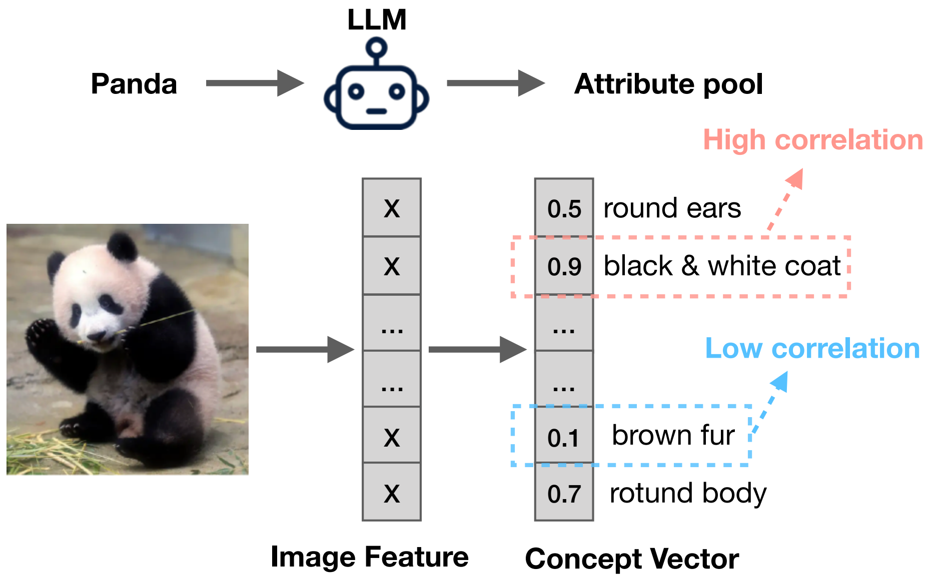


Figure 1.1. An explanation of concept bottleneck models. Concepts with high and low correlations to the input image are highlighted in red and blue color.

1.1 Preliminary: Concept Bottleneck Models

Concept bottleneck models [42] are a family of neural models that leverage natural language concepts to gain interpretability and interactivity. The idea is to build a concept layer, that projects raw image features into this layer, and use these concepts to predict labels.

There are two issues with the original concept bottleneck models proposed in [42]: First, for different tasks, one need to manually design concepts, which requires huge human effort. A general image classification task consists of hundreds of classes. For medical domain, it would also require expert knowledge from doctors.

The recent advance of foundation models creates new opportunities for building interpretable visual recognition models, as demonstrated by the powerful capabilities of models such as GPT-3 and ChatGPT in encoding world knowledge [9, 75]. A set of visual attributes that are related to visual classes can be effortlessly queried from LLMs. Then we can noisily label the correlation between an image and an attribute using VLMs (e.g., CLIP) by computing their similarity. Given a set of attributes, we can construct a semantic vector where each dimension corresponds to a visual attribute and the value represents the similarity between the image and the attribute. A high-level illustraion is shown in Figure 1.1. One recent work [115] shows that a large set of attributes from LLMs (e.g., 50 attributes per class) can achieve comparable performance to image features in a linear probing setting.

1.2 Learning a Concise Concept Space

two key observations motivate us to re-think this formulation: (1) A large number of attributes dramatically hurts the interpretability of a model. It is unrealistic to manually check thousands of attributes to fully understand model decisions. (2) We surprisingly find that when the number of attributes is large enough (e.g., the dimension of image features), random words drawn from the entire vocabulary can perform equally well as LLM-generated attributes. Moreover, reducing the number of random words by 25% can still attain competitive performance. This indicates that redundant and noisy information exists in the massive LLM-generated attributes.

1.2.1 Visual Concepts vs. Random Words

Conceptually, the semantic projection in concept bottleneck models resembles principal component analysis, where we aim to find a set of bases in the form of natural language, and by projecting the images into these bases we obtain a concept space where each dimension in the space corresponds to a visual concept. However, one would argue if the large set of attribute concepts we obtained from LLMs is the optimal language basis. As of today, LLMs are models that noisily condense world knowledge from the web, and are not optimized for visual recognition or visual reasoning tasks.

From a linear algebra view, given an image feature $X_I \in \mathbb{R}^N$, if we have N concepts where their vectors are orthogonal to each other, we can perfectly preserve the information and obtain same performance with those N concepts. Even though it is almost impossible to sample orthogonal embeddings from a language encoder, this still motivate us to find

Intuitively, most concepts in the large concept pool are irrelevant to classify a certain class.

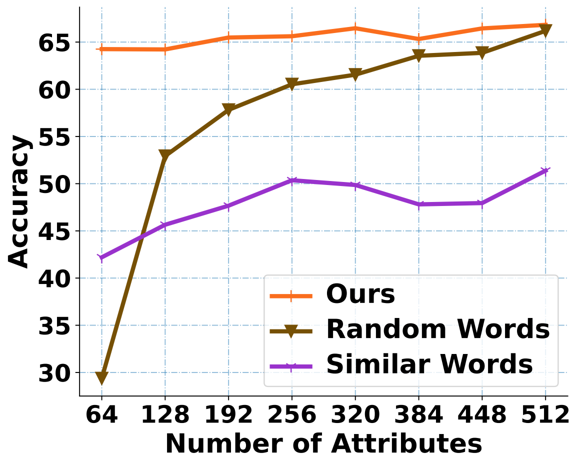


Figure 1.2. Performance comparison with random or similar words on CUB.

Table 1.1. Examples from Random (R), Silimlar (S), GPT-3 (G) attributes

	Examples

R	`boy champagne`
	`allied whose acrobat`
	`eight centered lobby heads`

S	`red,gray,snow wings`
	`orange wings`
	`lime,navy wings`

G	`sloping forehead`
	`distinctive white throat`
	`bright red head and breast`

For example, attributes that describe dogs are less likely to be suitable attributes to recognize birds or cars. Practically, formatting a compact attribute set is also helpful for humans to interact with the model and understand its behavior better. A small number of attributes is much easier for diagnostic purposes and making decisions with these neural models, which is the ultimate goal of building interpretable models.

Therefore, we propose **the concept space hypothesis**: there exist subsets of attributes that can still achieve high classification performance with a much smaller size.

To test this hypothesis, we start experiments by comparing the performance of LLM concepts with random words. Surprisingly, we find that LLM-generated attributes in a large quantity behave just like random words. We compare our method of using GPT-3 generated visual attributes with random or similar words. Here, we constructed random words by randomly choosing 1-5 words from the entire English vocabulary, and semantically similar words by combining 1-3 random colors with the noun "wings" as suffix. As shown in Figure 1.2, when $K = 512$, random words perform as well as GPT-3 attributes in terms of classification accuracy. Even reducing K from 512 to 256 does not significantly hurt its performance. But when K is small (e.g., 64), the performance of random words drops dramatically. We conjecture that it is

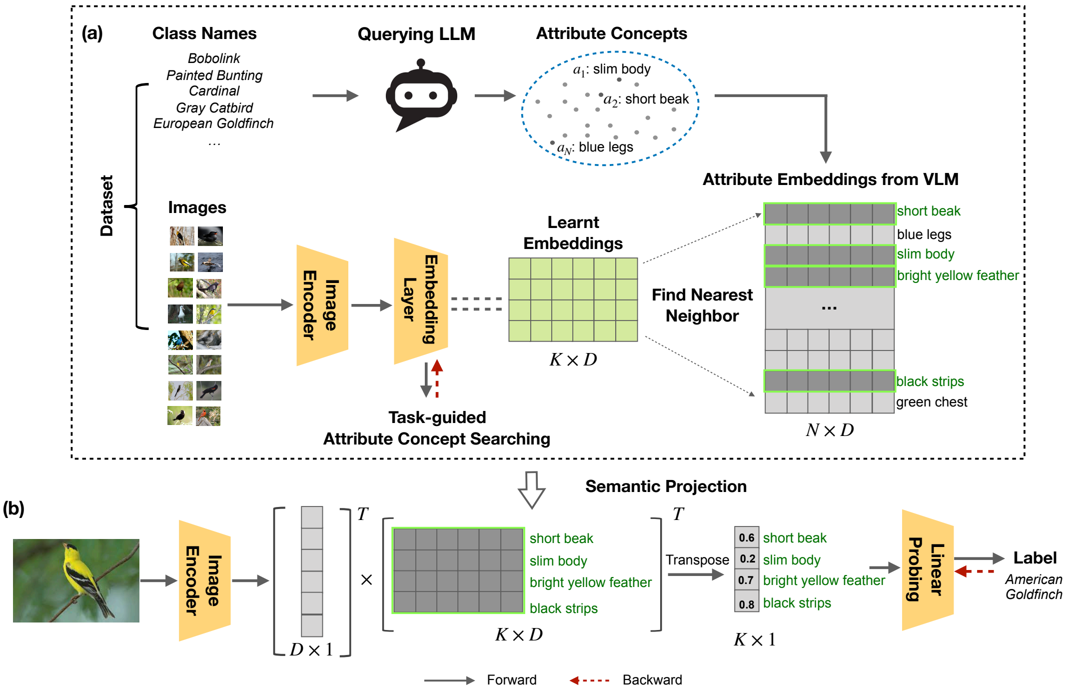


Figure 1.3. The framework of our model. (a) Querying attributes from LLMs and finding a concise set of representative attributes; (b) An example using the attributes for interpretable visual recognition.

because text embeddings randomly drawn from CLIP are nearly orthogonal bases [102]. Given an image feature $\in \mathbb{R}^D$, projection with a set of K=D orthogonal bases can perfectly preserve its information. We further explore how similar words (e.g., red wings, yellow wings) behave. Embeddings of similar words in a trained language model are not orthogonal bases hence the projection will lose information when K is large (e.g., intuitively it is hard to classify 200 bird species using only the color combination of wings). But as K gets smaller, since those similar words have close semantic meanings, they start to outperform random words. Overall, these findings motivate us to find a concise set of meaningful attributes while maintaining competitive performance.

1.2.2 Method: Learning to Find Concise Concepts

In this section, we introduce our key components for a new paradigm of visual recognition. It mainly consists of three modules: **First**, given an image domain, we query large language models to obtain a large set of visual attributes for the categories of a task. **Second**, we use a semantic transformation to project the image features into attribute features via a vision-language model, where each dimension in the new space corresponds to an attribute concept, and a higher value represents higher correlation between the image and the attribute. **Finally**, given the large space of attributes, we propose a novel learning-to-search method to efficiently prune the attributes into a much smaller subset to obtain a concise model for classification. See framework in Figure 1.3.

Generating Attributes from LLMs The first step of our framework is to obtain a set of appropriate attribute concepts. Given a dataset with different categories, (e.g., CUB with 200 bird classes), what are the distinctive visual attributes to recognize them? Manually labeling and designing these attribute concepts can be costly, and can not scale to large numbers of classes. Large Language Models (LLMs), such as GPT-3 [9] and ChatGPT, provide an alternative solution. We can view these language models as implicit knowledge bases with exceptional world knowledge on a variety of tasks and topics, which humans can easily interact with through natural language to query knowledge. To effectively query knowledge from LLMs with regard to classifying images, we use the following prompting template to query attributes for each class:

Q: What are the useful visual features to distinguish Y_c in a photo?

where Y_c corresponds to the name of class c in the form of natural language.

Semantic Projection After obtaining a pool consisting of N attribute concepts, the second challenge is how we can best leverage these attributes to build interpretable image classifiers. Recent advances of vision-language models such as CLIP bridge the gap between images and text, by pre-training models with large scale image-text pairs. Intuitively, converting from images to text is a discretization process that will unavoidably lose rich semantic information stored in an image.

To better preserve information, we use a semantic projection that transforms a visual feature into an attribute concept space. Given an image I, we convert the D-dimensional image feature $\mathbf{V} \in \mathbb{R}^D$ into an N-dimensional attribute concept vector $\mathbf{A} \in \mathbb{R}^N$:

$$\begin{aligned}
&\mathbf{V} = \Theta_V(I), \mathbf{T}_i = \Theta_T(a_i) \\
&s_i = \cos(\mathbf{V}, \mathbf{T}_i), i = 1, ..., N \\
&\mathbf{A} = (s_1, \ldots, s_N)^T
\end{aligned} \tag{1.1}$$

where $\cos(\cdot,\cdot)$ is the cosine similarity between two vectors, s_i is the cosine similarity between two vectors. Θ_V and Θ_T are the visual and text encoder of a VLM. $\mathbf{T}_i$ is the embedding of the i-th attribute in the attribute concept pool, $i \in \{1, \ldots, N\}$. $\mathbf{A}$ is the semantic vector of image I.

Task-Guided Concept Searching Given N attribute concepts, finding a subset of K attributes ($K \ll N$) to achieve the optimal classification performance is essentially a searching problem: the brute force solution is to exhaustively train different models from all possible combinations and find the one with the best performance, which is impractical due to high computational cost given the large search space.

Inspired by dictionary learning and vector quantization techniques [94], we present a learning-to-search method that learns a dictionary to approximate an expressive subset of attributes given fixed K. Specifically, we first define an embedding matrix $\mathbf{E} \in \mathbb{R}^{K \times D}$, where K is a K-way categorical that equals the number of attributes, and D is the dimensionality of embedding vectors $\mathbf{V}$ and $\mathbf{T}_i$ (i.e., the latent dimension of VLMs), where $\mathbf{V}$ and $\mathbf{T}_i$ is the image embedding and the i-th attribute embedding shown in Eq.(1.1). Since our goal is to find K attributes to be expressive, we propose a task-guided attribute concept searching method to optimize for a particular task. For visual recognition tasks, we use a classification head to project the dictionary into K_C classes and guide the learning process with the categorical cross-entropy loss:

$$\mathcal{L}_{ce} = -\frac{1}{M} \sum_{i=1}^{M} \sum_{c=1}^{K_C} y_{i,c} \log(p_{i,c}) \tag{1.2}$$

where M is the number of images in a mini-batch, $y_{i,c}$ is the binary indicator of the i-th image in the mini-batch belonging to class c, and $p_{i,c}$ is the predicted probability of the i-th image belonging to class c.

But simply training with the guidance of the cross-entropy loss is suboptimal, as the embeddings $\mathbf{E}$ are not in the same space of $\mathbf{T}$. Thus, we use the Mahalanobis distance as a constraint to encourage the embeddings to be optimized towards the latent space of vision-language models. Given a sampled probability distribution $\mathbf{T}$, the Mahalanobis distance of $\mathbf{E}_j$ from $\mathbf{T}$ is defined as

$$\mathscr{D}_{mah}^{j} = \sqrt{(\mathbf{E}_j - \boldsymbol{\mu})\mathbf{S}^{-1}(\mathbf{E}_j - \boldsymbol{\mu})} \tag{1.3}$$

where $\boldsymbol{\mu} = (\mu_1, ..., \mu_D)$ is the mean vector and $\mathbf{S}$ is the positive-definite covariance matrix of $\mathbf{T}$. Then the regularization term is defined as:

$$\mathscr{L}_{mah}^{j} = \frac{1}{K} \sum_{j=1}^{k} \mathscr{D}_{mah}^{j} \tag{1.4}$$

Overall, our model is optimized with a mixture of two losses:

$$\mathscr{L}_{loss} = \mathscr{L}_{ce} + \lambda \sum_{j=1}^{K} \mathscr{L}_{mah}^{j}. \tag{1.5}$$

After training, we have the embedding matrix $\mathbf{E}$ which will be used for searching the attributes from the attribute concept pool $\mathscr{C}$. Note that for $\mathbf{E} \in \mathbb{R}^{K*D}$, each row of $\mathbf{E}$ is a D-dimensional vector. We denote the j-th row of $\mathbf{E}$ as $\mathbf{E}_j$. We use greedy search as follows:

$$\begin{aligned} \mathbf{T}_j^* = \underset{i \in \{1,\cdots,N\}}{\arg\max} \cos(\mathbf{T}_i, \mathbf{E}_j), \\ \text{s.t. } \mathbf{T}_j^* \neq \mathbf{T}_k^*, \forall 1 \leq k < j, \\ \text{where } j \text{ is from 1 to } K, \end{aligned} \tag{1.6}$$

As j iterates from 1 to K, we can find K attribute embeddings $\mathbf{T}_j^*, j \in \{1, \cdots, K\}$, which corresponds to K expressive attribute concepts and are the condensed features containing the necessary knowledge for the task. With the selected attributes, we can calculate the semantic vector of each image as in Eq. (1.1), where each dimension of the vector is a similarity score between the image and an attribute. We evaluate the performance of these semantic vectors with linear probes, and the obtained linear model is used for inference and analysis.

1.2.3 Experimental Results

Datasets We conduct our experiments on 8 different image classification datasets, including: CUB [98], CIFAR-10 and CIFAR-100 [44], Food-101 [8], Flower [70], Oxford-pets [77], Stanford-cars [43], Imagenet [18]. For Imagenet, it is not trivial to analyze all 1000 diverse classes. So we narrow the scope to 397 animal classes, with 509,230/19,850 samples for train/test. We denote this subset as Imagenet-Animals. For other datasets, most of them include images within a specific domain (CUB, Flower, Food, Oxford-pets, Stanford-cars), while CIFAR-10 and CIFAR-100 contain broader classes that lie across domains.

Baselines We compare with state-of-the-art works that leverage attributes either from human annotations or from LLMs. For a fair comparison, we use linear probes to evaluate all methods: (1) **CompDL** [121] builds semantic vectors using CLIP scores between human-designed attributes and images. (2) **LaBO** [115] is a recent work that builds semantic vectors with a large set of attributes from LLMs. (3) **Human** [42]. Attribute labels for each image are annotated by humans. We compare with two versions: binary labels for each attribute, and calibrated labels with confidence scores given by annotators.

Comparison with previous work We first compare our method with LaBo [115]. It is designed to use M_c concepts per class with default number of 50, which corresponds to 10,000 attributes for CUB. For fair-comparison, we set M_c as 1 and 2 in the experiments. As shown in Table 1.2, our method outperforms LaBo with the same number of attributes on both the full and few-shot setting. Furthermore, our method can achieve similar accuracy with only a smaller number of attributes (e.g., 32 attributes for CUB). These results suggest that our learned

Table 1.2. Comparison with state-of-the-art. LaBo is designed to use at least as many attributes as classes. We use "–" to denote non-applicability.

Datasets	CUB			CIFAR-10			CIFAR-100			Flower		
K	32	200	400	8	10	20	64	100	200	32	102	204
LaBo	–	60.93	62.61	–	78.11	84.84	–	75.10	76.94	–	80.98	86.76
Ours	60.27	**63.88**	**64.05**	77.47	**80.09**	**87.99**	73.31	**75.12**	**77.29**	80.88	**87.26**	**89.02**
Datasets	Food			Oxford_Pets			Stanford_cars			Imagenet_Animals		
K	64	101	202	16	37	74	64	196	392	128	397	794
LaBo	–	79.95	81.33	–	76.91	84.33	–	72.33	74.39	–	74.88	75.49
Ours	78.41	**80.22**	**81.85**	76.29	**83.15**	**85.91**	72.07	**74.57**	**75.56**	74.48	**75.69**	**75.83**

Table 1.3. Comparison with human annotations on CUB.

K (# of attributes)	8	16	32	312
Human Binary [98]	4.02	7.31	10.11	47.38
Human Calibration [42]	3.75	7.15	9.78	43.37
CompDL [121]	12.64	26.41	28.69	52.60
Ours	**31.67**	**48.55**	**60.27**	**65.17**

attributes are discriminative enough to classify the images, despite given much fewer attributes.

We then further compare with human annotations from CUB. For $K < 312$, we select attributes based on their accumulated confidence score for all samples. As shown in Table 1.3, human annotated attributes are more noisy than CLIP similarities. With the same attributes, CLIP scores from CompDL build more expressive features. On top of that, our LLM-suggested attributes significantly outperform human designs, e.g. by using 16 attributes we achieve similar performance as 312 attributes defined by humans.

1.2.4 Knowledge Summarization with Concise Concepts

We show our learned descriptive attributes with $K = 8$ in Figure 1.4. Intuitively, we can observe these attributes are distinctive for each domain. Take birds recognition (CUB) as an example, the eight attributes covered most of the body parts of a bird (head, breast, legs, etc.). As we are condensing knowledge from hundreds of bird classes, each attribute broadly covers many categories. A bright red head and breast can be a noticeable visual attribute for many

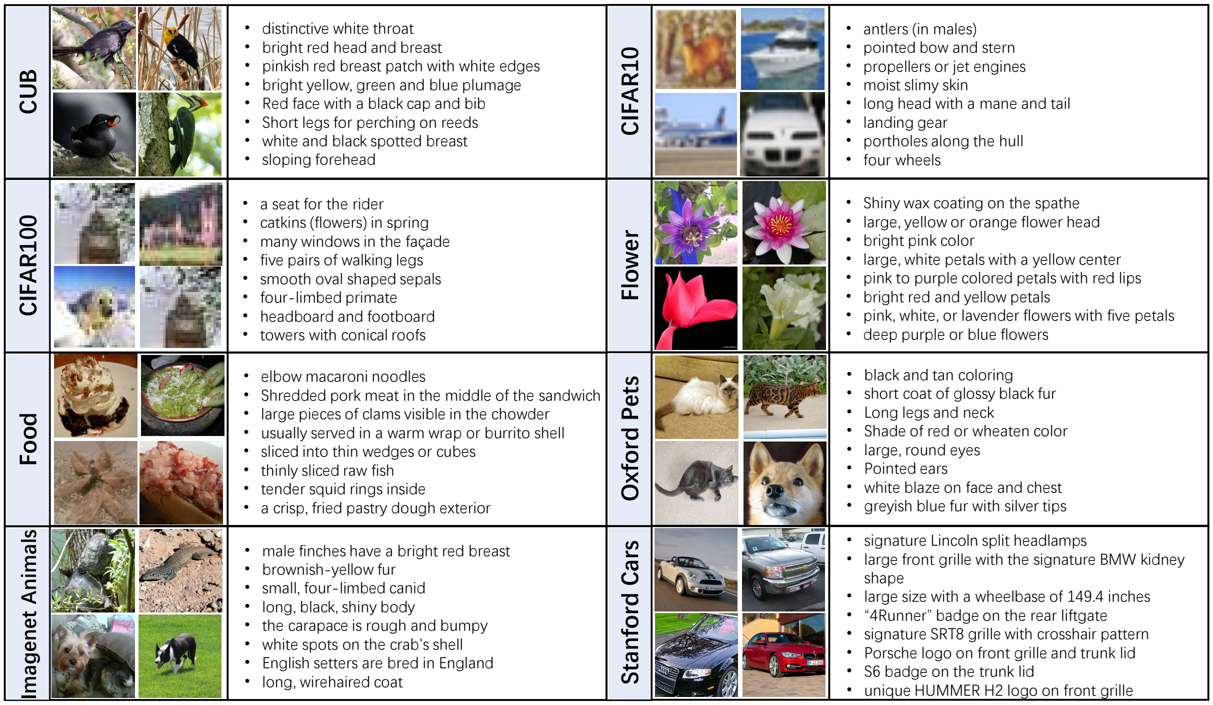


Figure 1.4. A concise set of 8 descriptive attributes learned for each dataset with sampled images.

bird species, such as the Northern Cardinal and the Vermilion Flycatcher. Overall, explaining a domain with a few descriptive attributes is challenging, even for an expert with sufficient domain knowledge. But our model is able to automatically provide a level of knowledge to help humans understand how visual recognition works.

1.3 Application to Medical Imaging

1.3.1 Concepts bring Robustness

Neural networks are prone to learn spurious correlations for classification tasks. In the non-medical domain, [87] found that models trained on the Waterbirds dataset correlate waterbirds with backgrounds containing water, and models trained on the CelebA dataset [60] correlate males with dark hair. This could be more of an issue for medical image classification, as confounding factors broadly exist and labeled data are often limited [16]. Take the classification of patient X-rays between Covid-19 and normal for instance, certain factors such as the hospitals

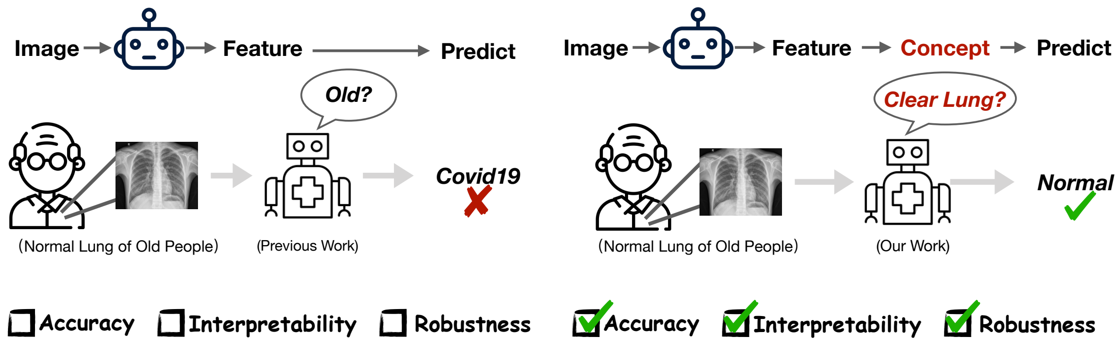


Figure 1.5. High level illustration of our framework which utilizes concepts for medical image classification to achieve interpretability and robustness while maintaining accuracy. **Left**: Classification with a visual encoder; **Right**: Classification with concepts. A Chest X-ray from a healthy old individual may be classified as Covid-19 because of the age, while our method can mitigate spurious correlation by classifying with clinical concepts.

where the X-rays are performed and the age of the patient strongly correlate with the target disease classification. To quantify this issue, we curated datasets of known confounding factors such as hospitals, age and gender, and found that standard visual classifiers and previous popular methods designed to mitigate spurious correlations often perform poorly and struggle to generalize in these datasets. As a concrete example, instead of learning to predict Covid or normal, the classifier might instead learn to predict if the X-ray is from a young or old patient.

Inspired by recent work and our own work above that uses attributes [42, 111, 23] or descriptions [63] to amplify image classification, in this section, we bring a new perspective to address spurious correlations in medical imaging through natural language concepts.

Specifically, we elicit medical knowledge from large language models (e.g., GPT-4) in a zero-shot manner to build a set of concepts, i.e., concise descriptors regarding each disease or pathology, and project visual features into the concept space using a vision-language model to connect two modalities, and finally classify medical images with the concept vector. By doing so, we explicitly tell the model to learn desired features rather than possible spurious correlations, hence improving the robustness of classification models. Empirically, we find that classification using concepts can alleviate spurious correlations and substantially improve

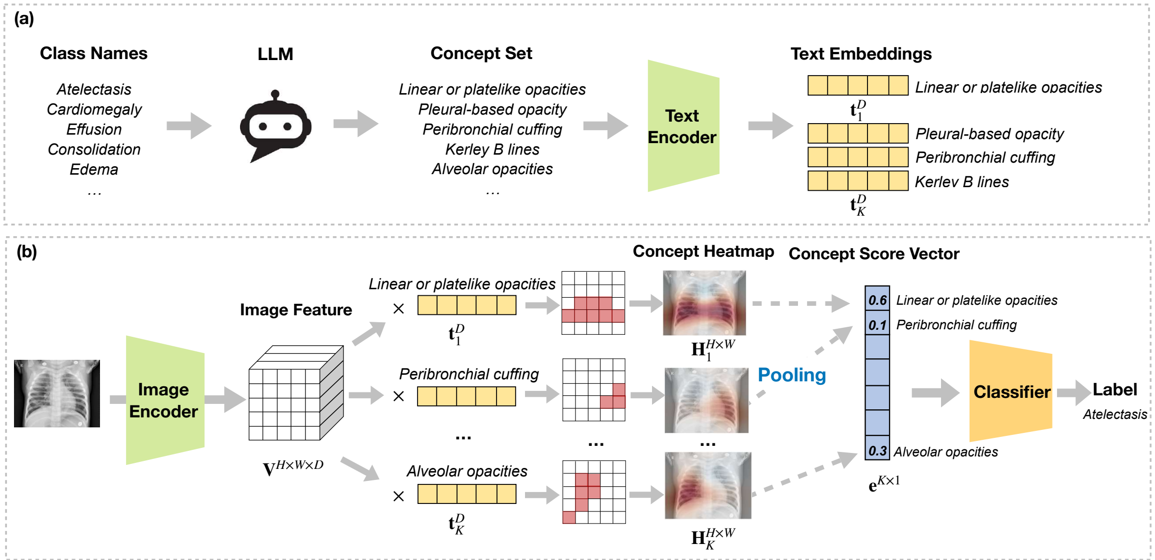


Figure 1.6. The overall framework of classification with concepts. (a) Eliciting medical knowledge from GPT-4. (b) Projecting visual features into the concept space for classification.

classification performance on challenging datasets. An illustration is shown in Figure 1.5.

1.3.2 Framework Overview

As shown in Figure 1.6, we first interact with a large language model, GPT-4, to generate useful medical concepts for target diseases or pathology. We further ask a board-certified radiologist to check if these concepts are clinically correct. Empirically, we find GPT-4 concepts to be comparable or slightly better then human designed descriptors for classification.

After obtaining a set of N useful concepts $\mathscr{C} = \{c_1, c_2, \ldots, c_N\}$ from GPT-4, we leverage a specialized medical VLM, BioViL [7], to connect medical concepts with images. Given an image I and a concept set $\mathscr{C} = \{c_1, c_2, \ldots, c_N\}$, we extract its feature map $\mathbf{V} \in \mathbb{R}^{H \times W \times D}$ and the text embeddings $\mathbf{t}_i \in R^D$ from the two stream visual and text encoders.

Given $\mathbf{V}$ and $\mathbf{t}_i$, one can obtain a heatmap $\mathbf{H}_i$ by computing their cosine distance. We then apply average pooling to obtain the similarity score between a concept and an image:

$$s_i = \frac{1}{H \cdot W} \sum_{j=1}^{H} \sum_{k=1}^{W} \mathbf{H}_i^{j,k} \tag{1.7}$$

We obtain a concept vector $\mathbf{e} = (s_1, \ldots, s_N)^T$, representing the similarity between an image and a set of concepts.

Finally, we learn a decision layer to predict M_c classes given concept vectors. To gain a level of interpretability [71], we normalize concept vectors e to a scale of 0 and 1, and use a fully connected layer $\mathbf{W}^F \in \mathbb{R}^{M_c \times N}$ without bias terms. Training is done with a categorical cross-entropy loss.

Since each logit $z_{i,c}$ (corresponding to $p_{i,c}$) is a linear combination of non-negative concept scores $\mathbf{e}_i = (s_1^{(i)}, \ldots, s_N^{(i)})^T$, i.e., $z_{i,c} = \sum_{j=1}^{N} \mathbf{W}_{j,c}^F s_j^{(i)}$. We can interpret the weights in W_F as importance scores for classifying class c with concepts. Higher absolute values in $W_{j,c}^F$ indicate that the concept is more important for classification. Negative weights can be interpreted as negations, i.e., the non-existence of this concept. Overall, this final linear layer offers a way for human to better understand and analyze model decisions.

1.3.3 Creating Diagnostic Datasets

With the framework designed above, we are ready to conduct experiments to verify its effectiveness. But before that, we need to find the right benchmarks, ideally with explicit confounding factors to make performance analysis easier.

To this end, We first create four datasets with strong confounding factors as follows: **(1) NIH-gender**: We use the meta data from NIH-CXR [101] to build a subset with male and female chest X-rays for classifying Atelectasis and Effusion. **(2) NIH-age**: We build a subset from NIH-CXR with young and old chest X-rays to classify normal and abnormal. **(3) NIH-agemix**: Similar to NIH-age, the normal cases in the training set consist of 90% young patients and 10% old paitents, while the abnormal cases consist of 90% old patients and 10% young patients. **(4) Covid-mix**: We create a dataset from various sources to classify Covid-19 and non-Covid Pneumonia. There can be several confounding factors in this dataset, for example, hospital, age and gender. For all datasets, we create train/test sets with balanced classes, hence report accuracy as the main metric.

We then evaluate models on four standard benchmarks. These datasets do not have explicit confounding factors, since the training and test samples are assumed to be randomly drawn from the same distribution: **(1) NIH-CXR** [101] collected from NIH Clinical Center. **(2) Covid-QU** [14] collected by Qatar University. **(3) Pneumonia** [40]: A public dataset for detecting pneumonia. **(4) Open-i** [17]collected by Indiana University Hospital.

1.3.4 Experimental Results

Baselines We compare our method with three types of baselines: domain generalization models designed to mitigate spurious correlations (ERMERM [89], Fish [89], LISA [117]), , visual encoders (BioViL Image Features), and recent methods that build Concept Bottleneck Models (Post-Hoc CBMs [120], Label-free CBMs [71]) for image classification.

Results on datasets with explicit confounding factors We first compare our method with baselines on the four curated datasets, as shown in Table 1.4. **First**, standard methods to mitigate spurious correlations, such as Fish and LISA, may fail to capture the domain shift in this challenging setting. Moreover, they also require explicit confounding labels to better learn domain invariant features, which need careful expert analysis and are often difficult to obtain in real-world scenarios. **Second**, image features can easily overfit to spurious correlations, even for a medical domain visual encoder such as BioViL. For example, it only has an accuracy of 9% on the test set of NIH-age, meaning it learns to predict whether the chest x-ray is taken for a young or old individual almost perfectly, instead of learning to predict normal or abnormal. **Finally**, our method also outperforms recent concept bottleneck models and attains much better robustness, demonstrating the effectiveness of our simple projection from visual features into a concept space. **Results on other benchmarks** We then evaluate the performance of our method on the other four datasets, which are popular benchmarks with no explicit confounding factors. As shown in Table 1.5, on two datasets, projecting visual features into a concept space still leads to slightly better classification performance than latent visual features. We conjecture the reason could be the implicit confounding factors in the data, even though those datasets are assumed to

Table 1.4. Performance comparison for robustness. Results are in percentage (%).

Models	NIH-gender	NIH-age	NIH-agemix	Covid-mix	Interpretability
ERM	21.70	3.30	13.80	51.73	✗
Fish	21.70	6.00	17.00	52.16	✗
LISA	23.00	2.30	14.20	51.30	✗
BioViL Image Features	71.60	9.40	13.70	51.08	✗
BioViL Image Features (dropouts)	70.20	19.00	28.60	49.57	✗
Post-Hoc CBM	77.40	13.70	16.70	51.08	✓
Label-free CBM	78.90	32.90	35.80	47.40	✓
Ours	**79.60**	**50.70**	**53.40**	**62.36**	✓

be collected in an unbiased way where training and testing are drawn from the same distribution. Hence classification with concepts can improve model robustness and performance even on datasets without explicit confounding factors. This indicates the potential of our method to serve as a universal framework for medical image classification, especially for real-world applications where distribution shift between training and testing is unavoidable.

Table 1.5. Performance comparison on standard datasets without explicit confounding factors.

Models	NIH-CXR	Covid-QU	Pneumonia	Open-i	Interpretability
BioViL Visual Encoder	**63.66**	78.14	86.70	71.01	✗
BioViL Visual Encoder (dropouts)	43.59	68.59	77.08	55.39	✗
Post-Hoc CBM	62.88	**79.09**	87.34	72.35	✓
Label-free CBM	62.40	72.23	88.30	71.91	✓
Ours	63.27	78.00	**88.46**	**72.80**	✓

1.4 Conclusions

In this chapter, we show a novel class of models: concept bottleneck models that builds intermediate concept layers for black-box neural models. We propose a way to learn a concise concept layer to achieve higher interpretability and interactivity. We further qualitatively show these learned concepts present a level of knowledge from visual datasets. Moreover, leveraging concepts naturally brings robustness into the model. We hope our line of work could inspire future research on concept learning, and also contributes to practical applications in various science and engineering domains that use deep learning models.

Chapter 1, in part, is a reprint of the material as it appears in the following publications:

"Learning Concise and Descriptive Attributes for Visual Recognition" by An Yan, Yu Wang, Yiwu Zhong, Chengyu Dong, Zexue He, Yujie Lu, William Wang, Jingbo Shang, Julian McAuley, published at *International Conference on Computer Vision 2023*. The dissertation author was the primary investigator and author of this paper.

"Mitigating Spurious Correlations for Medical Image Classification via Natural Language Concepts " by An Yan, Yu Wang, Petros Karypis, Zexue He, Amilcare Gentili, Chun-Nan Hsu, Julian McAuley, published at Medical Imaging Workshop in *Conference on Neural Information Processing Systems 2023*. The dissertation author was the primary investigator and author of this paper.

Chapter 2

Image Captioning Models

The task of generating textual descriptions of images tests a machine's ability to understand visual data and interpret it in natural language. In this chapter, we will discuss image captioning, one of the most important tasks for vision-language research. Our definition of image captioning is relaxed over the well-known MS-COCO task which does single-sentence captioning on a single image: given one or more images, with optional text input, the model is asked to generate desired text conditioned on these images.

We will present our work for four different image captioning tasks: (1) Describing visual differences between two images. We present a model to learn structured visual representations. (2) Generating radiology reports given Chest X-ray images. We show how contrastive learning can help regularize the text generation process, and encourage model to generate diverse, less-generic reports. (3) Generating personalized explanations for recommender systems. We show how to personalize text generation conditioned on user history and visual inputs. (4) Augmenting pure text generation with visual images. We show how a vision foundation model, e.g., stable diffusion [83], can augment text generation models such as BART [47] and GPT-2 [80].

Even though recent Large Multimodal Models, such as GPT-4V, have shown the benefits of scaling-up, and presented strong capabilities to unify all image captioning tasks, we will show that the ideas we discuss here still deliver insights to improve existing large-scale models and provide principles to design the next-generation vision-language models.

2.1 Describing Visual Differences

2.1.1 Motivation

Visual comparison is the task to describe visual differences between paired images [37, 93, 25]. To complete the task and generate comparative descriptions, a machine should understand the visual differences between a pair of images. Previous methods [37] often consider the pair of pre-trained visual features such as the ResNet features [31] as a whole, and build end-to-end neural networks to predict the description of visual comparison directly. In contrast, humans can easily reason about the visual components of a single image and describe the visual differences between two images based on their semantic understanding of each one. Humans do not need to look at thousands of image pairs to describe the difference of new image pairs, as they can leverage their understanding of single images for visual comparison.

Therefore, we believe that visual differences should be learned by understanding and comparing every single image's semantic representation. Another relevant work [126] conceptually supports this argument, where they show that low-level ResNet visual features lead to poor generalization in vision-and-language navigation, and high-level semantic segmentation helps the agent generalize to unseen scenarios.

2.1.2 Method: Learning-to-Compare

We present a novel framework in Figure 2.1, which consists of three main components. First, a *segmentation encoder* is used to extract structured visual features with strong semantic priors. Then, a *graph convolutional module* performs reasoning on the learned semantic representations. To enhance the understanding of each image, we introduce a *single-image captioning auxiliary loss* to associate the single-image graph representation with the semantic meaning conveyed by its language counterpart. Finally, a decoder generates the visual descriptions comparing two images based on differences in graph representations. All parameters are shared for both images and both tasks.

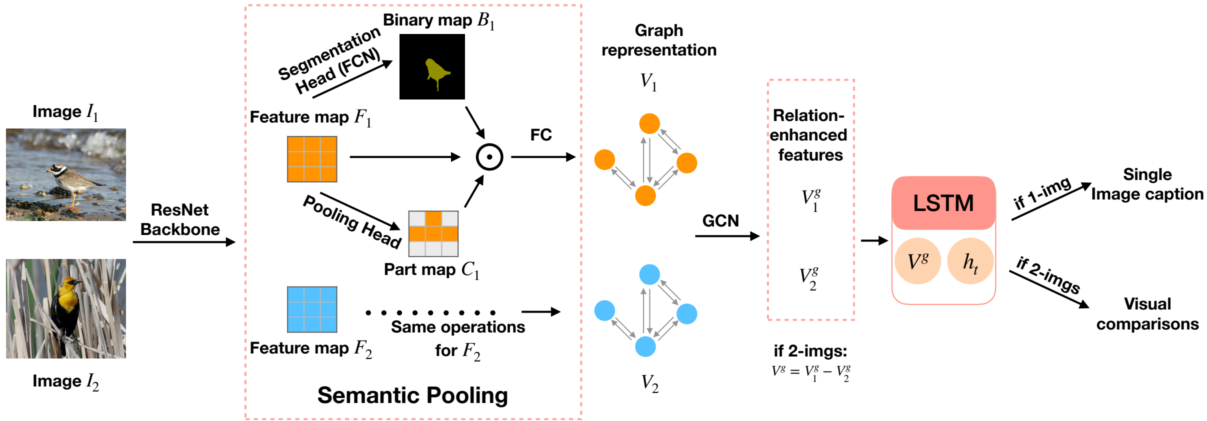


Figure 2.1. Our LEARNING-TO-COMPARE model. It consists of a segmentation encoder, a graph convolutional module, and an LSTM decoder with an auxiliary loss for single-image captioning.

Semantic Representation Construction To extract semantic visual features, we utilize pre-trained fully convolutional networks (FCN) [61] with ResNet-101 as the backbone. An image $\mathscr{I}$ is fed into the ResNet backbone to produce a feature map $\mathscr{F} \in \mathbb{R}^{D\times H\times W}$, which is then forwarded into an FCN head that generates a binary segmentation mask B for the bird class. However, the shapes of these masks are variable for each image, and simple pooling methods such as average pooling and max pooling would lose some information of spatial relations within the mask.

To address this issue and enable efficient aggregation over the area of interest (the masked area), we add a module after the ResNet to cluster each pixel within the mask into K classes. Feature map $\mathscr{F}$ is forwarded through this pooling module to obtain a confidence map $\mathscr{C}$ $\in \mathbb{R}^{K\times H\times W}$, whose entry at each pixel is a K-dimensional vector that represents the probability distribution of K classes.

Then a set of nodes $V = \{v_1, ..., v_K\}, v_k \in \mathbb{R}^D$ is constructed as following:

$$v_k = \sum_{i,j} \mathscr{F} \odot \mathscr{B} \odot \mathscr{C}_k \tag{2.1}$$

where i=1,$...H$, j=1,$...$,W,, $\mathscr{C}_k$ is the k-th probability map and $\odot$ denotes element-wise multiplication.

To enforce local smoothness, i.e., pixels in a neighborhood are more likely belong to one class, we employ total variation norm as a regularization term:

$$\mathscr{L}_{TV} = \sum_{i,j} |C_{i+1,j} - Ci,j| + |C_{i,j+1} - Ci,j| \tag{2.2}$$

Comparative Relational Reasoning Inspired by advances in visual reasoning and graph neural networks [11, 51], we introduce a relational reasoning module to enhance the semantic representation of each image. A fully-connected visual semantic graph $G = (V, E)$ is built, where V is the set of nodes, each containing a regional feature, and E is constructed by measuring the pairwise affinity between each two nodes v_i, v_j in a latent space.

$$A(v_i, v_j) = (W_i v_i)^T (W_j v_j) \tag{2.3}$$

where W_i, W_j are learnable matrices, and A is the constructed adjacency matrix.

We apply Graph Convolutional Networks (GCN) [41] to perform reasoning on the graph. After the GCN module, the output $V^o = \{v_1^o, ..., v_K^o\}, v_k^o \in \mathbb{R}^D$ will be a relationship enhanced representation of a bird. For the visual comparison task, we compute the difference of each two visual nodes from two sets, denoted as $V_{diff}^g = \{v_{diff,1}^o, ..., v_{diff,K}^o\}, v_{diff,k}^o = v_{k,1}^o - v_{k,2}^o \in \mathbb{R}^D$.

Learning to Compare while Learning to Describe After obtaining relation-enhanced semantic features, we use a Long Short-Term Memory (LSTM) [34] to generate captions. As discussed above, semantic understanding of each image is key to solve the task. However, there is no single dataset that contains both visual comparison and single-image annotations. Hence, we leverage two datasets from similar domains to facilitate training. One is for visual comparison, and the other is for single-image captioning. Alternate training is utilized such that for each iteration, two mini-batches of images from both datasets are sampled independently and fed into the encoder to obtain visual representations V^o (for single-image captioning) or V_{diff}^o (for visual comparison).

The LSTM takes V^o or V^o_{diff} with previous output word embedding y_{t-1} as input, updates the hidden state from h_{t-1} to h_t, and predicts the word for the next time step. The generation process of bi-image comparison is learned by maximizing the log-likelihood of the predicted output sentence. The loss function is defined as follows:

$$\mathcal{L}_{diff} = -\sum_t \log P(y_t | y_{1:t-1}, V^o_{diff}) \tag{2.4}$$

Similar loss is applied for learning single-image captioning:

$$\mathcal{L}_{single} = -\sum_t \log P(y_t | y_{1:t-1}, V^o) \tag{2.5}$$

Overall, the model is optimized with a mixture of cross-entropy losses and total variation loss:

$$\mathcal{L}_{loss} = \mathcal{L}_{diff} + \mathcal{L}_{single} + \lambda \mathcal{L}_{TV} \tag{2.6}$$

where λ is an adaptive factor that weighs the total variation loss.

2.1.3 Experimental Results

Datasets The Birds-to-Words (B2W) has 3347 image pairs, and each has around 5 descriptions of visual difference. This leads to 12890/1556/1604 captions for train/val/test splits. Since B2W contains only visual comparisons, We use the CUB-200-2011 dataset (CUB) [98], which consists of single-image captions as an auxiliary to facilitate the training of semantic understanding. CUB has 8855/2933 images of birds for train/val splits, and each image has 10 captions.

Evaluation Metrics Performances are first evaluated on three automatic metrics: BLEU-4 [76], ROUGE-L [55], and CIDEr-D [96]. Each generated description is compared to all five reference paragraphs. Note for this particular task, researchers observe that CIDEr-D is susceptible to common patterns in the data (See Table 2.1 for proof), and ROUGE-L is anecdotally correlated with higher-quality descriptions (which is noted in previous work [25]). Hence we consider ROUGE-L as the major metric for evaluating performances. We then perform a human evaluation to further verify the performance.

Table 2.1. Results for visual comparison on the Birds-to-Words dataset. *Most Frequent* produces only the most observed description in the dataset: "the two animals appear to be exactly the same". *Text-Only* samples captions from the training data according to their empirical distribution. *Neural Naturalist* is a transformer model in [25]. *CNN+LSTM* is a commonly-used CNN encoder and LSTM decoder model.

	Validation			**Test**		
Model	BLEU-4 ↑	ROUGE-L ↑	CIDEr-D ↑	BLEU-4 ↑	ROUGE-L ↑	CIDEr-D ↑
Most Frequent	20.0	31.0	**42.0**	20.0	30.0	**43.0**
Text-Only	14.0	36.0	5.0	14.0	36.0	7.0
Neural Naturalist	24.0	46.0	28.0	22.0	43.0	25.0
CNN+LSTM	25.1	43.4	10.2	24.9	43.2	9.9
L2C [B2W]	31.9	45.7	15.2	31.3	45.3	15.1
L2C [CUB+B2W]	**32.3**	**46.2**	16.4	**31.8**	**45.6**	16.3
Human	26.0	47.0	39.0	27.0	47.0	42.0

Performance Comparison As shown in Table 2.1, first, L2C[B2W] (training with visual comparison task only) outperforms baseline methods on BLEU-4 and ROUGE-L. Previous approaches and architectures failed to bring superior results by directly modeling the visual relationship on ResNet features. Second, joint learning with a single-image caption L2C[B2W+CUB] can help improve the ability of semantic understanding, thus, the overall performance of the model. Finally, our method also has a smaller gap between validation and test set compared to *neural naturalist*, indicating its potential capability to generalize for unseen samples.

2.2 Medical Report Generation

2.2.1 Motivation

Automated radiology report generation aims at generating informative text from radiologic image studies. It could potentially improve radiology reporting and alleviate the workload of radiologists. Recently, following the success of deep learning in conditional text generation tasks such as image captioning [97, 105], many methods have been proposed for this task [38, 54, 57, 12].

Unlike conventional image captioning benchmarks (e.g. MS-COCO [56]) where referenced captions are usually short, radiology reports are much longer with multiple sentences, which pose higher requirements for information selection, relation extraction, and content ordering. To generate informative text from a radiology image study, a caption model is required to understand the content, identify abnormal positions in an image and organize the wording to describe findings in images. However, the standard approach of training an encoder-decoder model with teacher forcing and cross-entropy loss often leads to text generation outputs with high frequency tokens or sentences appearing too often [82, 35]. This problem could be worse for chest X-ray report generation, since the task has a relatively narrow text distribution with domain-specific terminology and descriptions for medical images, and models often struggle to generate long and diverse reports [30, 6].

To tackle these challenging issues, we propose to introduce contrastive learning into chest X-ray report generation. However, simply using random non-target sequences as negative examples in a contrastive framework is suboptimal [46], as random samples are usually easy to distinguish from the correct ones. Hence, we further introduce a weakly supervised contrastive loss that assigns more weights to reports that are semantically close to the target. By exposing the model to these "hard" negative examples during training, it could learn robust representations which capture the essence of a medical image and generate high-quality reports with improved performance on clinical correctness for unseen images.

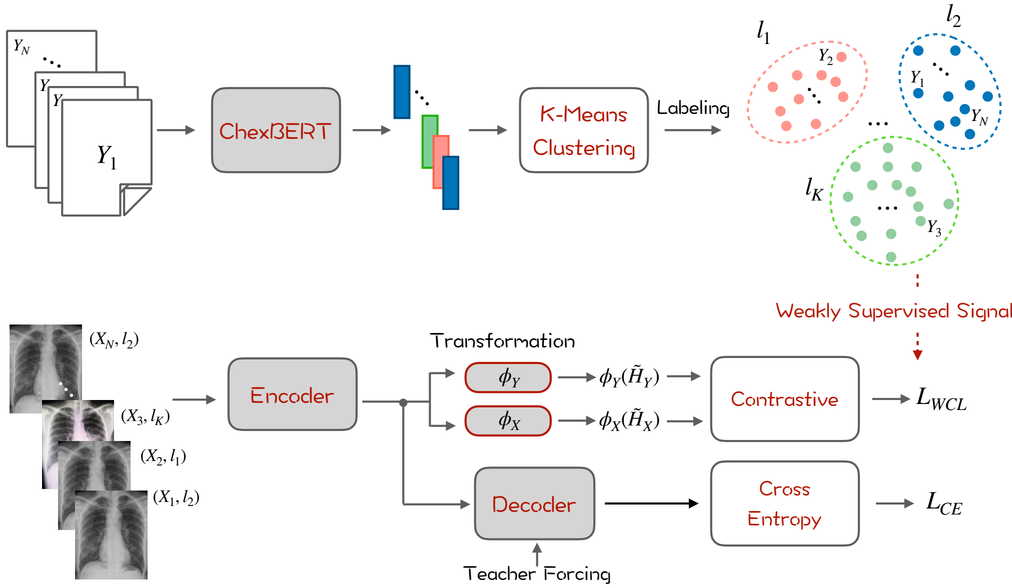


Figure 2.2. Illustration of our weakly supervised contrastive learning framework. We use a task-specific BERT model to label the reports, guiding the contrastive learning process during training.

2.2.2 Method: Contrastive Learning for Text Generation

Generating Reports with Transformer We leverage a memory-driven transformer proposed in [12] as our backbone model, which uses a memory module to record key information when generating long texts.

Given a chest X-ray image I, its visual features X are extracted by pre-trained convolutional neural networks (e.g. ResNet [31]). Then we use the standard encoder in transformer to obtain hidden visual features H_X. The decoding process at each time step t can be formalized as

$$\hat{y}_t = Decoder(H_X, y_1, \ldots, y_{t-1}). \tag{2.7}$$

We use a cross-entropy (CE) loss to maximize the conditional log likelihood $\log p_\theta(Y|X)$ for a given N observations $(X^{(i)}, Y^{(i)})_{i=1}^N$ as follows:

$$\mathcal{L}_{CE} = \sum_{i=1}^{N} \log p_\theta(Y^{(i)}|X^{(i)}). \tag{2.8}$$

Labeling Reports with Finetuned BERT As shown in Figure 2.2, we first extract the embeddings of each report from ChexBERT [90], a BERT model pretrained with biomedical text and finetuned for chest X-ray report labeling. We use the [CLS] embedding of BERT to represent report-level features. We then apply K-Means to cluster the reports into K groups. After clustering, each report Y is assigned with a corresponding cluster label l, where reports in the same cluster are considered to be semantically close to each other.

Weakly supervised Contrastive Learning To regularize the training process, we propose a weakly supervised contrastive loss (WCL). We first project the hidden representations of the image and the target sequence into a latent space:

$$z_x = \phi_x(\tilde{H}_X), z_y = \phi_y(\tilde{H}_Y), \tag{2.9}$$

where $\tilde{H}_X$ and $\tilde{H}_Y$ are the average pooling of the hidden states H_X and H_Y from the transformer, ϕ_x and ϕ_y are two fully connected layers with ReLU activation [66]. We then maximize the similarity between the pair of source image and target sequence, while minimizing the similarity between the negative pairs as follows:

$$\mathscr{L}_{WCL} = \sum_{i=1}^{N} \log \frac{\exp(s_{i,i})}{\sum\limits_{l_i \neq l_j} \exp(s_{i,j}) + \alpha \sum\limits_{l_i = l_j} \exp(s_{i,j})}, \tag{2.10}$$

where $s_{i,j} = sim({z_x}^{(i)}, {z_y}^{(j)})/\tau$, sim is the cosine similarity between two vectors, τ is the temperature parameter, α is a hyperparameter that weighs the importance of negative samples that are semantically close to the target sequence, i.e., with the same cluster label $l_i = l_j$ in Eq. (2.10). Empirically, we find that these samples are "hard" negative samples and the model would perform better by assigning more weights to distinguish these samples.

Overall, the model is optimized with a mixture of cross-entropy loss and weakly super-

Table 2.2. The performance of all baselines and our full model on the test sets of MIMIC-ABN and MIMIC-CXR datasets with respect to natural language generation (NLG) and clinical efficacy (CE) metrics. Results are reported in percentage (%). *ST* is CNN+LSTM with attention [105]. *HCR* [38] is a hierachical CNN-RNN model. *CVSE* [68] is a cross-modal retrieval model. *TopDown* [2] is a widely-used image captioning model. *MDT* is a memory-driven transformer proposed in [12].

Dataset	Model	NLG metrics				CE metrics		
		BLEU-1	BLEU-4	METEOR	ROUGE-L	Precision	Recall	F-1
MIMIC-ABN	*ST*	14.9	3.3	7.2	17.4	20.3	22.2	21.2
	HCR	8.4	1.9	5.9	14.9	26.1	15.7	19.6
	CVSE	19.2	3.6	7.7	15.3	31.7	22.4	26.2
	MDT	24.6	6.6	9.7	23.0	**34.0**	29.1	29.4
	MDT+WCL	**25.6**	**6.7**	**10.0**	**24.1**	33.2	**30.9**	**30.0**
MIMIC-CXR	*ST*	29.9	8.4	12.4	26.3	24.9	20.3	20.4
	TopDown	31.7	9.2	12.8	26.7	32.0	23.1	23.8
	MDT	35.3	10.3	14.2	**27.7**	33.3	27.3	27.6
	MDT+WCL	**37.3**	**10.7**	**14.4**	27.4	**38.5**	**27.4**	**29.4**

vised contrastive loss:

$$\mathcal{L}_{loss} = (1-\lambda)\mathcal{L}_{CE} + \lambda\mathcal{L}_{WCL}, \tag{2.11}$$

where λ is a hyperparameter that weighs the two losses.

2.2.3 Experiments

Datasets We conduct experiments on two datasets: (1) MIMIC-ABN, which was proposed in [68] and contains a subset of images of MIMIC-CXR with abnormal sentences only, with 26,946/3,801/7,804 reports for train/val/test sets. (2) MIMIC-CXR [39], the largest radiology dataset to date that consists of 222,758/1,808/3,269 reports for train/val/test sets.

Evaluation Metrics Performance is first evaluated on three automatic metrics: BLEU [76], ROUGE-L [55], and METEOR [19].

We then use the CheXpert labeler to evaluate the clinical accuracy of the abnormal findings reported by each model, which is a state-of-the-art rule-based chest X-ray report labeling system [36]. Given sentences of abnormal findings, CheXpert will give a positive and

Table 2.3. Generated samples from MIMIC-CXR dataset.

GT: there is moderate amount of right-sided subcutaneous emphysema which is similar in appearance compared to prior. right-sided chest tube is again visualized. there is no increase in the pneumothorax. bilateral parenchymal opacities are again visualized and not significantly changed. the tracheostomy tube is in standard location. right subclavian line tip is in the mid svc. **WCL**: tracheostomy tube tip is in unchanged position. right-sided port-a-cath tip terminates in the low svc. left-sided port-a-cath tip terminates in the proximal right atrium unchanged. heart size is normal. mediastinal and hilar contours are similar. innumerable bilateral pulmonary nodules are re- demonstrated better assessed on the previous ct. small right pleural effusion appears slightly increased compared to the prior exam. small left pleural effusion is similar. no new focal consolidation or pneumothorax is present. there are no acute osseous abnormalities.
GT: the lungs are mildly hyperinflated as evidenced by flattening of the diaphragms on the lateral view. diffuse interstitial markings compatible with known chronic interstitial lung disease are unchanged. there is no pleural effusion or evidence of pulmonary edema. there is no focal airspace consolidation worrisome for pneumonia. mild to moderate cardiomegaly is unchanged. the mediastinal and hilar contours are unremarkable. a coronary artery stent is noted. there is. levoscoliosis of the thoracic spine . **WCL**: lung volumes are low. heart size is mildly enlarged. the aorta is tortuous and diffusely calcified. crowding of bronchovascular structures is present without overt pulmonary edema. patchy opacities in the lung bases likely reflect areas of atelectasis. no focal consolidation pleural effusion or pneumothorax is present. there are no acute osseous abnormalities.

negative label for 14 diseases. We then calculate the Precision, Recall and Accuracy for each disease based on the labels obtained from each model's output and from the ground-truth reports.

Performance comparison We compare our approach to other methods on two datasets. As shown in Table 2.2, first, our method (*MDT+WCL*) outperforms previous retrieval (*CVSE*) and generation based models (*MDT*) on most text generation metrics. Second, our contrastive loss significantly improves clinical efficacy metrics, demonstrating its capability to accurately report abnormal findings. Finally, the relative difference between *MDT* and *MDT+WCL* is higher on MIMIC-CXR, which contains a larger training set for learning robust representations.

Generated Samples We present generation results from our model in Table 2.3.

2.3 Personalized Text Generation

2.3.1 Motivation

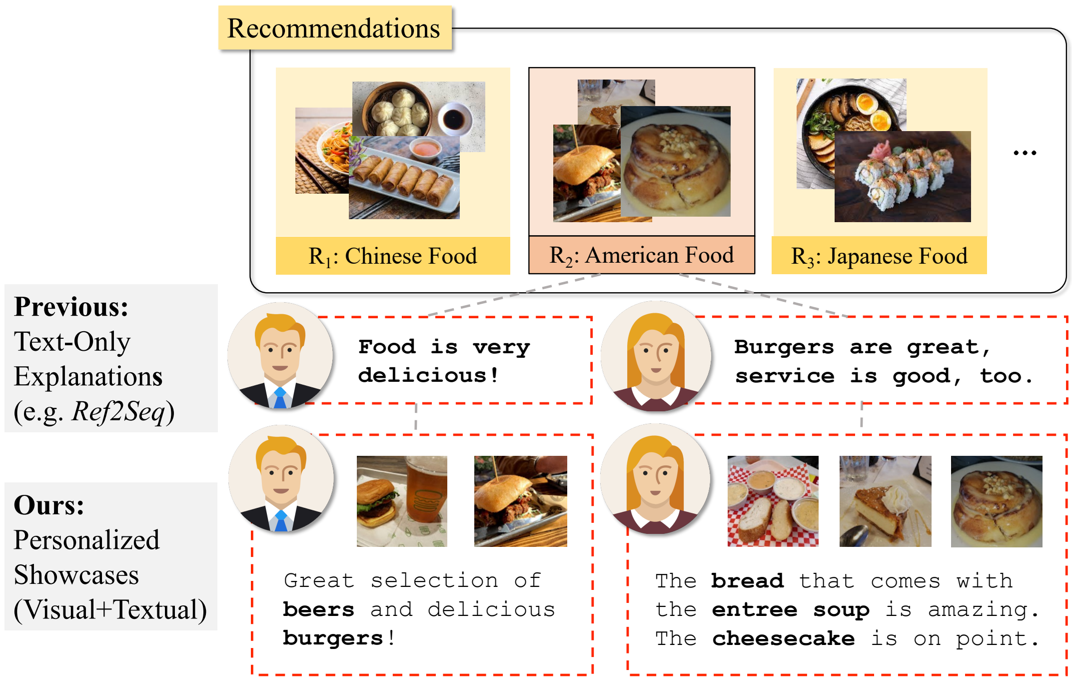


Figure 2.3. Illustration of previous text-only explanation and our personalized showcases for recommendations. Given a recommended item or business: (1) Text-only Explanation models only use historical textual reviews from user and item sides to generate textual explanations. (2) We propose a personalized showcases task to enrich the personalized explanations with multi-modal (visual and textual) information, which can largely improve the informativeness and diversity of generated explanations.

Personalized explanation generation models have the potential to increase the transparency and reliability of recommendations [106, 107]. Previous works [129, 122] considered generating textual explanations from users' historical reviews, tips [53] or justifications [69]. However, these methods still struggle to provide diverse explanations because a large amount of general sentences (e.g., 'food is very good!') exist in generated explanations and the text generation models lack grounding information (e.g., images) for their generation process. To further diversify and enrich explanations for recommendations, we propose a new explanation generation task named *personalized showcases* (shown in Figure 2.3). In this new task, we explain recommendations via both textual and visual information. Our task aims to provide a set of images that are relevant to a user's interest and generate textual explanations accordingly. Compared to previous works that generate only text as explanations, our showcases present diverse explanations including images and visually-guided text.

To this end, the first challenge of this task is building a *dataset*. Existing review datasets (e.g., Amazon [69] and Yelp) are largely unsuitable for this task with low diversity and insufficient image data. Thus, we first construct a large-scale multi-modal dataset, namely GEST, which is collected from Google Local Restaurants including review text and corresponding pictures. Then, to improve the quality of GEST for personalized showcases, we annotate a small subset to find highly matched image-sentence pairs. Based on the annotations, we train a classifier with CLIP [79] to extract visually-aware explanations from the full dataset. The images and text explanations from users are used as the learning target for personalized showcases.

For this new task, we design a new multi-modal explanation framework. To begin with, the framework selects several images from historical photos of the business that the user is most interested in. Then, the framework takes the displayed images and users' profiles (e.g., historical reviews) as inputs and learns to generate textual explanations with a multi-modal decoder. However, generating expressive, diverse and engaging text that will capture users' interest remains a challenging problem. First, different from previous textual explanation generation, the alignment between multiple images and generated text becomes an important problem for showcases, which poses higher requirements for information extraction and fusion across modalities. Second, a typical encoder-decoder model with a cross-entropy loss and teacher forcing can easily lead to generating repetitive and dull sentences that occur frequently in the training corpus (e.g., "food is great") [35].

To tackle these challenges, we propose a **P**ersonalized **C**ross-Modal **C**ontrastive **L**earning (PC^2L) framework by contrasting input modalities with output sequences. Contrastive learning has drawn attention as a self-supervised representation learning approach [72, 10]. However, simply training with negative samples in a mini-batch is suboptimal [46] for many tasks, as the randomly selected embeddings could be easily discriminated in the latent space. Hence, we first design a cross-modal contrastive loss to enforce the alignment between images and output explanations, by constructing hard negative samples with randomly replaced entities in the output. Motivated by the observation that users with similar historical reviews share similar interests, we further design a personalized contrastive loss to reweight the negative samples based on their history similarities. Experimental results on both automatic and human evaluation show that our model is able to generate more expressive, diverse and visually-aligned explanations compared to a variety of baselines.

2.3.2 A Large-Scale Dataset from Google Maps

Dataset Collection Our dataset is constructed from *Google Local* (i.e., maps) using a breadth-first-search algorithm with memorization. After collecting the review data, we filtered out reviews of length less than 5 words, which are less likely to provide useful information; we also removed reviews (2.13%) containing more than 10 images. We processed our dataset into two subsets as (1) GEST-s1 for personalized image set selection, and (2) GEST-s2 for visually-aware explanation generation. Statistics of our processed dataset are in Table 2.4,

Visual Diversity Analysis To distinguish our GEST from existing review datasets and show the usefulness of *personalized showcases*, we first define CLIP-based dis-similarity in three levels to measure the diversity of user-generated images in each business. Then, we compare the visual diversities between our GEST data with two representative review datasets, Amazon Reviews [62, 69] and Yelp.

First, similar to [79], we use the cosine similarity (denoted as *sim*) from pre-trained CLIP to define the dis-similarity between image i_m and i_n as $dis(i_m, i_n) = 1 - sim(i_m, i_n)$. Thus, we introduce visual diversity in three levels as *Intra-Business Div*, *Inter-User Div* and *Intra-User Div*, with higher scores meaning more visual diversity.

Then, we investigate the visual diversities for our GEST data as well as Amazon Reviews (using all categories *All* (A) and subcategories *Beauty* (B), *Clothing* (C), *Electronics* (E)) and

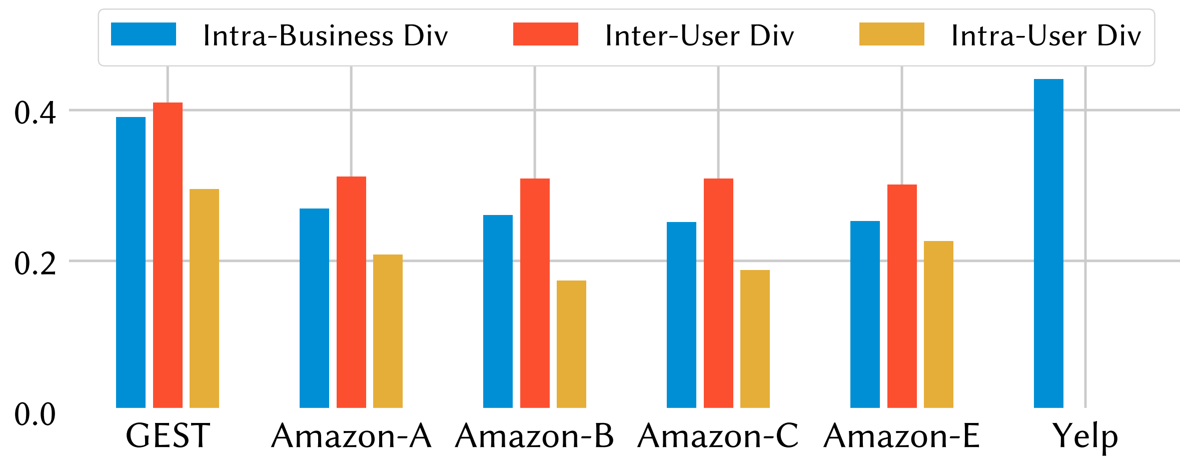


Figure 2.4. Visual Diversity Comparison with Amazon and Yelp. A, B, C, E in Amazon denote different categories of amazon review datasets, which are uniformly sampled from *All*, *Beauty*, *Clothing* and *Electronics*, respectively. Intra-/Inter- User Diversity for the Yelp dataset is unavailable since Yelp images lack user information.

Yelp. For Amazon, we treat each item page as a "business" because reviews are collected according to items. In our calculation, we sample 5,000 items with more than one user-uploaded image. Note that images in Yelp dataset do not have user information, so we cannot calculate user-level diversities for Yelp. From Figure 2.4, we have the following observations:

- **Diversities within datasets:** Figure 2.4 shows that for GEST and Amazon, *Inter-User Div* is the highest and *Intra-User Div* is the lowest. It indicates even for the same business (item), users focus on and present different visual information.

- **GEST *vs.* Amazon:** In Figure 2.4, three visual diversities of Amazon are consistently lower than GEST by a large margin. We try to explain this by discussing the difference of user behaviors on these two platforms. User-generated images on Amazon usually focus on the purchased item. Though the information they want to show differs, there is usually a single object in an image (i.e., the purchased item). Thus visual diversity is limited. While for GEST, as examples in Figure 2.5 show, reviews on restaurants allow users to share more diverse information from more varied items, angles or aspects. Compared with Amazon, using GEST should generate more informative *personalized showcases* according

Table 2.4. Data statistics for GEST. Avg. R. Len. denotes average review length and #Bus. denotes the number of Businesses. -raw denotes raw GEST. -s1 denotes GEST data for the first step, and -s2 denotes GEST data for the second step of our proposed framework.

Dataset	#Image	#Review	#User	#Bus.	Avg. R. Len.
GEST-raw	4,435,565	1,771,160	1,010,511	65,113	36.26
GEST-s1	1,722,296	370,563	119,086	48,330	45.48
GEST-s2	203,433	108,888	36,996	30,831	24.32

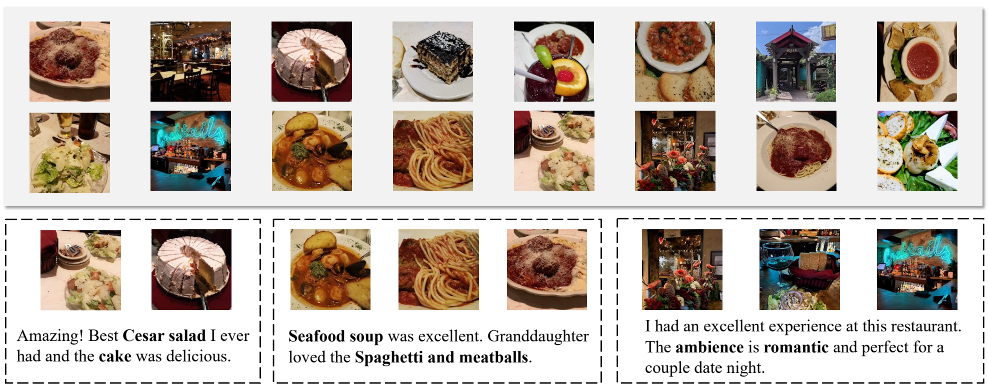


Figure 2.5. Example of business and user reviews in GEST. For a business (e.g., an Italian restaurant), GEST contains historical reviews and images from different users.

to different user profiles.

- **GEST *vs.* Yelp:** Yelp images are high-quality and the *intra-business div.* is higher (0.44) than GEST (0.39). Images in Yelp themselves are similar to images in GEST. However, Yelp images do not fit our task due to the lack of user information for personalization.

2.3.3 Method: Personalized Cross-Modal Contrastive Learning

Personalized Image Set Selection Our framework overview is in Figure 2.6. The first step of our framework is to select an image set as a visual explanation that is relevant to a user's interests, and is diverse. We formulate this selection step as diverse recommendation with multi-modal inputs.

Multi-Modal Encoder. We use CLIP [79], a state-of-the-art pre-trained cross-modal retrieval model as both textual- and visual-encoders. CLIP encodes raw images as image features,

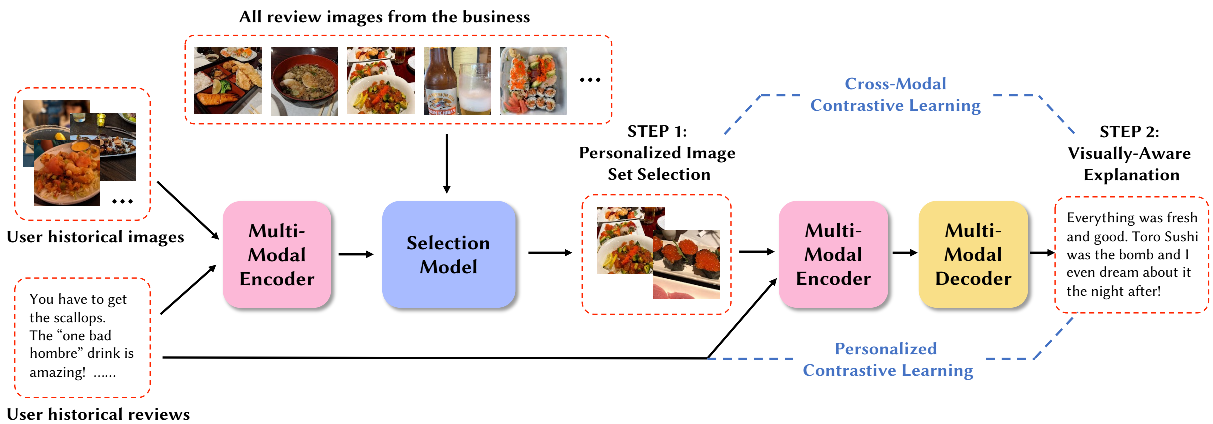


Figure 2.6. Illustration of our *personalized showcases* framework for a given business. We take user historical images and textual reviews as inputs. First, we select an image set that is most relevant to a user's interest. Then we generate natural language explanations accordingly with a multi-modal decoder. A cross-modal contrastive loss and a personalized contrastive loss are applied between each input modality and the explanations. Last, the selected images and generated textual explanations will be organized as multi-modal explanations to users.

and encodes user textual- and visual-profiles as user profile features.

Image Selection Model. We use a Determinantal Point Process (DPP) [45] to select the image subset, which has recently been used for different diverse recommendation tasks [104, 4]. Compared with other algorithms for *individual* item recommendation, DPP-based models are suitable for *multiple* image selection. Given user u and business b, we predict the image set $\hat{I}_{u,b}$ as follows:

$$\hat{I}_{u,b} = \mathrm{DPP}(I_b, u), \tag{2.12}$$

where I_b is the image set belonging to business b. In our design, we calculate user-image relevance using the CLIP-based user's profile features and image features. More details of the model are in [104].

Visually-Aware Explanation Generation After obtaining an image set, we aim to generate personalized explanations given a set of images and a user's historical reviews. Specifically, we build a multi-modal encoder-decoder model with GPT-2 [80] as the backbone.

Multi-Modal Encoder. Given a set of user u's (we omit the subscript u below for simplicity) historical reviews $X = \{x_1, x_2, \ldots, x_K\}$, we use the text encoder of CLIP to extract the review features $R = \{r_1, r_2, \ldots, r_K\}$. Similar operations are applied to the input images $I = \{i_1, i_2, \ldots, i_n\}$, where we use the CLIP visual encoder to extract visual features $V = \{v_1, v_2, \ldots, v_n\}$. Those features are then projected into a latent space:

$$Z_i^V = W^V v_i, Z_i^R = W_i^R r_i, \tag{2.13}$$

where W^V and W^R are two learnable projection matrices. Then we use a multi-modal attention (MMA) module with stacked self-attention layers [95] to encode the input features:

$$[H^V; H^R] = \mathrm{MMA}([Z^V; Z^R]), \tag{2.14}$$

where each H_i^V, H_i^R aggregate features from two modalities and $[;]$ denotes concatenation. This flexible design allows for variable lengths of each modality and enables interactions between modalities via co-attentions.

Multi-Modal Decoder. Inspired by recent advances of pre-trained language models, we leverage GPT-2 as the decoder for generating explanations. To efficiently adapt the linguistic knowledge from GPT-2, we insert the encoder-decoder attention module into the pre-trained model with a similar architecture in [12]. With this multi-modal GPT-2, given a target explanation $Y = \{y_1, y_2, ..., y_L\}$, the decoding process at each time step t can be formalized as:

$$\hat{y}_t = \mathrm{Decoder}([H^V; H^R], y_1, \ldots, y_{t-1}). \tag{2.15}$$

We use a cross-entropy (CE) loss to maximize the conditional log likelihood for N training samples $(X^{(i)}, I^{(i)}, Y^{(i)})_{i=1}^N$ as follows:

$$\mathscr{L}_{CE} = -\sum_{i=1}^{N} \log p_\theta(Y^{(i)} | X^{(i)}, I^{(i)}). \tag{2.16}$$

We use ground truth images from the user for training and images from our image-selection model for inference.

Personalized Cross-Modal Contrastive Learning Unlike image captioning tasks [105, 110] which mainly describe images, our task use multiple images as "visual prompts" to express personal feelings. To encourage expressive and visual-aligned generations, we propose **P**ersonalized **C**ross-Modal **C**ontrastive **L**earning (PC^2L). We first project the embeddings of images H^V, historical reviews H^R, and the target sequence H^Y into a latent space:

$$\tilde{H}^V = \phi_V(H^V),\ \tilde{H}^R = \phi_R(H^R),\ \tilde{H}^Y = \phi_Y(H^Y) \tag{2.17}$$

where ϕ_V, ϕ_R, and ϕ_Y consist of two fully connected layers with ReLU activation and average pooling over the hidden states H_V, H_R and H_Y from the last self-attention layers. With the InfoNCE loss [72, 10], we then maximize the similarity between the pair of source modality and target sequence, while minimizing the similarity between the negative pairs as follows:

$$\mathscr{L}_{CL} = -\sum_{i=1}^{N} \log \frac{\exp(s_{i,i}^{X,Y})}{\exp(s_{i,i}^{X,Y}) + \sum_{j \in K} \exp(s_{i,j}^{X,Y})}, \tag{2.18}$$

where $s_{i,j}^{X,Y} = sim(\tilde{H}_{(i)}^X, \tilde{H}_{(j)}^Y)/\tau$, *sim* is the cosine similarity between two vectors, τ is the temperature parameter, (i) and (j) are two samples in the mini-batch, K is the set of negative samples for sample (i).

One challenge of this task is the model is asked to describe multiple objects in a set of images [99]. To ensure the visual grounding between multiple image features and output text, we design a novel cross-modal contrastive loss. Specifically, given a target explanation $Y = \{y_1, y_2, ..., y_L\}$, we randomly replace the entities in the text with other entities presented in the dataset to construct a hard negative sample $Y^{ent} = \{y'_{ent1}, y_2, ...y'_{ent2}, ...y_L\}$ (i.e., "I like the sushi" to "I like the burger"), such that during training, the model is exposed to samples with incorrect entities regarding the images, which are non-trivial to distinguish from the original target sequence. Thus, we add the hidden representation of Y^{ent} as an additional negative sample *ent* to formulate the cross-modal contrastive loss:

$$\mathscr{L}_{CCL} = -\sum_{i=1}^{N} \log \frac{\exp(s_{i,i}^{V,Y})}{\exp(s_{i,i}^{V,Y}) + \sum\limits_{j \in K \cup ent} \exp(s_{i,j}^{V,Y})}, \tag{2.19}$$

On the other hand, to enhance the personalization of explanations, we re-weight negative pairs according to user personalities. The intuition is that users with more distinct personalities are more likely to generate different explanations. Motivated by this, we propose a weighted personalized contrastive loss:

$$\mathscr{L}_{PCL} = -\sum_{i=1}^{N} \log \frac{\exp(s_{i,i}^{R,Y})}{\exp(s_{i,i}^{R,Y}) + f(i,j) \sum\limits_{j \in K} \exp(s_{i,j}^{R,Y})}. \tag{2.20}$$

where negative pairs in a mini-batch are re-weighted based on user personality similarity function f. In our framework, user personalities are represented by their historical reviews. Specifically, we define f function as:

$$f(i,j) = \alpha^{(1 - sim(\tilde{R}_{(i)}, \tilde{R}_{(j)}))} \tag{2.21}$$

i.e., we reduce the weights of negative pairs with similar histories, and increase those with distinct histories. α ($\alpha > 1$) is a hyperparameter that weighs the negative samples, *sim* is the cosine similarity, $\tilde{R}_{(i)}$ and $\tilde{R}_{(j)}$ are the average features of two users' input historical reviews.

Overall, the model is optimized with a mixture of a cross-entropy loss and the two contrastive losses:

$$\mathscr{L}_{loss} = \mathscr{L}_{CE} + \lambda_1 \mathscr{L}_{CCL} + \lambda_2 \mathscr{L}_{PCL}, \tag{2.22}$$

where λ_1 and λ_2 are hyperparameters that weigh the two losses.

Table 2.5. Performance comparison. Results are in percentage (%). *GT* is ground truth.

Model	Input	N-Gram Metrics				Diversity Metrics		Embedding Metrics	
		BLEU-1	BLEU-4	METEOR	NIST	DISTINCT-1	DISTINCT-2	CLIP-SCORE	BERT-SCORE
GT	-	-	-	-	-	6.06	43.23	28.41	-
ST	*img*	8.24	0.28	3.41	28.08	2.74	17.41	24.31	85.20
R2Gen	*img*	6.47	0.22	3.10	36.55	3.23	22.45	24.28	85.89
Ref2Seq	*text*	7.09	**0.67**	3.80	30.78	0.92	5.89	23.83	84.71
Peter	*text*	8.89	0.44	3.28	34.45	0.38	1.27	23.27	86.94
Ours	*img*	9.92	0.32	3.64	37.35	3.37	26.37	**24.68**	88.03
	img+text	**10.40**	0.36	**3.83**	**50.64**	**3.58**	**28.58**	24.50	**88.23**

2.3.4 Experimental Results

Baselines. We compare our model with popular baselines from different tasks, including image captioning, report generation and explanation generation:

- *ST* [105] is a classic CNN+LSTM model for image captioning.
- *R2Gen* [12, 108] is a state-of-the-art memory-driven transformer specialized at generating long text with visual inputs.
- *Ref2Seq* [69] is a popular reference-based seq2seq model.
- *Peter* [52] is a recent transformer-based model which uses the user and item IDs to generate explanation.
- *img* and *text* refer to image and text features respectively.

For textual explanations, we first evaluate with n-gram metrics: BLEU (n=1,4) [76], METEOR [19] and NIST (n=4) [21]. For diversity, we report DINSTINCT [49]. CLIP-SCORE [33], BERT-SCORE [124] are two embedding metrics for visual alignment and semantic quality.

Performance Comparison Results for text explanation generation are presented in Table 2.5. First, the clear gap between text-input models and image-input models on diversity metrics validates the benefits of incorporating visual features. The setting of visually-aware generation is able to generate accurate and diverse explanations. Second, PC^2L shows substantial improvement on most of the metrics compared to other models. Though text-based models *Ref2Seq* and *Peter* achieve competitive results with our method on some n-gram metrics such as BLEU, their performance is much worse on diversity and embedding metrics.

Generation Results We present generated samples in Figure 2.7, and compare our personalized showcases to single-modal explanations from *Ref2Seq* and *Text GPT-2*. Overall, our visual explanations is able to recommend images that fit users' interest. This indicates the effectiveness of our image selection module and the selected images can be used as valid visual explanations. More importantly, these images can provide grounding information for text generation such that the textual explanations become more informative (i.e., specific dishes), which aligns with automatic metrics as well as human evaluation.

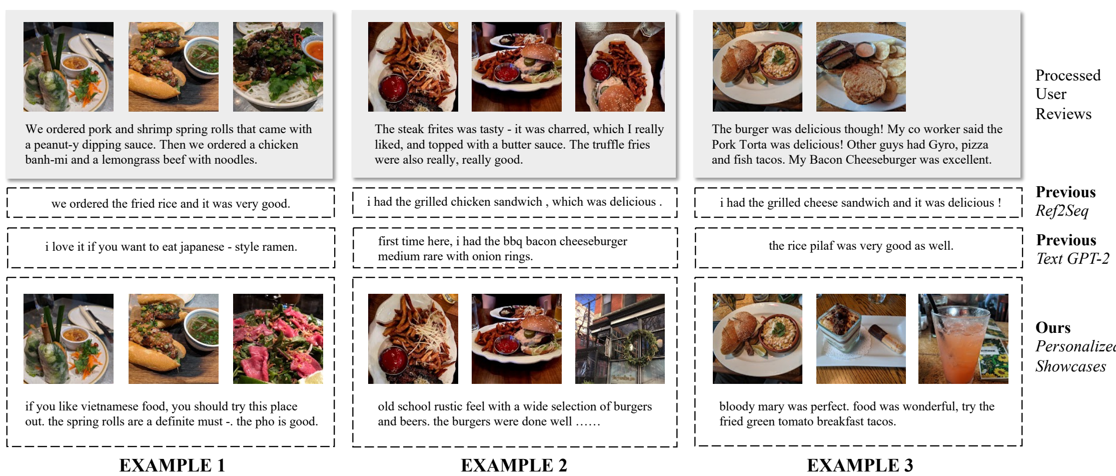


Figure 2.7. Comparison between text-only explanations (i.e., *Ref2Seq* and *Text GPT-2*) and our personalized showcases.

2.4 Visual-Augmented Text Generation

2.4.1 Motivation

In previous sections, we introduced three different image captioning tasks, where the model is asked to generate text conditioned on images. But what happens to pure natural language generation tasks? One great resource human writers cherish is the ability of imagination, with

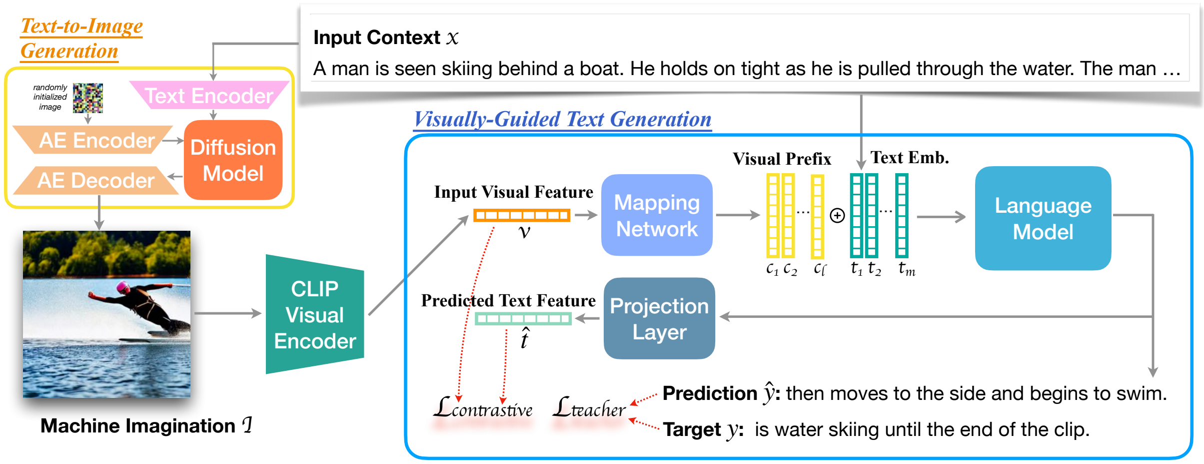


Figure 2.8. An overview of our model. Given an input context x, we first visualize the context with the text-to-image generation model. Then we use the machine-generated image I as the additional visual supervision to guide the language model in open-ended text generation.

which they render mental images about an actual or vicarious experience and link knowledge that would later make the writing more concrete, sensible, and intriguing. Cognitive studies show that visual imagery improves comprehension during language processing [26, 86], and that mental imagery facilitates humans' written language expression at young ages [27]. Inspired by such a cognitive process in humans, we ask the research question of whether we can endow machines with the same ability to utilize visual information and construct a general picture of the context to guide text generation.

Recent advances in text-to-image synthesis [83] make it possible to visualize imaginations for machines given some context. In the following section, we will introduce a way to generate machine imaginations and use them to augment

2.4.2 Method: Imagination-Guided Open-Ended Text Generation

Text-to-Image Rendering In this work, we propose to use images generated conditioning on the context by the machines as additional visual information to the LM. The text-to-image generation backbone is StableDiffusion [83], which mainly consists of a text encoder, a diffusion model, and an autoencoder. The text encoder is from the frozen CLIP ViT-L/14 [79] and encodes the input text to textual embeddings. The diffusion model uses UNet [84] to provide noise estimation. The UNet is modified so as to attend to the input textual embeddings. The encoder of the pretrained autoencoder encodes images into the lower-resolution latent maps z_T. At each step t, the diffusion model provides the noise estimation ε and modifies z_t correspondingly. The decoder of the pretrained autoencoder takes the final noise-free latent map z and generates the image prediction. StableDiffusion is trained with LAION-5B [88].

Visual Prefix Construction One can encode the visual information with the pre-trained visual models. However, such visual embedding may lie in a representation space different from the LM due to the discrepancy between models. One way of introducing features extracted by another network to the current model is through feature mapping [64]. With a dataset of image-text pairs (I', x'), we can pre-train a mapping network F for a given LM in an image captioning formulation. More specifically, we encode I' with the visual encoder $\text{Enc}_{\text{visual}}$ and receive its visual features v'. Then we apply the mapping network F over v', and receive a sequence of l visual prefixes:

$$c'_1, c'_2, \ldots, c'_l = F(v') = F(\text{Enc}_{\text{visual}}(I')) \tag{2.23}$$

We provide the list of visual prefix as input to the LM with the corresponding text x' as the target output. Such a pre-training process enables F to project visual features into the visual prefix that lies within the same embedding distributions as the LM. The mapping network is agnostic of the downstream task, and only depends on the visual source and the LM.

After generating a descriptive image I^i for the input context x^i, we use CLIP to encode I^i and receive its visual features v^i. We apply the pre-trained mapping network F over v^i, and receive the visual prefix c^i of length l:

$$c^i = \{c^i_1, c^i_2, \ldots, c^i_l\} = F(\text{CLIP}(I^i)) \tag{2.24}$$

Visually-guided Language Modeling We use the visual information to guide text generation in two ways, reflected in the following two training objectives. Firstly, we directly introduce the machine-generated visual information as input to the LM. We concatenate the visual prefix c^i and the text embeddings t^i for the input context x^i with m tokens. LM input can be denoted as $[c^i;t^i] = \{c^i_1,\ldots,c^i_l,t^i_1,\ldots,t^i_m\}$. With $y^i = \{y^i_1,y^i_2,\ldots,y^i_n\}$ denoting the target output of n tokens, and θ denoting the trainable parameters, we can list out the teacher forcing training objective as follows:

$$L_{\text{teacher}} = -\sum_{j=1}^{n} \log p_\theta(y^i_j|c^i;t^i;y^i_{<j}) \tag{2.25}$$

In addition, we leverage a contrastive objective to enforce the generated text to be semantically similar to the input visual supervision with the InfoNCE loss [72, 108]:

$$L_{\text{contrastive}} = -\log \frac{\exp(\text{sim}(v^i,\hat{t^i})/\tau)}{\sum_{j\neq i}\exp(\text{sim}(v^i,\hat{t^j})/\tau)} \tag{2.26}$$

in which $\hat{t}$ is the projected representation of the decoder's last layer's output, and can be viewed as the sentence-level representation of the generated text. Here $\text{sim}(\cdot,\cdot)$ first normalizes the two vectors, then compute their cosine similarity, and τ is the temperature. An overview of our model is in Figure 2.8.

Training & Inference

We first pre-train the mapping network on the pre-training dataset with the teacher-forcing objective. Such pre-training is agnostic of the downstream task, and only depends on the type of base LM.

When applying our model on downstream tasks, we train the base LM with the teacher forcing objective for the first $N_{\text{no\_contra}}$ epochs. Then, we introduce the contrastive objective and tune the base LM together with the mapping network and projection layer by minimizing the following loss L. Here ep denotes the epoch and λ is the factor:

$$L = \begin{cases} L_{\text{teacher}}, & ep < N_{\text{no\_contra}}, \\ L_{\text{teacher}} + \lambda L_{\text{contrastive}}, & ep > N_{\text{no\_contra}}, \end{cases} \tag{2.27}$$

During inference, we provide the context and machine-generated image to the LM. We use beam search during decoding with a beam width of 10.

2.4.3 Experimental Results

Sentence Completion is a task of finishing the sentence in a commonsense inference scenario. We conduct experiments on the ActivityNet [32] subset of HellaSwag [123], which is a benchmark for commonsense natural language inference that ask the model to predict the most likely follow-up among several choices given a specific context. We compare with StoryEndGen [29] which encodes the given context incrementally and attends to the one-hop knowledge graph retrieved from ConceptNet for the context tokens. We implement our method on top of the GPT-2 [80], which by nature, can generate the follow-up for an arbitrary input in a zero-shot manner.

Story Generation requires the model to compose a story based on the given title or context. We conduct experiments on the widely used story generation benchmark ROCStories [65]. Each data item consists of a story title and a human-written five-sentence everyday life story that incorporates commonsense related to the title. We provide the story title and the story's first sentence as the input context, and ask the LM to predict the following four sentences. We consider the following methods as baselines: Action-Plan [24] first predicts the premise of a story with the convolutional LM [15], then use fusion mechanism to encourage a convolutional seq2seq model [28] to generate the story from the premise. Plan-and-Write [118] first plans a storyline that consists of keywords, then generate the story conditioned on the storyline. SimCTG [92] proposes a contrastive training objective that encourages the LM to learn discriminative and

Table 2.6. Generation quality scores for few-shot text completion on the ActivityNet and few-shot story generation on ROCStories. "Human" shows the human performance and "GPT2 *no finetune*" denotes the vanilla GPT2 model without tuning. All the other listed models are trained with 1% of the training data. "+ours" denotes introducing machine-generated images on top of the base LM.

Task	*	Setting	rep-2 ↓	rep-3 ↓	rep-4 ↓	diversity ↑	distinct-2 ↑	MAUVE↑	BERTScore↑
Sentence Completion	0	Human	0.45	0.05	0.01	99.50	77.32	-	-
	1	GPT2 *no finetune* [80]	6.71	6.87	10.13	78.07	74.83	44.19	22.57
	2	StoryEndGen [29]	39.53	35.11	39.30	34.12	44.57	0.45	-47.29
	3	GPT2 *text-only finetune*	4.20	4.03	5.53	86.85	75.14	49.45	24.13
	4	GPT2 +ours	**2.43**	**2.61**	**3.57**	**91.63**	**75.92**	**60.30**	**24.25**
Story Generation	5	Human	1.76	0.38	0.15	97.71	56.34	-	-
	6	GPT2 *no finetune*	37.65	22.76	21.92	45.67	43.42	0.43	-7.77
	7	Action-Plan [24]	52.05	35.58	28.11	26.97	21.43	0.41	-18.32
	8	Plan-and-Write [118]	45.22	32.86	23.34	30.71	20.83	0.41	-37.35
	9	SimCTG [92]	28.72	24.02	20.61	43.00	42.06	0.43	18.01
	10	GPT2 *text-only finetune*	25.41	18.51	14.41	52.10	46.60	9.10	21.23
	11	GPT2 +ours	**10.73**	**5.64**	**3.42**	**81.36**	**51.91**	**35.94**	**23.03**

isotropic token representations, and is implemented on GPT-2 [80].

Evaluation For sentence completion and story generation, we follow previous work and evaluate the quality of the generated text from the aspect of model degeneration level (rep-n, diversity, distinct-n), text distribution divergence (MAUVE), and semantic similarity (BERTScore): (1) rep-n = 1.0 - $\frac{|\text{unique } n\text{-grams}|}{|\text{total } n\text{-grams}|}$ measures sequence level repetition by computing the portion of duplicate n-grams [103]. (2) diversity = $\prod_{n=2}^{4}(1-\text{rep-}n)$ measures the diversity of n-grams [91]. (3) distinct-n = $\frac{|\text{unique } n\text{-grams}|}{|\text{length of text}|}$ measures the portion of distinct n-grams in the text [49]. (4) MAUVE measures the learned distributions divergence between the generated text and human-written text [78]. We report MAUVE with gpt2-large as the base model. A low MAUVE indicates a great difference between the distributions of generated text and human text. (5) BERTScore assesses contextual text similarity between two pieces of texts by computing the cosine similarities between their tokens' embeddings [124]. We report BERTScore with roberta-large as base model. A low BERTScore means the generated text is contextually different from the ground-truth.

Performance Comparison Open-ended text generation is a broad topic with flexible and inexhaustible setups, many of which have low resources. Collecting annotations is often extremely expensive and time-consuming. Therefore, we report few-shot results to check if our framework can rapidly adapt to new task setups with a few examples, which is more practical in real-life.

More specifically, we report few-shot open-ended text generation results with 1% of the training data. For sentence completion and story generation tasks, the base LM is GPT2-base [80]. For concept-to-text, we test it with BART-base [47] as the base LM.

For sentence completion, as shown in Table 2.6, StoryEndGen (#2) suffers from degeneration with the highest rep-n and the lowest diversity. Training with only 1% of the training data improves GPT2's performance on all metrics (#3 vs. #1). Under the same few-shot setting, adding additional machine-generated images with ours (#4) further alleviate model degeneration. The improvement on MAUVE also indicates that introducing visual input can aid GPT2 in generating text that is more similar to the human-written ones.

For story generation, as shown in Table 2.6, for the story generation task that requires the LM to compose longer text, we see the vanilla GPT2 without tuning suffering from more severe degeneration compared to rendering a sentence ending (#6 vs. #1). The high rep-n scores indicate that the two non-Transformer-based baselines Action-Plan (#7) and Plan-and-Write (#8) stammer with repetitive tokens, which greatly differs from the human-written text (leads to low MAUVE) and does not have concrete meanings (leads to low BERTScore). The models based on GPT-2 (#9-#10) yield more complete sentences with concrete meanings (BERTScore gets higher). However, they keep repeating the same sentence, which is still quite different from human language (MAUVE remains low). Applying our method to GPT-2 leads to minor degeneration and has the best performance on all metrics (#11).

Generated Samples We present generated results in Figure 2.9.

Context:	Two girls are standing in a yard wearing cheerleading outfits. A girl…

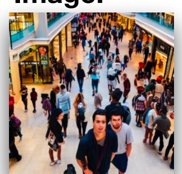
Generated Image:

GPT2 (text-only)	is *standing in a yard* with a group of girls.
GPT2 +iNLG	is wearing a white bikini and blue shorts.

(a) Sentence Completion

Context:	Barry at the Mall. Barry and his friends spent a day at the mall.

Generated Image:



SimCTG	Barry was at the mall with his friends. They were all very excited to see him at the mall. *They all had a great time at* the mall. *They all had a great time at* work. *They all had a great time at* school
GPT2 (text-only)	They went to the local mall and *bought a new car*. *They drove to the mall and bought a new car*. *They drove to the mall and bought a new car*. Barry's friends were very happy they *bought a new car*.
GPT2 +iNLG	They were swamped by the crowds. Barry was scared to go home. They tried to avoid the mall, but they were stopped by the people at the mall.

(b) Story Generation

Figure 2.9. Sentence ending and stories generated by GPT2-based methods tuned with 1% of the training data. *Repetitive contents* are underlined. The sentence ending and story written by our model is coherent with the context, related to the machine-generated image, and has minor degeneration. More demonstrative examples can be found in the Appendix.

2.5 Conclusions

In this chapter, we revisited some of the classical image captioning tasks via our previous work. These tasks present the broad application cases for models capable of generating text given images. For the technical takeaways, we show the benefit of multi-task learning for single-image and multiple-image captioning, which is still not fully solved in today's most advanced open-source vision-language models. We also show a possible issue of next token prediction that it tend to learn the head distributions from data, and the potential of incorporating contrastive learning into training language models to mitigate this problem. Finally, visual features can potentially benefit text generation, and vision (e.g., stable diffusion, DALLE [81]) and language foundation models (e.g., GPT-3, ChatGPT) may be unified together to create stronger models for vision language understanding and generation.

Chapter 2, in part, is a reprint of the material as it appears in the following publications:

“Describing Visual Differences Needs Semantic Understanding of Individuals” by An Yan, Xin Wang, Tsu-Jui Fu, William Wang, published at *European Chapter of the Association for Computational Linguistics*,2021 The dissertation author was the primary investigator and author of this paper.

“Weakly Supervised Contrastive Learning for Chest X-Ray Report Generation” by An Yan, Zexue He, Xing Lu, Jiang Du, Eric Chang, Amilcare Gentili, Julian McAuley, Chun-Nan Hsu, published at *Empirical Methods in Natural Language Processing*, 2021 The dissertation author was the primary investigator and author of this paper.

“Visualize Before You Write: Imagination-Guided Open-Ended Text Generation” by Wanrong Zhu, An Yan, Yujie Lu, Wenda Xu, Xin Eric Wang, Miguel Eckstein, William Wang, published at *European Chapter of the Association for Computational Linguistics*, 2023. The dissertation author was one of the primary authors of this paper.

“Personalized Showcases: Generating Multi-Modal Explanations for Recommendations” by An Yan, Zhankui He, Jiacheng Li, Tianyang Zhang, Julian McAuley, published at *The International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023. The dissertation author was the primary investigator and author of this paper.

Chapter 3

The Future of Vision-Language

3.1 Is GPT-4V all we need?

In the previous section, we talked about vision-language models with "small language decoders", from LSTM to BART and GPT-2. As of today, large multimodal models such as GPT-4V, has shown strong capabilities and generalizability for various tasks. It is able to complete all the tasks we mentioned above with much better performance than previous state-of-the-art models. Consequently, it makes us wonder, is GPT-4V all we need for image captioning or visual-conditioned text generation?

The answer is no. Even though GPT-4V is an amazing model, there are still issues to solve and long way to go along this direction. Researchers have shown there are still problems with GPT-4V [113] when being tested in various applications [116]. In our early exploration with GPT-4V, we also obeserved similar findings that indicate GPT-4V still needs improvement in certain aspects. We summarize some of the typical problems here:

1. Hallucination: GPT-4V may hallucinate, especially when the question or text is not best aligned with the images. For example, when the model is asked about some objects that do not exist in the current image, it tend to generate inaccurate and incorrect outputs.

2. Visual grounding: GPT-4V still has issues with visual grounding. It may fail to find the right location of some objects, even with the help of visual prompting.

3. Navigating dynamic environments: GPT-4V is good at single image understanding or even a few interleaved image-text inputs (with a current maximum of four images as inputs), but it still needs to improve reasoning and capability in a dynamic environment, for example, robotic operations in real-world, or smartphone navigation with a sequence of screens.

3.2 Future Work: Datasets

Data has been one of the most important building blocks in the era of deep learning or data-drive machine learning, and it has been more important than ever in the recent trend of Large Language Models. Unlike training LLMs, where supervised text data written by human is freely available online in a large-scale (yes, human-written text is supervised data), high-quality image-text data is rare and hard to collect.

One of the most famous and most large-scale image-text dataset is LAION-5B [88], which is the training source for many open-source text-image understanding models (e.g., openclip [13]) text-to-image diffusion models. However, there are two reasons we need to look for new data source: first, the data consists of single image-caption pairs, and we need interleaved image-text [130] in practice to train strong models. Second, the captions are noisy and loosely-connected to the images, even after filtering, the dataset is still suboptimal for training multimodal LLMs [59, 100].

More high-quality image-text datasets are needed to train and evaluate future vision-language models. Our assumption is that dense captioning with spatial information can be one final solution to connect vision and language, though it still needs verification from experiments. And in the process of creating these datasets, human annotation may be inevitable to some degree.

3.3 Future Work: Models

Model is another important building block other than data. Although since the invention of self-attention, the design choice of neural networks has been less important with transformers being all you need for deep representation learning, i.e., vision encoders and language decoders built with the transformer architecture has ruled over vision and language research [22, 20, 75]. There is still an on-going discussion on what could be the best architecture to efficiently connect vision-language. For open-sourced models, projection layers [59, 58, 100], cross-attention layers [1, 3], and visual tokens [50] have been proposed.

Compared with GPT-4V, one capability that current open-sourced models are missing is interleaved image-text understanding. There has been some effort along this direction [48, 119, 127], from model design to data preparation. However, we find these models still bad at describing visual differences [110] or perform multi-modal in-context learning.

Another less-discussed direction is efficiency, with multi-modal LLMs or video diffusion models, how do we efficiently understand long videos or sequences of images and vice versa? To this point, GPT-4V only supports four image inputs at once, how can we build models that can consume long sequences of pixels, e.g., watching a full movie or taking an online class, and write reviews or learn things as humans do?

With improved datasets and more explorations in the future, we will have a better understanding of model designs for vision-language.

3.4 Future Work: Applications

Large Multimodal models such as GPT-4V has shown promising generalizability in various applications, for example, autonomous driving [128], evaluating vision-language tasks [125], web navigation [128], etc.

We are among the earliest to explore new applications with GPT-4V, where we focused on smartphone GuI navigation [113]. Specifically, we use Set-of-Mark prompting [114] with self summarization to build a smartphone agent with GPT-4V. An high-level illustration of our framework is shown in Figure 3.1.

Figure 3.2 shows qualitative results of using GPT-4V to recursively process an episode

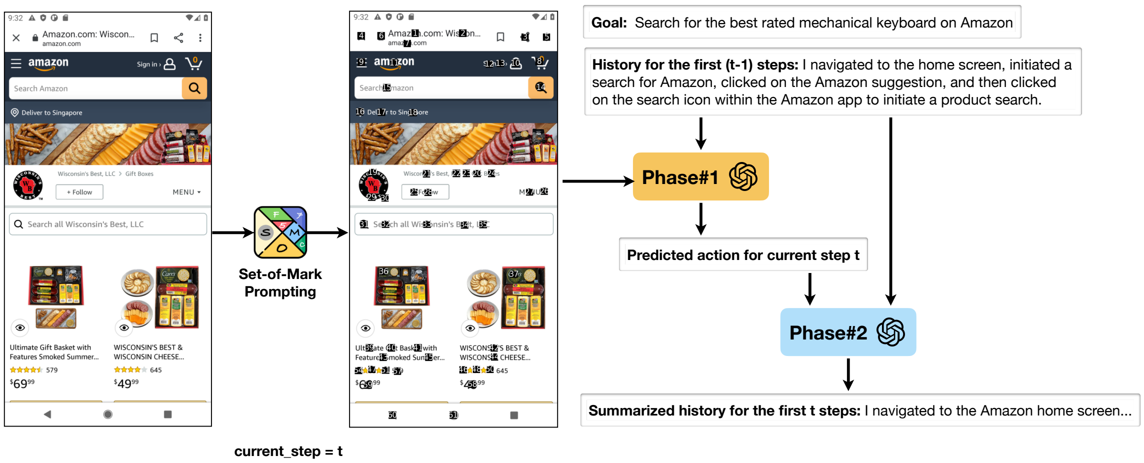


Figure 3.1. Illustration of our framework that builds smartphone navigation agents at one time step. The agent is asked to predict an action given current screen and history actions, then update the summarized history recursively.

of screen inputs for the task of shopping online. In each step, we include the objective, "You are asked to shop for a milk frother, your budget is between \$50 and \$100." and its previous action in the prompt to GPT-4V. We show that the model can effectively perform multi-step reasoning to accomplish the given instruction. Large Multimodal models like GPT-4V has the potential to perform many daily tasks for human from driving on streets to navigating cellphones and the web.

Other than building multimodal agents to assist humans, large visual generative models [85, 74] which can create images or videos following human instructions, also have the potential to revolutionize content generation. A recent model, Sora, is able to generate videos up to a minute long with high visual quality. The potential of simulating the physical world with machine learning models is unlimited.

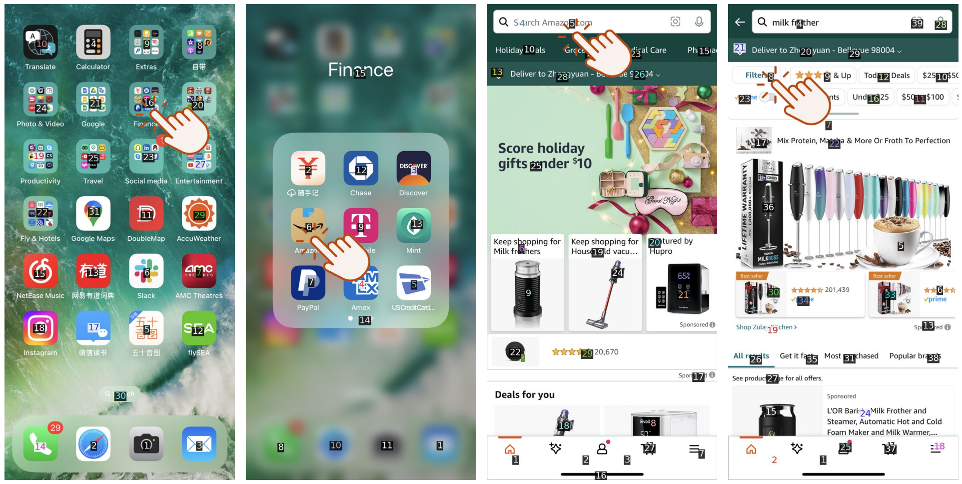
Overall, We believe there are many more exciting applications with large multimodal models for us to explore and to improve.

Instruction query:

You are asked to shop for a milk frother, your budget is between $50 and $100.

Prompt for all steps:

Imagine that you are a robot operating an iPhone.
You are asked to shop for a milk frother, your budget is between $50 and $100.
Your previous action is:
… .
Below is what you see on the screen, with numeric IDs added. For the next action, which ID would you click on the screen next?

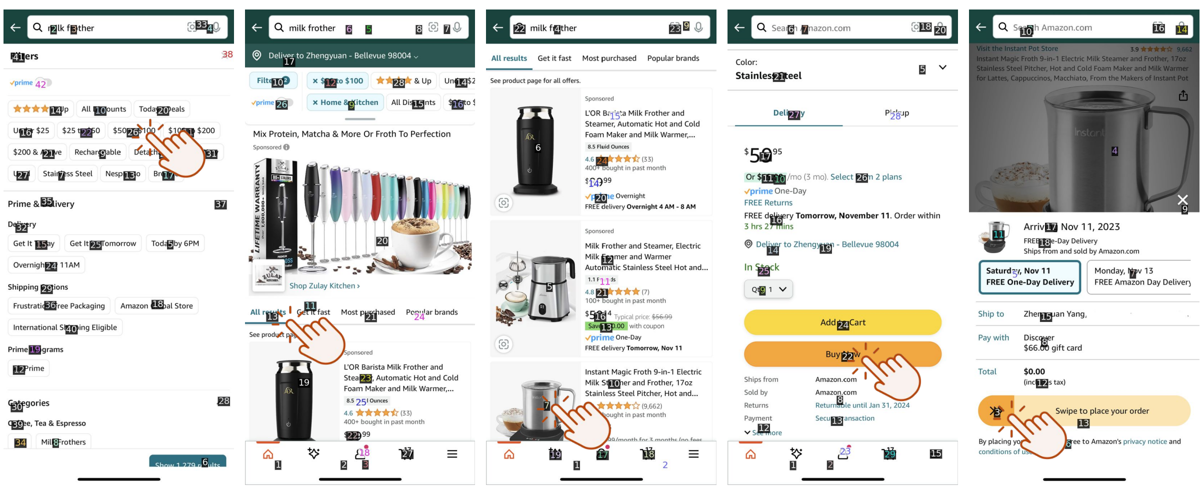


(1) GPT-4V:
To shop for a milk frother within the specified budget using the iPhone interface shown in the image, I would select an application that is capable of online shopping. In this scenario, I would click on ID 16 to use the Amazon app, as it is a comprehensive and user-friendly platform for purchasing a wide variety of products, including kitchen appliances like milk frothers.

(2) GPT-4V:
To proceed with shopping for a milk frother within the specified budget using the Amazon app, the next step would be to click on ID 6, which represents the Amazon app on the screen. This will open the app and allow me to search for milk frothers within the budget range.

(3) GPT-4V:
I would click on the search bar at the top of the Amazon app interface, which is typically where you would enter a search query. This search bar is labeled with ID 5. By clicking on this, I can then type in "milk frother" and set the price filter to show results that are between $50 and $100.

(4) GPT-4V:
The next step would be to refine the search results by setting the price filter. In this interface, the ID most likely to represent the filter or sorting options is 8 labeled "Filters," which seems to be the right place to start refining the search results according to the budget. So the next action would be to click on ID 8 to access filtering options, including setting the desired price range.



(5) GPT-4V:
The logical next step would be to click on the price filter that matches your budget range. Looking at the screen, you would choose the price filter that includes $50 to $100. The closest available options are ID 25 for "$25 to $50" and ID 26 for "$50 to $100". Since your budget starts at $50, you would click on ID 26 to filter the search results to show only milk frothers that are priced between $50 and $100.

(6) GPT-4V:
The next action would be to click on ID 13, "All results," to view all available milk frothers within the selected price range of $50 to $100.

(7) GPT-4V:
The next action would be to click on ID 7, which shows the "Instant Magic Froth 9-in-1 Electric Milk Steamer and Frother" listed at $59.95. This item is within the specified budget of $50 to $100.

(8) GPT-4V:
The next action would be to click on ID 22, which is the "Buy Now" button for the item that was selected in the previous action, the "Instant Magic Froth 9-in-1 Electric Milk Steamer and Frother" priced at $59.95.

(9) GPT-4V:
The next action would be to click on ID 3, which is the button to "Swipe to place your order". This would finalize the purchase of the milk frother selected.

Figure 3.2. A complete episode on iOS screen navigation with GPT-4V. Best viewed by zooming in on the screen.

3.5 Conclusions

So far, we have discussed recent advances of vision and language from different aspects, which in sum contribute to the latest progress of Large Multimodal Models.

All together, we are excited about the progress that has been made in this field for the past few years and happy to be able to contribute to it. At the same time, we also deeply believe there is still a long way to go towards general vision-language models and agents, and there are still enormous challenges and research questions ahead for us to solve, as mentioned above.

GPT-4V is not the end of vision-language research. On the contrary, it opens a bright future for vision and language. There is a new chapter ahead for us to write.

Chapter 3, in part, is currently being prepared for submission for the publication of the material:

"GPT-4V in Wonderland: Large Multimodal Models for Zero-Shot Smartphone GUI Navigation" by An Yan, Zhengyuan Yang, Wanrong Zhu, Kevin Lin, Linjie Li, Jianfeng Wang, Jianwei Yang, Yiwu Zhong, Julian McAuley, Jianfeng Gao, Zicheng Liu, Lijuan Wang. The dissertation author was the primary investigator and author of this paper.

Bibliography

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[3] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

[4] Jinze Bai, Chang Zhou, Junshuai Song, Xiaoru Qu, Weiting An, Zhao Li, and Jun Gao. Personalized bundle list recommendation. *The World Wide Web Conference*, 2019.

[5] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[6] William Boag, Tzu-Ming Harry Hsu, Matthew McDermott, Gabriela Berner, Emily Alesentzer, and Peter Szolovits. Baselines for chest x-ray report generation. In *Machine Learning for Health Workshop*, pages 126–140. PMLR, 2020.

[7] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 1–21. Springer, 2022.

[8] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 - mining discriminative components with random forests. In *ECCV (6)*, volume 8694 of *Lecture Notes in Computer Science*, pages 446–461. Springer, 2014.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[11] Xinlei Chen, Li-Jia Li, Li Fei-Fei, and Abhinav Gupta. Iterative visual reasoning beyond convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7239–7248, 2018.

[12] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*, 2020.

[13] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023.

[14] Muhammad EH Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al Emadi, et al. Can ai help in screening viral and covid-19 pneumonia? *Ieee Access*, 8:132665–132676, 2020.

[15] Yann Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *ICML*, 2017.

[16] Marleen De Bruijne. Machine learning approaches in medical image analysis: From detection to diagnosis, 2016.

[17] Dina Demner-Fushman, Sameer Antani, Matthew Simpson, and George R Thoma. Design and development of a multimodal biomedical information retrieval system. *Journal of Computing Science and Engineering*, 6(2):168–177, 2012.

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE Computer Society, 2009.

[19] Michael Denkowski and Alon Lavie. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the sixth workshop on statistical machine translation*, pages 85–91, 2011.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[21] George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145, 2002.

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[23] Jessica Echterhoff, An Yan, Kyungtae Han, Amr Abdelraouf, Rohit Gupta, and Julian McAuley. Driving through the concept gridlock: Unraveling explainability bottlenecks. *arXiv preprint arXiv:2310.16639*, 2023.

[24] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *ACL*, pages 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[25] Maxwell Forbes, Christine Kaeser-Chen, Piyush Sharma, and Serge Belongie. Neural naturalist: Generating fine-grained image comparisons. *arXiv preprint arXiv:1909.04101*, 2019.

[26] Linda B Gambrell and Ruby J Bales. Mental imagery and the comprehension-monitoring performance of fourth-and fifth-grade poor readers. *Reading Research Quarterly*, pages 454–464, 1986.

[27] Linda B Gambrell and Patricia S Koskinen. Imagery: A strategy for enhancing comprehension. *Comprehension instruction: Research-based best practices*, pages 305–318, 2002.

[28] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann Dauphin. Convolutional sequence to sequence learning. In *ICML*, 2017.

[29] Jian Guan, Yansen Wang, and Minlie Huang. Story ending generation with incremental encoding and commonsense knowledge. In *AAAI*, 2019.

[30] Philipp Harzig, Yan-Ying Chen, Francine Chen, and Rainer Lienhart. Addressing data bias problems for chest x-ray image report generation. *arXiv preprint arXiv:1908.02123*, 2019.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[32] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. *CVPR*, pages 961–970, 2015.

[33] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

[34] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[35] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

[36] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019.

[37] Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. *arXiv preprint arXiv:1808.10584*, 2018.

[38] Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*, 2017.

[39] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

[40] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131, 2018.

[41] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[42] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020.

[43] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013.

[44] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[45] Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Found. Trends Mach. Learn.*, 5:123–286, 2012.

[46] Seanie Lee, Dong Bok Lee, and Sung Ju Hwang. Contrastive learning with adversarial perturbations for conditional text generation. *arXiv preprint arXiv:2012.07280*, 2020.

[47] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics.

[48] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023.

[49] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*, 2015.

[50] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[51] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4654–4662, 2019.

[52] Lei Li, Yongfeng Zhang, and Li Chen. Personalized transformer for explainable recommendation. In *ACL/IJCNLP*, 2021.

[53] Piji Li, Zihao Wang, Zhaochun Ren, Lidong Bing, and Wai Lam. Neural rating regression with abstractive tips generation for recommendation. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017.

[54] Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Hybrid retrieval-generation reinforced agent for medical image report generation. In *Advances in neural information processing systems*, pages 1530–1540, 2018.

[55] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[56] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[57] Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare Conference*, pages 249–269. PMLR, 2019.

[58] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

[59] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[60] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018.

[61] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[62] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52, 2015.

[63] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*, 2022.

[64] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

[65] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *NAACL*, pages 839–849, San Diego, California, June 2016. Association for Computational Linguistics.

[66] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.

[67] Soroush Nasiriany, Fei Xia, Wenhao Yu, Ted Xiao, Jacky Liang, Ishita Dasgupta, Annie Xie, Danny Driess, Ayzaan Wahid, Zhuo Xu, et al. Pivot: Iterative visual prompting elicits actionable knowledge for vlms. *arXiv preprint arXiv:2402.07872*, 2024.

[68] Jianmo Ni, Chun-Nan Hsu, Amilcare Gentili, and Julian McAuley. Learning visual-semantic embeddings for reporting abnormal findings on chest x-rays. *arXiv preprint arXiv:2010.02467*, 2020.

[69] Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, 2019.

[70] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, pages 722–729. IEEE Computer Society, 2008.

[71] Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*, 2023.

[72] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[73] OpenAI. Gpt-4 system card. https://openai.com/research/gpt-4v-system-card, 2023.

[74] OpenAI. Video generation models as world simulators. https://openai.com/research/video-generation-models-as-world-simulators, 2024.

[75] TB OpenAI. Chatgpt: Optimizing language models for dialogue. *OpenAI*, 2022.

[76] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[77] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505. IEEE Computer Society, 2012.

[78] Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828, 2021.

[79] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[80] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[81] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[82] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*, 2015.

[83] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CVPR*, pages 10674–10685, 2022.

[84] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[85] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.

[86] Mark Sadoski and Allan Paivio. Imagery and text: A dual coding theory of reading and writing. Lawrence Erlbaum Associates Publishers, 2000.

[87] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR, 2020.

[88] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[89] Yuge Shi, Jeffrey Seely, Philip HS Torr, N Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*, 2021.

[90] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*, 2020.

[91] Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. Language models can see: Plugging visual controls in text generation. *ArXiv*, abs/2205.02655, 2022.

[92] Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. A contrastive framework for neural text generation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *NeurIPS*, 2022.

[93] Hao Tan, Franck Dernoncourt, Zhe Lin, Trung Bui, and Mohit Bansal. Expressing visual relationships via language. *arXiv preprint arXiv:1906.07689*, 2019.

[94] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[95] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[96] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[97] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[98] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[99] Shaokun Wang, Tian Gan, Yuan Liu, Jianlong Wu, Yuan Cheng, and Liqiang Nie. Micro-influencer recommendation by multi-perspective account representation learning. *IEEE Transactions on Multimedia*, 2022.

[100] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.

[101] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[102] Zihan Wang, Chengyu Dong, and Jingbo Shang. " average" approximates" first principal component"? an empirical analysis on representations from neural language models. *arXiv preprint arXiv:2104.08673*, 2021.

[103] Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*, 2019.

[104] Mark Wilhelm, Ajith Ramanathan, Alexander Bonomo, Sagar Jain, Ed H. Chi, and Jennifer Gillenwater. Practical diversified recommendations on youtube with determinantal point processes. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018.

[105] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[106] An Yan, Shuo Cheng, Wang-Cheng Kang, Mengting Wan, and Julian McAuley. Cosrec: 2d convolutional neural networks for sequential recommendation. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 2173–2176, 2019.

[107] An Yan, Chaosheng Dong, Yan Gao, Jinmiao Fu, Tong Zhao, Yi Sun, and Julian McAuley. Personalized complementary product recommendation. In *Companion Proceedings of the Web Conference 2022*, pages 146–151, 2022.

[108] An Yan, Zexue He, Xing Lu, Jiang Du, Eric Chang, Amilcare Gentili, Julian McAuley, and Chun-Nan Hsu. Weakly supervised contrastive learning for chest x-ray report generation. *arXiv preprint arXiv:2109.12242*, 2021.

[109] An Yan, Zhankui He, Jiacheng Li, Tianyang Zhang, and Julian McAuley. Personalized showcases: Generating multi-modal explanations for recommendations. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2251–2255, 2023.

[110] An Yan, Xin Eric Wang, Tsu-Jui Fu, and William Yang Wang. L2c: Describing visual differences needs semantic understanding of individuals. *arXiv preprint arXiv:2102.01860*, 2021.

[111] An Yan, Yu Wang, Yiwu Zhong, Chengyu Dong, Zexue He, Yujie Lu, William Yang Wang, Jingbo Shang, and Julian McAuley. Learning concise and descriptive attributes for visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3090–3100, 2023.

[112] An Yan, Yu Wang, Yiwu Zhong, Zexue He, Petros Karypis, Zihan Wang, Chengyu Dong, Amilcare Gentili, Chun-Nan Hsu, Jingbo Shang, et al. Robust and interpretable medical image classifiers via concept bottleneck models. *arXiv preprint arXiv:2310.03182*, 2023.

[113] An Yan, Zhengyuan Yang, Wanrong Zhu, Kevin Lin, Linjie Li, Jianfeng Wang, Jianwei Yang, Yiwu Zhong, Julian McAuley, Jianfeng Gao, et al. Gpt-4v in wonderland: Large multimodal models for zero-shot smartphone gui navigation. *arXiv preprint arXiv:2311.07562*, 2023.

[114] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023.

[115] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. *arXiv preprint arXiv:2211.11158*, 2022.

[116] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023.

[117] Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 25407–25437. PMLR, 2022.

[118] Lili Yao, Nanyun Peng, Ralph M. Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. Plan-and-write: Towards better automatic storytelling. In *AAAI*, 2019.

[119] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

[120] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480*, 2022.

[121] Tian Yun, Usha Bhalla, Ellie Pavlick, and Chen Sun. Do vision-language pretrained models learn primitive concepts? *arXiv preprint arXiv:2203.17271*, 2022.

[122] Hongyu Zang and Xiaojun Wan. Towards automatic generation of product reviews from aspect-sentiment scores. In *INLG*, 2017.

[123] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *ACL*, pages 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics.

[124] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[125] Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. Gpt-4v (ision) as a generalist evaluator for vision-language tasks. *arXiv preprint arXiv:2311.01361*, 2023.

[126] Yubo Zhang, Hao Tan, and Mohit Bansal. Diagnosing the environment bias in vision-and-language navigation. *arXiv preprint arXiv:2005.03086*, 2020.

[127] Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. Mmicl: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint arXiv:2309.07915*, 2023.

[128] Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. Gpt-4v (ision) is a generalist web agent, if grounded. *arXiv preprint arXiv:2401.01614*, 2024.

[129] M. Zhou, Mirella Lapata, Furu Wei, Li Dong, Shaohan Huang, and Ke Xu. Learning to generate product reviews from attributes. In *EACL*, 2017.

[130] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems*, 36, 2024.

[131] Wanrong Zhu, An Yan, Yujie Lu, Wenda Xu, Xin Eric Wang, Miguel Eckstein, and William Yang Wang. Visualize before you write: Imagination-guided open-ended text generation. *arXiv preprint arXiv:2210.03765*, 2022.