Title	Knowledge Transferability in Vision-and-language Models and Its Applications
Author(s)	陳, 天偉
Citation	大阪大学, 2025, 博士論文
Version Type	VoR
URL	https://doi.org/10.18910/101753
rights	
Note	

Knowledge Transferability in Vision-and-language Models and Its Applications

Submitted to

Graduate School of Information Science and Technology

Osaka University

January, 2025

Tianwei CHEN

List of Publications

Journal Publications

1. **Tianwei Chen**, Noa Garcia, Liangzhi Li, Yuta Nakashima: Exploring Emotional Stimuli Detection in Artworks: A Benchmark Dataset and Baselines Evaluation. *Journal of Imaging 10, no. 6: 136, 2024/6/4, DOI: 10.3390/jimaging10060136.* (Chapter 3)

2. **Tianwei Chen**, Noa Garcia, Mayu Otani, Chenhui Chu, Yuta Nakashima, Hajime Nagahara: Exploring Knowledge Transferability between Vision-and-Language Tasks. *Journal of Imaging 10, no. 12: 300, 2024/11/22, DOI: 10.3390/jimaging10120300.* (Chapter 2)

International Conference

1. **Tianwei Chen**, Noa Garcia, Liangzhi Li, Yuta Nakashima: Retrieving Emotional Stimuli in Artworks. *The Annual ACM International Conference on Multimedia Retrieval (ICMR) 2024: 515-523* (Chapter 3)

2. **Tianwei Chen**, Yusuke Hirota, Mayu Otani, Noa Garcia, Yuta Nakashima: Would Deep Generative Models Amplify Bias in Future Models? *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2024: 10833-10843* (Chapter 4)

Domestic Conference (related to this thesis)

1. **Tianwei Chen**, Noa Garcia, Mayu Otani, Chenhui Chu, Yuta Nakashima, Hajime Nagahara: Exploring Knowledge Transferability between Vision-and-Language Tasks. *Meeting on Image Recognition and Understanding (MIRU) 2021* (Chapter 2)

Invited talks

1. Tianwei Chen, Would Deep Generative Models Amplify Bias in Future Models? The 3rd International Workshop on Multimodal Human Understanding for the Web and Social Media (MUWS) 2024 (Chapter 4)

2. Tianwei Chen, Would Deep Generative Models Amplify Bias in Future Models? Meeting on Image Recognition and Understanding (MIRU) 2024 (Chapter 4)

3. Tianwei Chen, 画像生成AIが将来のモデルにおける社会的なバイアスを増強するか？ Forum on Information Technology (FIT) 2024 (Chapter 4)

Fellowship

1. Tianwei Chen, 分野横断イノベーションを創造する情報人材育成フェローシップ (2021-2023)

Abstract

In this paper, we explore the knowledge transferability in recent vision-and-language models. Recently, transferring knowledge from pre-trained vision-and-language models to handle a new task has become a common idea in solving tasks related to both visual and linguistic data. However, the knowledge transfer strategy of current vision-and-language models is not always effective. On the one hand, some knowledge may not be helpful for knowledge transfer. On the other hand, harmful knowledge, such as social bias, may be involved in the pre-trained models. For both cases, the knowledge transferability in vision-and-language models is limited, as these models may not surely solve new tasks with their knowledge and may provide unfair performance to different social groups. To explore the limitations of the current knowledge transfer strategy, analyze the reason, and further improve the models' performance in solving vision-and-language tasks, I choose the exploration of knowledge transferability in vision-and-language tasks as my PhD topic and make an exhaustive analysis on this topic. First, we explored the knowledge transferability between 12 vision-and-language tasks to verify that some knowledge in one task may not always be helpful for other tasks. We then explore the knowledge transferability from large-scale pre-trained models to a new detection task related to emotions and artwork, and we confirm that large-scale pre-trained models may still not have enough knowledge to solve a task in a specific field. At last, we explore how the harmful knowledge in deep generative models, such as Stable Diffusion, can affect knowledge transferability. Our experiments show a possible way to utilize knowledge from such deep generative models if we can apply proper control to filter out harmful knowledge from these models.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Vision-and-language models are considered a foundation for artificial general intelligence, as they are designed to solve tasks that people usually face (problems with multiple modalities, such as image and text) and do how people usually do: solve problems by seeing, reading, and thinking. Such models are required to understand not only the two different modalities (*i.e.*, images and texts) but also the relations between them. Although training vision-and-language models requires related image-and-text pairs, which means the data collection is more expensive and laborious, many efforts have been made in collecting both training data and evaluation data in related vision-and-language tasks such as visual question answering [2–6], multi-modal verification [7–9], and referring expression [10–13]. Research on these tasks also helps with some real-world challenges, such as blind people assistant [14, 15] and text-guided object detector [16, 17]. These tasks can hardly be solved by models that can only deal with one modality.

Further breakthroughs have been made for solving more challenging tasks, such as zero-shot image-text retrieval [18, 19] and image generation [20–22]. These models not only solve existing problems but also create new applications, such as virtual try-On [23, 24].

Recently, researchers have been seeking more and more data for training vision-and-language models, as the recent trend shows that these models can solve more complex tasks if they are trained on more data [18, 25–27]. However, annotating and constructing training data for one certain task is expensive in both time and cost. To handle this, researchers have proposed many methodologies and smarter training strategies (*e.g.*, meta-learning [28–30] and self-supervised pre-training [18, 31–34]) to reduce the training data requirement. Among them, the *knowledge transferring strategy* is the basic but surely effective one, which (1) first prepares a model that has learned knowledge from other datasets and (2) transfers the model to a certain target task via the fine-tuning or adapting process. This strategy is successful in solving vision-and-language tasks, as the pre-trained models may have learned basic knowledge that may help the models quickly adapt to the target tasks. Some recent work [18, 19] even shows the possibility that, if a model learns enough knowledge, it may be able to solve certain tasks even without the adapting process.

However, the knowledge transferring strategy of current vision-and-language models is not always effective. On the one hand, some knowledge may not be helpful to some tasks. A pioneer work in vision-and-language tasks [35], as well as some recent works conducted on vision-only tasks [36, 37], have shown that tasks may not always help other tasks in getting a better performance. Besides, the experiments show some limitations of the knowledge transfer, as some tasks different from the training data may not get help from the pre-training. On the other hand, many recent studies [38–40] on fairness also show that models have learned and can even amplify social bias when solving target tasks, which brings risks to the further utilization of models and limits the benefits of recent improvements fairly towards every person. In each of the two cases, given the effort of collecting knowledge, vision-and-language models are still facing challenges in real-world applications, as their performance on challenging tasks is still unknown, and they may not benefit all humans fairly.

To explore the limitations of the current knowledge transfer strategy, analyze the reason, and further improve the models' performance in solving vision-and-language tasks, I mainly work on three topics in my PhD research, including (1) knowledge transferability between different vision-and-language tasks, (2) knowledge transferability toward tasks with specific knowledge, and (3) the effect of harmful knowledge transfer toward future models.

In the rest of this thesis, I will introduce my work on each of the topics, as shown in Fig-

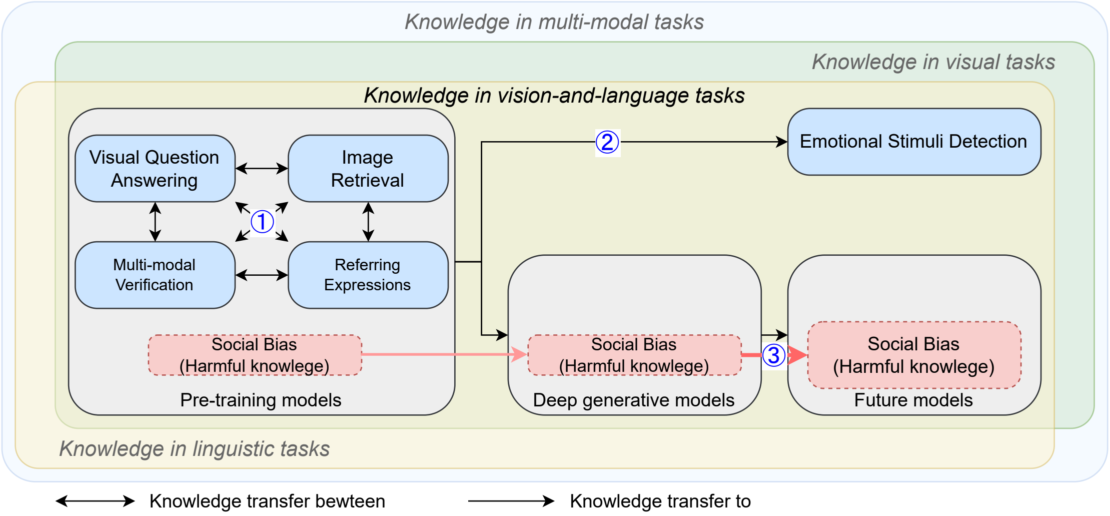

Knowledge in multi-modal tasks
Knowledge in visual tasks
Knowledge in vision-and-language tasks
Visual Question Answering
Image Retrieval
①
Multi-modal Verification
Referring Expressions
Social Bias (Harmful knowlege)
Pre-training models
②
Emotional Stimuli Detection
Social Bias (Harmful knowlege)
Deep generative models
③
Social Bias (Harmful knowlege)
Future models
Knowledge in linguistic tasks
Knowledge transfer bewteen
Knowledge transfer to

Figure 1.1: The overview of the thesis. We explore ① the knowledge transferability between vision-and-language tasks, ② knowledge transferability from pre-trained models to a challenging task: emotional stimuli detection, and ③ how harmful knowledge (such as social bias) in deep generative models affects future models.

ure 1.1. In Chapter 2, I will introduce the explorations of how vision-and-language tasks can help each other. We conduct an exhaustive analysis based on hundreds of cross-experiments on twelve vision-and-language tasks categorized into four groups. We further evaluate four factors that may affect the knowledge transferability, which are the random seeds, the data scale, the training stage, and the dataset similarity.

In Chapter 3, we explore how recent large pre-trained models (*e.g.*, VilBERT [31] and CLIP [18]) can be directly applied to challenging tasks. We first propose a task and annotate an evaluation dataset to detect the artwork regions that provoke certain emotions, and this task requires both knowledge of artwork and emotions. We then evaluate eight baseline models on this task, including a weakly-supervised model that we proposed for this task. Furthermore, we explore how the recent deep generative model, Stable Diffusion [22], can understand emotional stimuli.

In Chapter 4, we explore how social bias, a kind of harmful knowledge, can affect future models. Recent studies [38–40] show that deep generative models (*e.g.*, Stable Diffusion [22]) can generate biased images without intention. However, the generated images are increasing on the internet and may become training data for future models. To explore how deep generative models will affect future models, we conduct simulation experiments of dataset contamination by replacing the original image with the generated images. We then evaluate both the bias changes and model performance changes to evaluate how the future models are affected by the deep generative models. Furthermore, we make an analysis during the experiments and point out some factors that may affect the bias changes of data contamination.

In conclusion, my PhD research focuses on knowledge transferability in vision-and-language tasks in three different situations. The contributions could be summarized as follows.

1. We explore the knowledge transferability between twelve vision-and-language tasks. Our experiments indicate that knowledge from different tasks can not always help each other to improve performance. Task similarity, dataset size, pre-training stage, and other factors will affect the transferability.

2. We explore the knowledge transferability from large pre-trained vision-and-language models (*e.g.*, CLIP [18]) toward a challenging task: detecting regions in the artwork that evoke human emotions. Our experiments show that it is still hard for these models to solve this task, which indicates that although large pre-trained models are considered to learn plenty of knowledge, they still have their limitations in challenging tasks.

3. We explore the effect of harmful knowledge related to social bias in deep generative models toward future vision-and-language models. Our experiments confirm the existence of social bias in deep generative models. However, our experiments also show that such deep generative models do not only cause bias amplification but also other kinds of bias changes, such as bias mitigation. Such results show the possibility that the knowledge in deep generative models could be helpful to other vision-and-language tasks if we apply proper control to filter out harmful knowledge from these models.

Chapter 2

Knowledge Transferability in Vision-and-Language Tasks

2.1 Overview

Is learning more knowledge always better for vision-and-language models? In this chapter, we study knowledge transferability in multi-modal tasks. The current tendency in machine learning is to assume that by joining multiple datasets from different tasks, their overall performance improves. However, we show that not all the knowledge transfers well or has a positive impact on related tasks, even when they share a common goal. We conduct an exhaustive analysis based on hundreds of cross-experiments on 12 vision-and-language tasks categorized into 4 groups. While tasks in the same group are prone to improve each other, results show that this is not always the case. In addition, other factors, such as dataset size or the pre-training stage, may have a great impact on how well the knowledge is transferred.

The more data for learning, the better seems to be the current *motto* in machine learning, as large language models get exceptional results on previously unseen tasks by being trained on hundreds of millions of samples crawled from the Internet [41–44]. Following the path led by natural language processing research, the computer vision community is gradually adopting

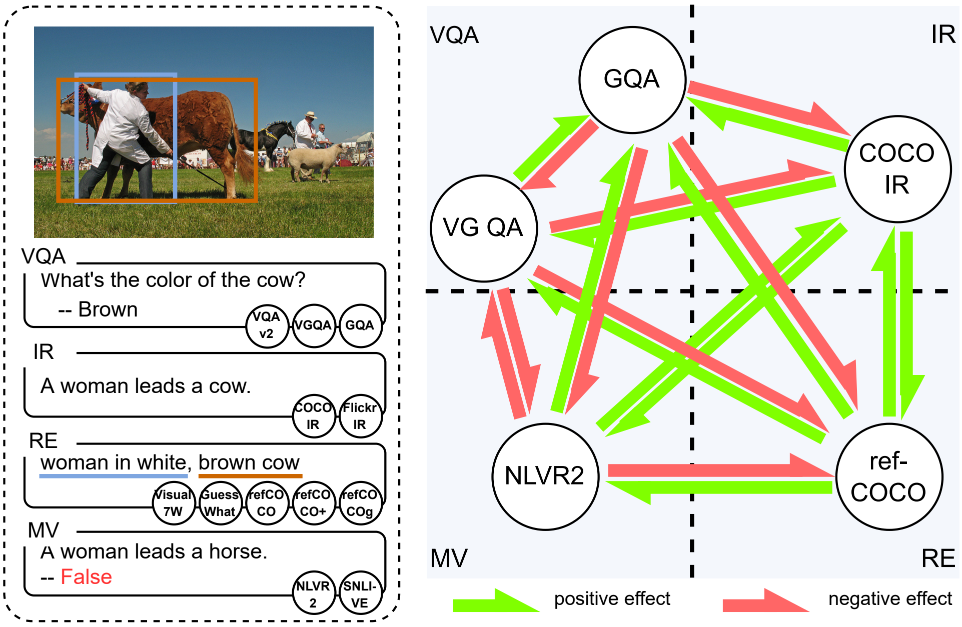


Figure 2.1: We explore the transferability among 12 vision-and-language tasks in 4 different groups: visual question answering (VQA), image retrieval (IR), referring expression (RE), and multi-modal verification (MV). Here, we illustrate the transferability among 5 tasks. Different tasks have different effects (positive or negative) on the other tasks.

Transformer-based models trained on web-scale datasets to achieve high performance in zero-shot settings [18, 45]. This is conducted by leveraging huge amounts of image-caption pairs available online to let the models learn the correspondences between the language semantics and the visual appearance of objects.

The problem with using hundreds of millions of samples for training is that the analysis, maintenance, processing, and particularly understanding of the data is beyond human means. With rising concerns about large models encoding and perpetuating harmful representations towards historically discriminated groups [46, 47], how data is handled acquires a crucial role. Knowing which data is being used, why, and for what means is now more important than ever.

We try to answer *whether more data is always better* by systematically analyzing the transferability within vision-and-language, which is the subset of tasks that require both visual and language understanding to be solved. For example, image captioning [48] or visual question answering [3]. In the last decade, dozens of high-quality vision-and-language datasets were collected, cleaned, and used as *de facto* benchmarks for human-like reasoning [49, 50]. Now, some of these datasets created with diverse motivations and purposes are coming together to train large vision-and-language models [31, 32].

While some tasks can improve their performance when a model is trained in a multi-dataset and multi-task protocol [35], it is still unclear to what extent and whether all vision-and-language tasks can benefit from this. Our goal is to shed light on this question and explore the transferability of knowledge within vision-and-language tasks in a similar way as [36, 37] do for vision-only datasets. Specifically, we conduct hundreds of cross-experiments in which the performance of a target task trained under a dozen different initializations from different pre-trained source tasks are compared.

Following [35], we divide vision-and-language tasks into 4 groups: visual question answering (VQA), image retrieval (IR), referring expression (RE), and multi-modal verification (MV), and we study both intra- and inter-group transferability. As illustrated in Figure 2.1, our results indicate that there is not yet a magic formula to consistently improve performance on all the datasets by transferring knowledge between tasks. In other words, while some target tasks benefit from a specific source task pre-training, others get harmed. Even within target tasks that are similar in terms of datasets and goals, different behavior is observed when the same source knowledge is transferred. Conversely, similar behaviors happen when different knowledge is transferred. This leads to the conclusion that more data is not always necessarily better for higher performance since it depends on the training dataset's goal, nature, and size.

From the experiments, we acquired several insights about the transferability of knowledge between vision-and-language models, which are summarized as follows:

- Tasks in the same group are more likely to help each other to improve performance. However, negative results show that tasks with shared goals do not always contribute positively to one another. This indicates that having a shared goal is favorable, but not enough.
- In the inter-group experiments, we find that the RE tasks tend to have a positive effect on

most of the tasks in other groups, while the MV group tends to receive a positive effect from other groups.

- While the best improvement is often given when knowledge is transferred within the same group, the worst results are concentrated on specific tasks, particularly GQA [6]. We study why and how this happens.
- We detect that different random seeds strongly affect the numeric performance of each task, sometimes even more than the transfer learning itself. This urges to report of vision-and-language results on multiple random configurations.
- We explore the effect of the data scale of the source task by down-sampling a large-scale task. The results show increasing performance on all of the smaller-scale tasks, which indicates that the dataset size is an important but not always a positive factor in knowledge transferability.
- We also explore how different stages of training affect the performance of the target tasks. We discover that in some cases, transferring knowledge at the early stages of pre-training can benefit the target task. When the model learns too much, the performance on the target task drops.
- Finally, we analyze the similarity between the 12 tasks' datasets and explore how the similarity between these datasets relates to knowledge transferability. We discover that tasks with different datasets can help each other, while tasks with similar datasets can bring negative effects. These results show that the dataset similarity may not strongly affect the vision-and-language tasks.

2.2 Related work

Knowledge transferability focuses on how a model that learns knowledge from source tasks can adapt to a new task. Existing research on this topic includes transfer learning [45, 54], multi-task learning [55, 56], and meta-learning [30, 57]. Ideally, the more knowledge a model learns, the better performance it has. However, in practice, models are affected by several phenomena,

Table 2.1: Dataset statistics for the 12 tasks used in our experiments. From the left, the first and second columns are the number of samples in the train and validation (Train +Val) set and test set, respectively. The third column is the metric to evaluate the corresponding task. The fourth column is the name of the test set. The fifth column is the number of images in the train and validation set. The last column is the source dataset from which the images of the corresponding task come from.

	Train + Val samples	Test samples	Evaluation metric	Evaluation set	Train + Val image	Image source
VQA v2 [3]	542,104	447,793	Accuracy	test-dev	98,861	MSCOCO [48]
VG QA [5]	1,294,255	5,000	Accuracy	validation	92,147	MSCOCO [48] + YFCC100M [51]
GQA [6]	962,928	12,578	Accuracy	test-dev	69,868	Visual Genome [5]
COCO IR [48]	487,600	1,000	Recall@5	test	99,435	MSCOCO [48]
Flickr30K IR [52]	140,485	1,000	Recall@5	test	29,077	Flickr30K [53]
NLVR2 [8]	86,373	6,967	Accuracy	test-P	29,808	NLVR2 [8]
SNLI-VE [9]	512,396	17,901	Accuracy	test	95,522	Flickr30K [53]
Visual7w [10]	93,813	57,265	Accuracy	test	16,415	MSCOCO [48]
GuessWhat [11]	100,398	23,785	Accuracy	test	51,291	MSCOCO [48]
refCOCO [12]	96,221	10,752	Accuracy	test	14,481	MSCOCO [48]
refCOCO+ [12]	95,852	10,615	Accuracy	test	14,479	MSCOCO [48]
refCOCOg [13]	65,514	9,602	Accuracy	test	17,903	MSCOCO [48]

such as catastrophic forgetting [58,59], that limit their performance. Our work is mainly related to the following two topics:

2.2.1 Transferability analysis

Transferability analysis studies how well the knowledge from a source task benefits a target task. Zamir *et al.* [36] proposed a method to analyze and utilize the transferability among 24 vision-only tasks on a single indoor scenes dataset. They pre-trained models in the source tasks, transferred them to the target tasks, and calculated the transferability by evaluating how well the model performed in the target task. Following this idea, Mensink *et al.* [37] studied the transferability between 20 real-world vision-only tasks. They analyzed three main factors: the image domain similarity between source and target tasks, the task type, and the data size.

While studies in [36, 37] were conducted on vision-only tasks, we aim to explore multi-modality transferability in the vision-and-language domain. The particularity of multi-modal datasets is that knowledge needs to be transferred not only across tasks but also between modalities, which adds an extra layer of difficulty to the problem.

2.2.2 Paradigm of solving vision-and-language tasks

The most popular paradigm of knowledge transfer is to pre-train a model on a large dataset and transfer it to a downstream task [31, 32, 60–67]. For example, Lu *et al.* [31] proposed a BERT-based vision-and-language model, and pre-trained it with three self-supervised tasks to learn knowledge from Google's Conceptual Captions dataset [68]. Following this work, many contributions were made in applying better text modeling [65], better visual feature extraction [66], and contrastive learning [18, 32]. Besides, there has been some work analyzing the knowledge transferability in specific tasks such as video question answering [69].

Recently, CLIP [18] has shown a remarkable capacity to understand both vision and language data, by applying a specific image-text contrastive learning strategy. In this model, images and texts are processed by separate image encoder and text encoder, and the model is trained to match the image feature and text feature that belong to one pair. The specific design of CLIP makes it good for making zero-shot scenarios of vision-and-language tasks. Following CLIP, many studies explore how to utilize CLIP to improve models' performance in existing vision-and-language tasks. Song *et al.* [70] explored the possibility of using the CLIP model directly for the vision-and-language tasks in the scenario of few-shot learning. Tsimpoukelli *et al.* [71] and Shen *et al.* [72] explored the possibility of utilizing CLIP's visual encoder and text encoder to extract more useful features for vision-and-language tasks. Li *et al.* [73] applied CLIP's image-text contrastive learning strategy to the pre-training process of large vision-and-language models and explored how the training strategy benefits the pre-training model.

In general, our work is similar to CLIP in that we are also concerned about how downstream tasks can benefit from the pre-training. However, CLIP is different from us as we primarily focus on how the related tasks can help each other, while CLIP focuses more on how to train a model from an unrelated general dataset.

Our work is also close to multi-task vision-and-language learning [35, 74–76]. Nguyen *et al.* [75] proposed a multi-task learning model with three vision-and-language tasks by choosing the best layers for each task. In [35], a training strategy to prevent learning too much knowledge from converged tasks is proposed, resulting in a model trained on 12 vision-and-language tasks. Following this idea, Hu *et al.* [76] designed a unified transformer that can learn from either vision or text data. This model enables multi-task learning among vision-only, text-only, and vision-and-language tasks, and thus extends the knowledge that the vision-and-language model can learn.

None of the above work conducts a formal analysis of how the different tasks affect each other. Conversely, we thoroughly explore knowledge transferability among vision-and-language tasks and uncover insights that may be useful when applying knowledge transfer methods to vision-and-language.

2.2.3 Vision-and-language tasks

This chapter mainly explores knowledge transferability in four types of tasks: visual question answering [3,5,6], image retrieval [48,52], referring expressions [10–13], and multi-modal verification [8,9]. There are many other types of vision-and-language tasks, such as image captioning [48, 77], text-to-image generation [20, 22, 78], and visual language navigation [79, 80]. Furthermore, there are also other interesting tasks related to other modalities, such as video [81–83], and voice [84]. Evaluating more tasks could provide more insights about knowledge transferability, but we only focus on four types of tasks, following [35], that take both image and text as input.

2.3 Vision-and-language tasks in this work

Visual question answering (VQA). Given an image and a related question, VQA requires a model to select an answer from several candidates. The setting of VQA aims to not only explore the model's capacity to understand both visual and linguistic data but also the capacity of knowledge reasoning, which is also known as a "visual Turing challenge" [85]. As the example shown in Figure 2.1, when the VQA task gives a question "What's the color of the cow?" that relates to the given image, the model not only needs to understand both the image and question but also needs to check the color of the cow to give a proper answer "brown" to the given question. Beginning with the idea of the "visual Turing challenge," Malinowski *et al.*propose the classic "questing-to-image" formula and firstly release a small-scale (about $12K$ question-answer pairs) but workable dataset DAQUAR [85] for both training and evaluation. To solve the data-scale problem, Ren *et al.*generate question-answer pairs based on the COCO caption dateset [86] to construct the dataset COCOQA [86], which enlarges the scale of training data to about $82K$ question-answer pairs. To further make a reliable dataset, large annotation projects [2, 3, 5] on visual question answering are launched and result in currently the most widely used datasets VQA v2 [3] and Visual Genome QA (VG QA) [5].

When the standard visual question answering task shows the possibility for a model to answer visual questions, many studies start to explore the visual question answering models. For example, Hudson *et al.*propose a dataset GQA [6] that requires the model to focus more on the relations between visual contents. There are also studies that try to use visual question answering models to solve real-world challenges such as blind people caring [14].

In this chapter, our exploration involves the following three VQA tasks: VQA v2 [3], VG QA [5], and GQA [6].

- **VQA v2** is a classic visual question answering task towards solving multi-modal problems. It contains $204K$ images from MSCOCO [48] with $614K$ human-annotated natural language question-answer (QA) pairs. In the evaluation process, models are required to predict one answer from all answer candidates around the whole dataset, *i.e.*, each question has thousands of answer candidates.

- **Visual Genome QA (VG QA)** has a similar target as VQA v2, but has a larger dataset with $108K$ images, $1.7M$ QA pairs, as well as $5.4M$ region descriptions and $2.3M$ object relationships, which provide rich evidence for the analysis of visual question answering. Similar to VQA v2, this task requires the models to predict one answer from all answer candidates around the whole dataset.
- **GQA** is more concerned with the models' capacity on visual reasoning. It contains a dataset with $113K$ images and $22M$ QA pairs, which leverage the scene graph information from VG QA [5] to generate more challenging questions that need multiple reasoning steps to arrive at the answer. During the evaluation process, GQA also requires the models to predict one answer from all answer candidates around the whole dataset.

Image retrieval (IR). Given a caption, image retrieval requires the model to select the most representative image from a pool of images. The target of image retrieval is challenging as the images and sentences may be highly related to each other. Furthermore, image retrieval is also challenging when the task scale increases, as the time complexity of calculating the image-sentence matching score is about $O(n^2)$. As shown in Figure 2.1, given the text of "A woman leads a cow.", the model should find the related image as shown in the top-left part. The challenge of this task is, although the image and sentence in one pair clearly match each other, many of the images and the sentences are very similar. The similarity between image-sentence pairs makes it difficult to distinguish the correct image by the given sentence. In the area of image retrieval, COCO IR [48] and Flickr30K IR [52] are the two of the most widely used datasets, which are both for exploring models' capacity to retrieve correct images. In recent years, many valuable challenges have been proposed in the formula of image retrieval, such as artwork retrieval [87] and food retrieval [88].

In this chapter, our exploration involves the following two IR tasks: COCO IR [48] and Flickr30K IR [52].

- **COCO IR** in an image retrieval task based on the COCO caption dataset [48]. In this task, there are $123K$ images with $567K$ related human-annotated captions. To evaluate models' performance, COCO IR provides three accuracy scores with different recall scales: the accuracy on top one retrieval (Recall@1), the accuracy on top five retrievals (Recall@5), and the accuracy on top ten retrievals (Recall@10). In this chapter, we use Recall@5 as the main metric for the evaluation.

- **Flickr30K IR** is an image retrieval task based on Flickr30K dataset [52]. It has $31K$ images with $146K$ human-annotated captions. Similar to COCO IR, Flickr30K IR uses Recall@1, Recall@5, and Recall@10 to evaluate the models' performance. In this chapter, we also use Recall@5 as the main evaluation metric.

Referring expressions (RE). Referring expression concerns the relation between linguistic expressions (*i.e.*, texts) and visual contents (*i.e.*, objects), which can be divided into two directions: 1) detecting visual contents based on the expressions, or 2) generating expressions by the given visual contents. In this chapter, we mainly focus on the first direction of referring expression. Given a text and an image, referring expressions require the model to detect the corresponding region in the image described by the text. In contrast to image retrieval, referring expressions do not focus on retrieving one image from a group. Instead, it focuses on detecting the related region from one image. Thus, compared to image retrieval, the referring expression takes more concern about the specific objects in one image. As the example shown in Figure 2.1, instead of the text about the whole image (*e.g.*, "A woman leads a cow."), texts such as "woman in white" and "brown cow" are given as the targets, which are related to certain objects. Referring expression is used to be a classic natural language processing task that has been studied since the 1970s [89]. From that time, researchers kept interested in how models can work as humans to link the texts and the visual contents and added more challenges, such as applying real-world images [12, 13], enriching object categories [11], and more difficulty in visual reasoning [10].

In this chapter, our exploration involves the following RE tasks: Visual7w [10], GuessWhat [11], and RefCOCO(+/g) [12, 13].

- **Visual7w** is a referring expression task based on the Visual 7w dataset [10]. In this task, there are $25K$ images with $151K$ region-text pairs. The evaluation method is the accuracy of if a model predicts a region that has an Intersection over Union (IoU) score higher than 50%.
- **GuessWhat** is a referring expression task based on the GuessWhat dataset [11]. In this dataset, there are $66K$ images with $137K$ region-text pairs. The same as Visual7w, GuessWhat evaluates models' performance by the object prediction accuracy with IoU$> 50\%$.
- **RefCOCO(+/g)** [12, 13] are three similar referring expression tasks that leverage the image and object information from the COCO dataset [48]. Among these tasks, refCOCO and refCOCO+ are collected by ReferitGame [12], which is an interactive game between two players as one player expresses an object and the other player points it out. The refCOCO+ is more challenging than refCOCO as it restricts the player to not using location words during the game. While refCOCOg [13] collects its data in a non-interactive setting, which asks annotators to express the given object directly. In general, refCOCO has $19K$ images with $131K$ expressions, refCOCO+ has $19K$ images with $130K$ expressions, and refCOCOg has $25K$ images with $90K$ expressions. The same as the above referring expression tasks, refCOCO(+/g) evaluates models' performance by the object prediction accuracy with IoU$> 50\%$.

Multi-modal verification (MV). Given one or more images and a referred text, multi-modal verification requires the model to decide whether the text is correct or not. Different from the rest of the three groups of tasks, multi-modal verification tasks usually have a very limited number of answers, e.g., NLVR2 [8] only has two candidate answers, and SNLI-VE [9] only has three candidate answers. However, multi-modal verification is challenging in the view that it requires more visual reasoning capacity to verify if the text and image conflict. As the example shown in Figure 2.1, multi-modal verification requires the model to judge if “a woman leads a horse" in the image. The verification is considered challenging as understanding “a woman" and “a horse" cannot directly lead to the correct answer. The idea of recent multi-modal verification studies, such as CLEVR [90] and [7], are motivated by the progress of visual question answering, which shows the evidence that deep learning models are capable of making knowledge reasoning. The results of these two tasks are encouraging and further motivating the tasks of NLVR2 [8] and SNLI-VE [9], which are currently the most widely studied multi-modal verification tasks.

In this chapter, our exploration involves the following two MV tasks: NLVR2 [8] and SNLI-VE [9].

- **NLVR2** [8] is a multi-modal verification task that requires models to verify if one sentence is true in both two images, *i.e.*, the task takes one text and two images as the input and requires the models to give a binary answer (true or false). This specific design is to verify if the model can make reasoning across not only the modality but also different data in the same modality. The dataset of this task contains $103K$ real-world images with $93K$ human-annotated texts. The evaluation process involves calculating the accuracy of the models' binary predictions.
- **SNLI-VE** [9] is another multi-modal verification task to verify if a hypothesis (text) is accurate (entailment), partly accurate (neutral), or wrong (contradiction) to a given premise (image). The dataset of this task contains $31K$ images with $548K$ texts, and the evaluation metric is the accuracy of the triplet classification.

2.4 Methodology

We study how the knowledge from a source task affects a target task. Formally, we define the problem as follows:

Given a set $\mathcal{T}$ of vision-and-language tasks, we pick out a source task $s \in \mathcal{T}$ and a target task $t \in \mathcal{T}$. We train a *direct* model m_t by training a model m with the target task t. We also train a *one-hop* model $m_{s \to t}$ by pre-training m with s, and then with t. As shown in Figure 2.2, the performances of a pair of models $(m_t, m_{s \to t})$ are compared for all possible combinations of s and t in $\mathcal{T}$. Tasks are categorized into groups according to their main goal, so tasks with

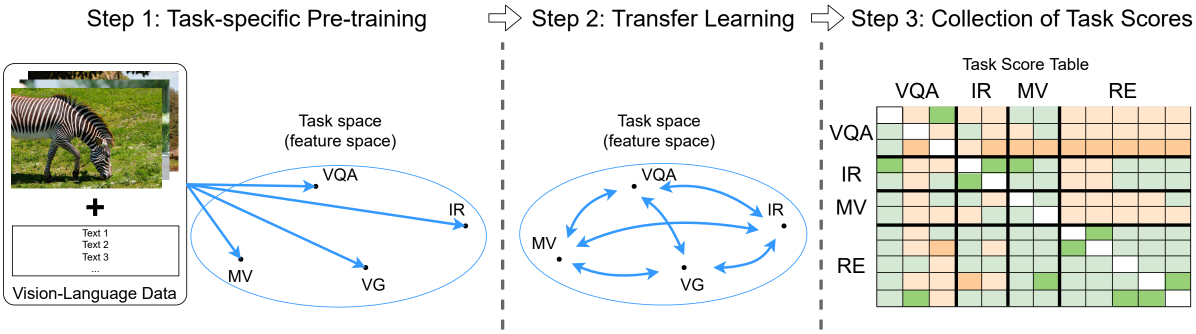


Figure 2.2: Analysis of transferability relationships between tasks. In Step 1, we train 12 vision-and-language tasks independently. In Step 2, we use the models from Step 1 and fine-tune them on each of the other tasks. In Step 3, we form a transferability relation table for the 12 vision-and-language tasks in four groups: visual question answering (VQA), image retrieval (IR), multi-modal verification (MV), and referring expressions (RE).

similar goals are assigned to the same group.

2.4.1 Tasks selection

As introduced in Section 2.3, we study 12 vision-and-language tasks categorized into four groups. The feature of each task is listed in Table 2.1. Please note that the number of samples and images in the Train + Val set is the number after the cleaning in Section 2.5.

2.4.2 Model

We follow the model structure in [35], consisting of a unified multi-modal encoder based on VilBERT [31] with 12 different task-specific heads for corresponding tasks. The training goal is:

$$\arg\min_{\theta_e,\theta_t} L_t(\psi_{\theta_t}(\phi_{\theta_e}(V_t, S_t))), \tag{2.1}$$

where V_t and S_t are the image and text in the dataset of task t, and θ_e and θ_t are the parameters of the encoder ϕ and the task t's head ψ, respectively. L_t is the loss of the task t.

Table 2.2: Results of *direct* model m_t in the 12 tasks.

10 different random seeds		avg ± std	max	min
Task	VQA v2	70.3 ± 0.56	70.71	69.18
	VG QA (Val)	33.5 ± 0.48	34.17	32.86
	GQA	58.1 ± 0.53	58.65	57.10
	COCO IR	90.4 ± 0.77	91.02	89.12
	Flickr30K IR	86.5 ± 0.87	87.24	84.80
	NLVR2	73.4 ± 0.50	74.11	72.34
	SNLI-VE	75.3 ± 0.16	75.64	75.04
	Visual7w	80.4 ± 0.19	80.63	80.04
	GuessWhat	62.3 ± 0.17	62.68	62.14
	refCOCO	77.7 ± 0.30	78.15	77.13
	refCOCO+	69.1 ± 0.57	69.68	67.81
	refCOCOg	71.6 ± 0.63	72.50	70.50

2.4.3 Workflow

The workflow, as shown in Figure 2.2, is split into three steps: 1) task-specific pre-training, 2) transfer learning, and 3) collection of scores.

Task-specific pre-training. We pre-train each of the 12 tasks independently, *i.e.*, each task $s \in \mathcal{T}$ is trained by its corresponding dataset and does not see any dataset from other tasks. We collect the trained models m_s from each task as the pre-trained models, which learned task-specific knowledge from the source task. We also evaluate each *direct* model m_t as baselines for non-transferred knowledge.

Transfer learning. We fine-tune, again, each pre-trained model m_s. Given m_s and the target task t, we get a final model $m_{s\rightarrow t}$ by fine-tuning m_s with all of the training samples in task t.

Collection of scores. We categorize tasks into groups and evaluate all *direct* models m_t and *one-hop* models $m_{s\rightarrow t}$ for all possible task pairs. Results are discussed in Section 2.5.2.

2.5 Experiments on VilBERT

Datasets. We use the same set of datasets as [35], including the training and test sets of the 12 tasks. The overlapping samples from the different tasks were removed from the training sets to prevent leaking data from the test set into the training set. Note that the original test sets were not changed during this cleaning process. For training and validation sets, VQA v2, VG QA, COCO IR, and NLVR2 have about 100,000 images; GQA and GuessWhat have about 60,000 images; Flickr30K IR and SNLI-VE have about 30,000 images; and refCOCO, refCOCO+. refCOCOg and Visual7w have about 15,000 images.

Experimental settings. We follow most of the settings in [35]. We modify the batch size to $1/4$ to fit the training in our server.[1] Pre-trained models m_s are trained for 6 epochs, which is enough for convergence. To ensure models $m_{s \to t}$ learn task-specific knowledge well, we use the models with the best performance in the validation set, except for VG QA, which is evaluated on the validation set, and thus the model at the 6th epoch is used. All of the models are seen converged in their corresponding tasks. We train every model with three different random seeds and report results by their mean and standard deviation.

Evaluation metrics. We use accuracy for tasks in the VQA group and the MV group. For the IR group, we use Recall@5. For the RE group, we follow [12, 13, 35] and compute the score based on the intersection over union (IOU) between the ground truth and the prediction.

2.5.1 Random seed

Preliminary results showed large variations in performance when models are trained under different random initializations, as also shown in [91]. Thus, before proceeding with the transferability experiments, we first explore the instability of vision-and-language tasks and their sensibility to random seeds. We train each direct model, m_t for all $t \in \mathcal{T}$, 10 times with different random seeds. The results are shown in Figure 2.3. Although most of the tasks present a gap larger than 1% between the maximum and the minimum score, most of the scores in each task

[1] We use a single server with 4 16GB NVIDIA P100 GPUs.

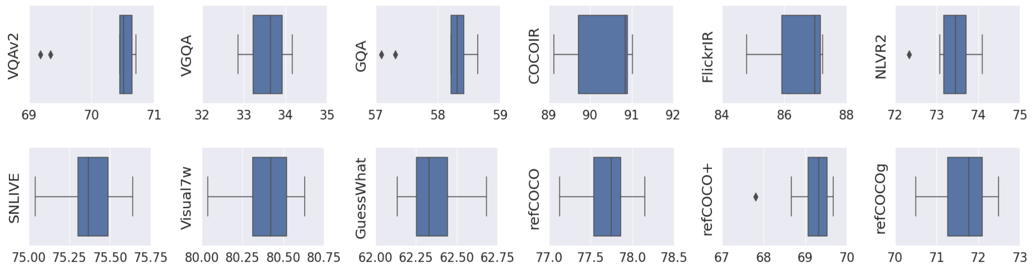


Figure 2.3: Box plots of the 12 tasks trained with 10 random seeds showing a big gap between the best and the worst scores.

are concentrated in a small region. More details are shown in the box plots in Table 2.2. Nine tasks have a gap larger than 1%. Among them, Flickr30K IR is the one that fluctuates the most, with a gap of 2.44% and a standard deviation of 0.87. This reveals that experiments on a single run may not be reliable enough to extract conclusions about model performance. In general, we found that the random seed has a big impact on the evaluation of vision-and-language tasks. To ensure our results are reliable, we run each experiment three times.

2.5.2 Results by group

For the transferability experiments, we collected results from 12 *direct* models m_t and 132 *one-hop* models $m_{s \to t}$ and present them herein in Tables 2.3. We used the results of m_t (Row "*direct* model m_t") as the baseline. We relied on a color scheme to illustrate the comparative performance of the transferred models $m_{s \to t}$: deep green for the best scores of each column, *i.e.*, the best results of each task in transfer learning, and deep orange for the worst results. For the rest of the entries of the tables, light green and light orange indicated better and worse performance than the baseline, *i.e.*, positive or negative transfer of the knowledge.

Visual question answering group. Columns 1–3 (VQA v2, VG QA (Val), and GQA) in Table 2.3 show the results in the VQA group. VQA v2 and GQA benefit from each other, but they do not improve the VG QA performance. In fact, GQA has the worst effect on VG QA among the 12 tasks. VG QA achieves its best performance with the help of refCOCOg, indicating that even though it is commonly seen as a VQA task, it may be closer to the RE group. When tasks in the VQA group are the target tasks, the source tasks have a consistent effect on each of them, *e.g.*, COCO IR gives the best effect to VQA v2, while giving a negative effect to both VG QA and GQA. In contrast, GuessWhat gives the worst effect on GQA, while giving a positive effect on VQA v2. More specifically, VQA v2 and VG QA show contrary behavior: VQA v2 obtains a positive effect from all of the tasks outside the VQA group, while only refCOCO and refCOCO+ give VG QA a positive effect. This indicates that although VQA, VG QA, and GQA have the same type of training goal, their underlying knowledge may be very different, and thus receive different contributions from the same source task. Finally, even though tasks in the VQA group are the largest in terms of training samples when they act as the source task, they tend to have a negative impact on the other group tasks (Row 1-3 in Tables 2.3), indicating that large training sets are not a guarantee for a better transfer.

Image retrieval group. Columns 4–5 (COCO IR and Flickr30K IR) in Table 2.3 summarize the performance of the IR group. Both tasks in this group help each other. Also, as source tasks, they show similar behavior, with a tendency to improve other tasks. However, the results in this group show the largest variance. On one hand, as target tasks, only the VQA group has a consistently negative impact on Flickr30K IR. On the other hand, the standard deviation scores in COCO IR and Flicker30K IR are usually larger than in other groups. The standard deviation scores of $m_{\text{COCO IR}\rightarrow\text{Flickr30K IR}}$ and $m_{\text{Flickr30K IR}\rightarrow\text{COCO IR}}$ tend to be larger than tasks in other groups.

Multi-modal verification group. Columns 6–7 (NLVR2 and SNLI-VE) in Table 2.3 list the performance of the MV group. Except for GQA, most of the source tasks have a positive effect on the two MV tasks. NLVR2 and SNLI-VE also improve each other, but the effect is not larger than the ones from COCO IR and refCOCO+. This may be in part because NLVR2 and SNLI-VE are considerably different: NLVR2 is a binary classification task that verifies if a comment describes a fact among multiple images, while SNLI-VE is a ternary classification

Table 2.3: Knowledge transferability results per group. Results of $m_{row \to column}$ (rows 2–13, green or red color) are compared with the *direct* model m_{column} (row 1, blue color) and assigned green (when the average score is higher than m_{column}, *i.e.*, positive transfer) or red (when the average score is lower than m_{column}, *i.e.*, negative transfer). Deep green/red shows the best/worst score in each column.

avg ± std		Target Task t											
		VQA v2	VG QA (Val)	GQA	COCO IR	Flickr30K IR	NLVR2	SNLI-VE	Visual7w	GuessWhat	refCOCO	refCOCO+	refCOCOg
direct model m_t		69.6 ± 0.71	33.7 ± 0.74	57.5 ± 0.59	89.2 ± 0.16	85.3 ± 0.48	72.8 ± 0.48	75.3 ± 0.11	80.1 ± 0.13	62.3 ± 0.08	77.3 ± 0.19	68.5 ± 0.72	70.8 ± 0.35
Source task s	VQA v2	-	33.5 ± 0.55	58.2 ± 0.21	89.0 ± 1.43	84.8 ± 1.04	73.7 ± 0.60	75.4 ± 0.23	79.5 ± 0.51	60.8 ± 0.62	76.8 ± 0.62	67.7 ± 1.12	70.5 ± 0.12
	VG QA	70.0 ± 0.33	-	57.5 ± 0.57	90.0 ± 0.11	84.3 ± 0.77	72.1 ± 0.69	75.7 ± 0.05	79.9 ± 0.91	60.9 ± 0.66	76.9 ± 0.57	68.0 ± 0.81	70.8 ± 0.33
	GQA	69.7 ± 0.16	33.0 ± 0.76	-	88.9 ± 1.17	82.9 ± 1.46	72.1 ± 0.76	75.3 ± 0.67	79.0 ± 0.12	60.7 ± 0.26	76.7 ± 0.32	67.2 ± 0.22	70.0 ± 0.42
	COCO IR	70.5 ± 0.56	33.6 ± 0.52	57.4 ± 0.49	-	86.8 ± 1.56	75.3 ± 0.24	76.1 ± 0.11	79.5 ± 0.34	62.0 ± 0.36	77.4 ± 0.59	69.4 ± 0.33	72.0 ± 0.25
	Flickr30K IR	70.3 ± 0.32	33.4 ± 0.69	57.6 ± 0.35	90.1 ± 1.30	-	74.0 ± 0.65	75.8 ± 0.22	79.8 ± 0.17	62.3 ± 0.16	77.3 ± 0.18	68.7 ± 0.22	71.3 ± 0.23
	NLVR2	69.9 ± 0.34	33.4 ± 0.35	57.5 ± 0.43	89.7 ± 1.16	84.8 ± 1.09	-	75.9 ± 0.06	79.4 ± 0.27	62.0 ± 0.17	77.1 ± 0.40	68.4 ± 0.40	70.9 ± 0.10
	SNLI-VE	69.9 ± 0.68	33.3 ± 0.51	57.3 ± 0.32	89.2 ± 0.51	85.5 ± 1.80	73.9 ± 0.24	-	79.2 ± 0.60	61.2 ± 0.40	76.8 ± 0.43	67.2 ± 0.41	70.4 ± 0.65
	Visual7w	70.2 ± 0.25	33.5 ± 0.94	57.7 ± 0.57	89.8 ± 0.45	85.6 ± 1.06	73.9 ± 0.71	76.1 ± 0.26	-	63.0 ± 0.36	78.1 ± 0.39	69.4 ± 0.06	72.8 ± 0.30
	GuessWhat	69.7 ± 0.61	33.5 ± 0.06	56.9 ± 0.33	89.5 ± 0.45	85.1 ± 2.09	73.2 ± 0.31	75.9 ± 0.34	80.8 ± 0.05	-	78.1 ± 0.13	69.1 ± 0.13	72.2 ± 0.06
	refCOCO	70.2 ± 0.22	33.7 ± 0.29	57.4 ± 0.21	90.1 ± 0.83	85.4 ± 1.33	73.7 ± 0.28	76.0 ± 0.33	80.3 ± 0.03	62.6 ± 0.29	-	69.5 ± 0.23	72.4 ± 0.29
	refCOCO+	70.1 ± 0.44	33.3 ± 0.05	57.2 ± 0.41	88.8 ± 1.93	84.9 ± 2.27	74.4 ± 0.22	76.1 ± 0.14	80.4 ± 0.17	62.5 ± 0.29	77.8 ± 0.35	-	73.0 ± 0.19
	refCOCOg	69.7 ± 0.30	34.0 ± 0.46	57.4 ± 1.07	89.1 ± 2.19	84.9 ± 1.24	74.1 ± 0.59	75.7 ± 0.20	80.5 ± 0.35	62.8 ± 0.15	78.3 ± 0.14	69.7 ± 0.35	-

task that verifies how well a comment describes an image. Another reason may be because of the data distributions: NLVR2's images are from ILSVRC 2014 [92], while SNLI-VE's are from Flickr30K [53].

Referring expressions group. Columns 8–12 (Visual7w, GuessWhat, refCOCO, refCOCO+, and refCOCOg) in Table 2.3 lists the scores of the RE group. All the tasks in this group benefit from transferred knowledge in the same group. The improvements within this group, especially among refCOCO, refCOCO+, and refCOCOg, are larger than those from tasks in other groups. However, all tasks in the VQA and the MV groups have a negative effect on the RE group (except $m_{\text{NLVR2}\rightarrow\text{refCOCOg}}$). RE tasks also receive the worst effect from the GQA task. Tasks in this group usually have a positive outcome on the tasks in other groups, according to rows 8–12 in Table 2.3. This may be because of the nature of the group: RE tasks aim to find image regions given a text, which can be helpful to VQA, IR, and MV.

2.5.3 Main observations

Observation 1. Intra-group analysis: tasks in the same group tend to improve each other, but not always. Tasks in the IR, MV, and RE groups help other tasks in the same group to get better performance. However, tasks in the VQA group show different behavior: only half of the intra-class relationships are positive. This indicates that: 1) the defined task groups based on shared goals may be superficial and not a good representation of the internal type of knowledge in each task, and 2) having a shared goal may be favorable, but it is not enough for successfully transferring knowledge between tasks.

Observation 2. Inter-group analysis: some groups are more prone to help, while others do disservice. For example, while tasks in the RE group usually give a positive effect on most of the tasks that are in other groups, tasks in the VQA group produce no benefit to the tasks in the RE group, and only one task in the MV group (NLVR2) give slightly positive effect to a task in the RE group (refCOCOg). This indicates that the knowledge in certain groups, such as RE, may be more general, and thus easier to transfer, than task-specific knowledge from other, *e.g.*, VQA, groups.

Observation 3. Benefits in knowledge transferability are not reciprocal. For example, VQA v2 receives a positive effect from all of the other 11 tasks, but it contributes negatively to most of these tasks, except GQA, NLVR2, and SNLI-VE. The same happens between the MV and RE groups. RE consistently improves the MV group, including the best effect on SNLI-VE from refCOCO+. However, the MV group harms all the tasks in the RE group except $m_{\text{NLVR2}\rightarrow\text{refCOCOg}}$. This is consistent with the observations in [35].

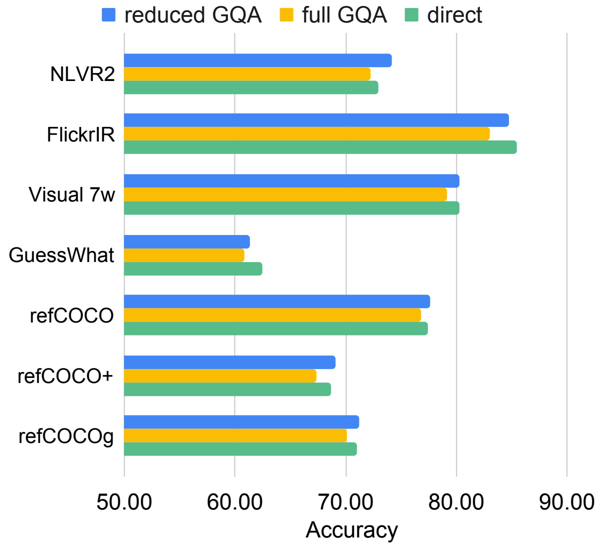


Figure 2.4: Accuracy of seven tasks pre-trained with a smaller set of GQA (reduced GQA), the full set of GQA (full GQA), and without pre-training (direct).

Observation 4. The best effect tends to come within the group, while the worst effect is usually from GQA. The best results for each task usually are from a source task in the same group, which reinforces the idea that tasks with the same target tend to benefit each other more (Observation 1). The worst results, however, are usually caused by GQA. Many reasons may cause this, such as the difference in the data scale between GQA and the rest of the tasks, or the knowledge for solving GQA may be too specific. To better understand the phenomena, we conduct additional experiments in Section 2.5.4 and Section 2.5.5.

2.5.4 Data scale

Next, we investigate the effect of the data scale on the transferability of knowledge. As discussed in Section 2.5.3, GQA pre-training tends to harm many of the rest tasks. We speculate that one of the reasons may be because GQA has a much larger training set than the other tasks. To investigate this hypothesis, we use GQA as the source task and downsample its training set

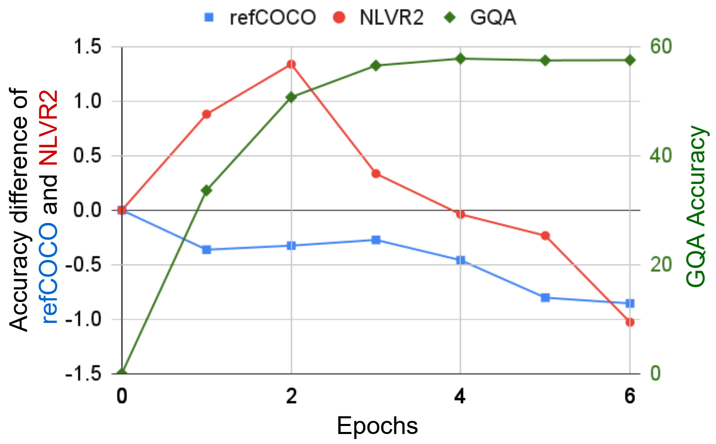


Figure 2.5: Accuracy on refCOCO (■) and NLVR2 (●) fine-tuned with m_{GQA} after different epochs of pre-training. As a reference, the accuracy of GQA (⧫) is also shown.

from $962,928$ to $96,221$, which is close to the scale of seven tasks: refCOCO, refCOCO+, refCOCOg, Visual7W, GuessWhat, NLVR2, and Flickr30K IR. We pre-train models with the reduced and full GQA training sets. The full GQA model and the reduced GQA model are then trained again on the seven tasks above in the same way as in Section 2.4.3. We also compare them against their direct models.

Figure 2.4 shows the accuracy of these seven tasks pre-trained on the reduced GQA, the full GQA, and the direct models. All models pre-trained on the reduced GQA get better performance than those pre-trained on the full GQA, which indicates that the data scale is a crucial factor in the transferability. When comparing the models derived from the reduced GQA with the direct models, the reduced models improve the performance for four of seven tasks (NLVR2, refCOCO, refCOCO+, and refCOCOg), showing that GQA can also contribute positively as a source task. The results show some similar phenomena to [93], as the large data scale may not necessarily generate better results.

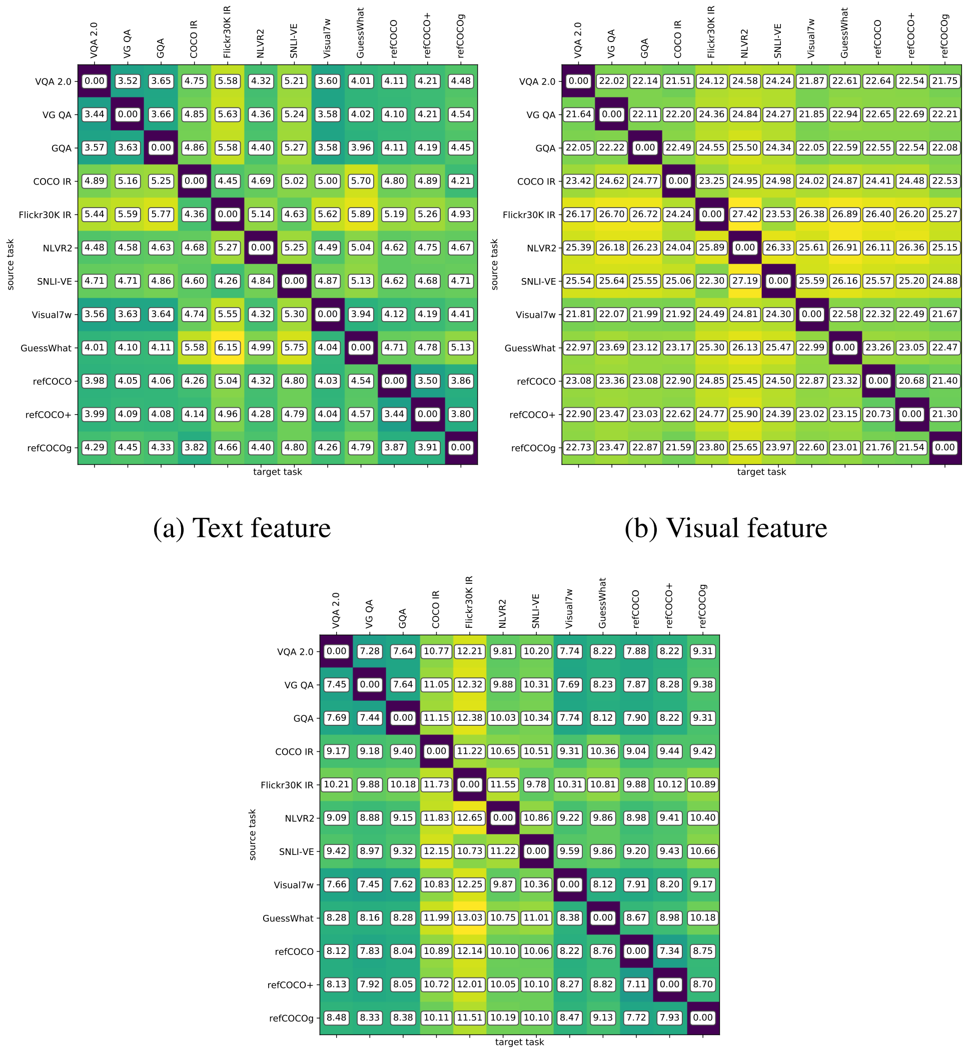
2.5.5 Training epoch

We finally explore the relationship between the number of training epochs of the source task and the success of the knowledge transferred in the target task. We conjecture that, for a target task that receives a negative effect from a source task, the more a model learns from the source task, the worse the model performs on the target task. To verify this, we use GQA as the source task, which tends to give the most negative effect to the other tasks. Furthermore, we choose refCOCO and NLVR2 as the target tasks, which get the worst performance from GQA. We get six pre-trained m^e_{GQA} models, where the number of epochs is $e = \{1, \cdots, 6\}$. The higher the epoch, the more knowledge from GQA m^e_{GQA} learns. We transfer these models to refCOCO and NLVR2.

The results are illustrated in Figure 2.5. The blue ■ and red ● are the scores for refCOCO and NLVR2, respectively, which for visibility are shown as the difference with respect to the direct model, *i.e.*$a = \text{Acc}^{(e)}_{\text{GQA}\to t} - \text{Acc}_{\text{GQA}}$, where $\text{Acc}^{(e)}_{\text{GQA}\to t}$ is the accuracy of model pre-trained with GQA for e epochs and fine-tuned with task t; $\text{Acc}^{(0)}_{\text{GQA}\to t}$ is the model that has no training on GQA, *i.e.*the direct model; and t is either refCOCO and NLVR2. For comparison, we also show the GQA accuracy ($\text{Acc}^{(e)}_{\text{GQA}}$, green ⧫). Both tasks get lower scores than the direct model when using a model trained on GQA for more than four epochs. In the case of refCOCO, it gets an inferior performance in all training epochs. In contrast, NLVR2 is improved by more than 1% from GQA pre-trained for two epochs, showing that the knowledge from GQA does not always have a negative effect.

2.5.6 Data domain similarity

Data domain distance We explore how the similarity of the data domain affects the transferability between vision-and-language tasks. On the one hand, some tasks take images from the same image dataset, *e.g.*, images in VQA 2.0, COCO IR, Visual7W, and GuessWhat are all from the MSCOCO [94] dataset. On the other hand, some tasks (*e.g.*, refCOCO, refCOCO+, and refCOCOg) have similar text data. Intuitively, two tasks with similar data domains may face



	VQA 2.0	VG QA	GQA	COCO IR	Flickr30K IR	NLVR2	SNLI-VE	Visual7w	GuessWhat	refCOCO	refCOCO+	refCOCOg
VQA 2.0	0.00	3.52	3.65	4.75	5.58	4.32	5.21	3.60	4.01	4.11	4.21	4.48
VG QA	3.44	0.00	3.66	4.85	5.63	4.36	5.24	3.58	4.02	4.10	4.21	4.54
GQA	3.57	3.63	0.00	4.86	5.58	4.40	5.27	3.58	3.96	4.11	4.19	4.45
COCO IR	4.89	5.16	5.25	0.00	4.45	4.69	5.02	5.00	5.70	4.80	4.89	4.21
Flickr30K IR	5.44	5.59	5.77	4.36	0.00	5.14	4.63	5.62	5.89	5.19	5.26	4.93
NLVR2	4.48	4.58	4.63	4.68	5.27	0.00	5.25	4.49	5.04	4.62	4.75	4.67
SNLI-VE	4.71	4.71	4.86	4.60	4.26	4.84	0.00	4.87	5.13	4.62	4.68	4.71
Visual7w	3.56	3.63	3.64	4.74	5.55	4.32	5.30	0.00	3.94	4.12	4.19	4.41
GuessWhat	4.01	4.10	4.11	5.58	6.15	4.99	5.75	4.04	0.00	4.71	4.78	5.13
refCOCO	3.98	4.05	4.06	4.26	5.04	4.32	4.80	4.03	4.54	0.00	3.50	3.86
refCOCO+	3.99	4.09	4.08	4.14	4.96	4.28	4.79	4.04	4.57	3.44	0.00	3.80
refCOCOg	4.29	4.45	4.33	3.82	4.66	4.40	4.80	4.26	4.79	3.87	3.91	0.00

(a) Text feature

	VQA 2.0	VG QA	GQA	COCO IR	Flickr30K IR	NLVR2	SNLI-VE	Visual7w	GuessWhat	refCOCO	refCOCO+	refCOCOg
VQA 2.0	0.00	22.02	22.14	21.51	24.12	24.58	24.24	21.87	22.61	22.64	22.54	21.75
VG QA	21.64	0.00	22.11	22.20	24.36	24.84	24.27	21.85	22.94	22.65	22.69	22.21
GQA	22.05	22.22	0.00	22.49	24.55	25.50	24.34	22.05	22.59	22.55	22.54	22.08
COCO IR	23.42	24.62	24.77	0.00	23.25	24.95	24.98	24.02	24.87	24.41	24.48	22.53
Flickr30K IR	26.17	26.70	26.72	24.24	0.00	27.42	23.53	26.38	26.89	26.40	26.20	25.27
NLVR2	25.39	26.18	26.23	24.04	25.89	0.00	26.33	25.61	26.91	26.11	26.36	25.15
SNLI-VE	25.54	25.64	25.55	25.06	22.30	27.19	0.00	25.59	26.16	25.57	25.20	24.88
Visual7w	21.81	22.07	21.99	21.92	24.49	24.81	24.30	0.00	22.58	22.32	22.49	21.67
GuessWhat	22.97	23.69	23.12	23.17	25.30	26.13	25.47	22.99	0.00	23.26	23.05	22.47
refCOCO	23.08	23.36	23.08	22.90	24.85	25.45	24.50	22.87	23.32	0.00	20.68	21.40
refCOCO+	22.90	23.47	23.03	22.62	24.77	25.90	24.39	23.02	23.15	20.73	0.00	21.30
refCOCOg	22.73	23.47	22.87	21.59	23.80	25.65	23.97	22.60	23.01	21.76	21.54	0.00

(b) Visual feature

	VQA 2.0	VG QA	GQA	COCO IR	Flickr30K IR	NLVR2	SNLI-VE	Visual7w	GuessWhat	refCOCO	refCOCO+	refCOCOg
VQA 2.0	0.00	7.28	7.64	10.77	12.21	9.81	10.20	7.74	8.22	7.88	8.22	9.31
VG QA	7.45	0.00	7.64	11.05	12.32	9.88	10.31	7.69	8.23	7.87	8.28	9.38
GQA	7.69	7.44	0.00	11.15	12.38	10.03	10.34	7.74	8.12	7.90	8.22	9.31
COCO IR	9.17	9.18	9.40	0.00	11.22	10.65	10.51	9.31	10.36	9.04	9.44	9.42
Flickr30K IR	10.21	9.88	10.18	11.73	0.00	11.55	9.78	10.31	10.81	9.88	10.12	10.89
NLVR2	9.09	8.88	9.15	11.83	12.65	0.00	10.86	9.22	9.86	8.98	9.41	10.40
SNLI-VE	9.42	8.97	9.32	12.15	10.73	11.22	0.00	9.59	9.86	9.20	9.43	10.66
Visual7w	7.66	7.45	7.62	10.83	12.25	9.87	10.36	0.00	8.12	7.91	8.20	9.17
GuessWhat	8.28	8.16	8.28	11.99	13.03	10.75	11.01	8.38	0.00	8.67	8.98	10.18
refCOCO	8.12	7.83	8.04	10.89	12.14	10.10	10.06	8.22	8.76	0.00	7.34	8.75
refCOCO+	8.13	7.92	8.05	10.72	12.01	10.05	10.10	8.27	8.82	7.11	0.00	8.70
refCOCOg	8.48	8.33	8.38	10.11	11.51	10.19	10.10	8.47	9.13	7.72	7.93	0.00

(c) Vision-and-language feature

Figure 2.6: Domain distance between 12 vision-and-language tasks. We calculate the distances of the vision-and-language feature (*i.e.*, fused feature), text feature, and visual feature. Each of the blocks shows the domain distance of $D_{row \rightarrow column}$. Please note that the distance of $D_{row \rightarrow column}$ and $D_{row \rightarrow column}$ may not be the same, as 2.2 is finding the closest source task sample for each target task sample.

smaller domain shifts when fine-tuning and thus tend to have a higher probability of helping each other. To explore this, we randomly take $1,000$ samples from the training set of each task, use VilBERT [31] to extract the features from the samples, and calculate the domain distance as Mensink *et al.* [37] did:

$$D(t|s) = \frac{1}{|t|} \sum_{z_t} (\underset{z_s}{min}\ d(f_{z_t}, f_{z_s})), \tag{2.2}$$

where $d()$ is the Euclidean distance, z_s and z_t are the samples from source task s and t, respectively. Please note that VilBERT [31] is not pre-trained on any of the 12 tasks.

In contrast to the analysis of vision-only tasks in [37], vision-and-language tasks concerning the visual and text data, as well as the hidden relation between both data. Thus, our domain distance exploration involves the analysis of the vision feature, text feature, and fused vision-and-language feature, respectively. The results are illustrated in Figure 2.6.

In general, on the one hand, some tasks are close to other tasks from the data domain view, *e.g.*, refCOCO, refCOCO+, and refCOCOg are closed the each other in all three figures. The closed distance of these three tasks is because of the same image origination and the similar text collection process. Since refCOCO, refCOCO+, and refCOCOg are in the same data scale, the closed distance of the data domain becomes one of the factors that these three tasks can help each the get better performance.

Figure 2.6c illustrates that Flickr30K IR has a larger distance as the target task compared with other tasks. This indicates that the data from other tasks are more different from Flickr30K IR, and the transfer learning toward Flickr30K IR may bring fewer benefits. As shown in Columns 4-5 in Table 2.3, GuessWhat and NLVR2, the largest and the second largest distance toward Flickr30K IR, while having the same data scale as Flickr30K IR, show the negative effect on Flickr30K IR.

However, the domain distance is not as strong a decisive factor as the data scale. Figure 2.6c illustrates that GQA is not much different from most of the other tasks. Instead, GQA has the top-5 closest data domain to NLVR2, Visual7W, GuessWhat, refCOCO, refCOCO+, and refCOCOg. However, GQA causes the worst results for these tasks. This fact indicates that the data scale may have a more decisive effect on knowledge transferability.

Appearance distribution Furthermore, we explore the data domain distribution in the embedding feature space, namely, the appearance distribution of all 12 tasks. The analysis is based on the sample feature in the last experiments, *i.e.*, we randomly take $1,000$ samples from the training set of each task and use VilBERT [31] to extract the features from the samples. To make a better visualization of the distribution, we use t-SNE to plot the samples into 2D figures, as shown in Figures 2.7 and 2.8.

Figure 2.7 illustrates the data appearance distribution of all 12 tasks. We find that in many cases, a sample from one task may be closer to the samples in another task, even though the two tasks have different data origination refer to Table 2.1 (*e.g.*, although NLVR2 (the brown squares) and VQA 2.0 (the blue dots) have different data origination, their samples are closed to each other). This indicates that the 12 tasks have close appearance distribution, and they are very similar to each other.

The similarity of data distribution not only appears in the vision-and-language feature (Figure 2.7c) but also in the vision-only feature (Figure 2.7a) and text-only feature (Figure 2.7b). This indicates that the 12 tasks are similar to each other from both the image and text domains.

Figure 2.8 further shows the appearance distribution of vision-and-language features for each task group. On the one hand, tasks in the VQA group and RE group have similar appearance distribution to other tasks in the same group, which indicates that tasks in these task groups are similar from the data domain view. On the other hand, tasks in the IR group and MV group have different appearance distribution from each other tasks, which indicate that the data of these tasks are different from each other and the knowledge from one task may not be much help to the other task. However, we can still observe from Table 2.3 that COCO IR and Flickr30K IR help each other to get the best performance, and NLVR2 and SNLI-VE help each other to improve. These results show that task similarity may be more decisive than data domain similarity in improving the model performance, *i.e.*, task similarity affects knowledge transferability more than data domain similarity.

2.5.7 Visual results

Figure 2.9 shows predictions on refCOCO with the direct model m_{refCOCO}, one-hop model $m_{\text{refCOCO+}\to\text{refCOCO}}$, one-hop model $m_{\text{GQA}\to\text{refCOCO}}$, and the ground truth. The IOU between the prediction and the ground truth is shown under the respective image. Whereas refCOCO+ helps to find more accurate regions and to obtain higher IOU, GQA misleads the task to smaller or even wrong regions. For example, in the image in the middle, although the direct model finds the region with the right person, refCOCO+ helps to find a more accurate region, but GQA predicts the wrong person. The same behavior can be observed in the last two images.

Figure 2.10 shows examples of the GQA validation set with the direct model m_{GQA}, one-hop model $m_{\text{VQA v2}\to\text{GQA}}$, and one-hop model $m_{\text{GuessWhat}\to\text{GQA}}$. We show the confidence of prediction for the ground truth class (Conf. of GT) under each example. VQA v2 gives the most positive effect to GQA, while GuessWhat gives the most negative effect. For example, in the second and third images from the left, GuessWhat induces wrong answers, whereas, in the last two images, VQA v2 helps to find the correct answers and improve the prediction with respect to the direct model.

2.6 Experiments on ViLT

In this section, we introduce our experiments on ViLT [33], which is also widely applied to multiple vision-and-language tasks (VQA v2 [3], COCO IR [48], Flickr30K IR [52], and NLVR2 [8]). To explore more tasks, we follow the instructions of ViLT and expand the model to support VG QA [5] and SNLI-VE [9]. Since ViLT directly uses the whole image (instead of the regions of the image) as the visual input, this model is not able to do referring expressions tasks. In general, we conduct a knowledge transferability exploration based on ViLT with six different vision-and-language tasks in three types:

- VQA: VQA v2 [3] and VG QA [5]
- IR: COCO IR [48] and Flickr30K IR [52]

Table 2.4: Knowledge transferability results in ViLT model. Results of $m_{row\,\rightarrow\,column}$ (row 2-7, green or red color) are compared with the *direct* model m_{column} (row 1, blue color) and assigned green (when the average score is higher than m_{column}) or red (when the average score is lower than m_{column}). Deep green/red shows the best/worst score in each column.

ViLT		Target task t					
		VQA v2	VG QA (Val)	COCO IR	Flickr30K IR	NLVR2	SNLI-VE
direct model m_t		71.32	35.10	93.10	88.80	76.55	72.58
Source task s	VQA v2	-	35.20	90.60	78.82	73.29	73.06
	VG QA	69.48	-	89.24	76.26	70.34	72.29
	COCO IR	71.10	34.99	-	79.74	72.31	72.94
	Flickr30K IR	69.93	34.79	89.14	-	73.59	73.28
	NLVR2	70.87	34.66	89.46	78.20	-	72.67
	SNLI-VE	65.97	33.25	85.46	74.60	63.46	-

- MV: NLVR2 [8] and SNLI-VE [9]

We follow the same experimental setting of ViLT [2] to fine-tune the model on each task and make the experiments with the same methodology in Section 2.4, *i.e.*, train *direct* models m_t and *one-hop* models $m_{s \rightarrow t}$, then compare their performance. We also use the same evaluation data and metrics as listed in Table 2.1.

2.6.1 Main observations

The experimental results are listed in Table 2.4. We have the following observations from these results:

Observation 1. Most of the tasks tend to get bad effects from other tasks. Compared to the performance of *direct* model m_t, the performance of *one-hop* models $m_{s\,\rightarrow\,t}$ in VQA v2, COCO IR, Flickr30K, and NLVR2 show significant decreases (more than 3%). These

[2] https://paperswithcode.com/method/vilt

decreases may related to the catastrophic forgetting since ViLT is trained on more datasets than VilBERT. The result may indicate that catastrophic forgetting is one of the factors to affect knowledge transferability.

Observation 2. VG QA and SNLI-VE get help from some tasks to get better performance. These results indicate that the knowledge from other vision-and-language tasks still may benefit models in solving some target tasks, even though catastrophic forgetting mitigates benefits.

Observation 3. SNLI-VE tends to get benefits from other tasks while having bad effects on them. This observation is similar to VilBERT's results in Observation 3 in section 2.5.3, which indicates that the benefits of knowledge transferability are not reciprocal. The phenomenon may also indicate that the knowledge in SNLI-VE is different from other tasks, but other tasks may involve the knowledge for solving SNLI-VE.

2.7 Limitations and future work

More complex transfer scenarios. In this chapter, we mainly focus on one-to-one transfer learning as it is the most widely used strategy in knowledge transferability. In our future work, we would like to make comprehensive explorations on other knowledge transfer strategies (*e.g.*, more-to-more knowledge transfer scenario) in our future work.

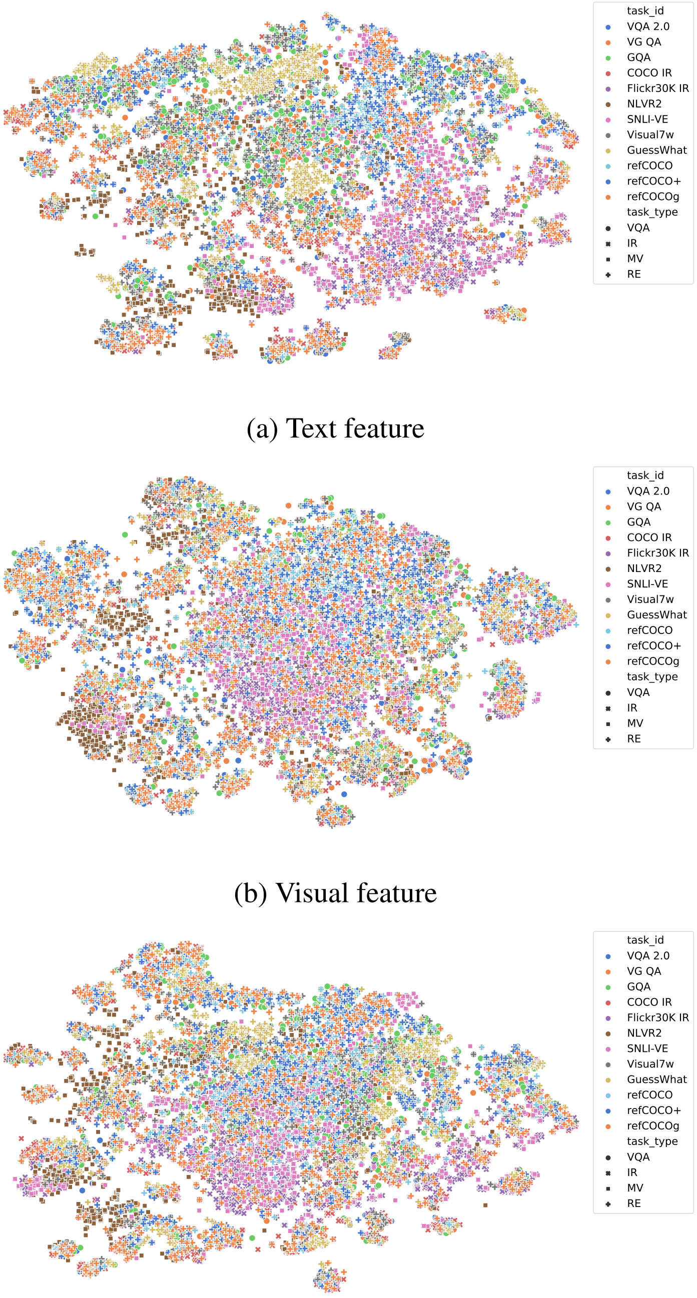
Optimal transfer point. From our experiments, we observe that although GQA usually brings negative effects to other tasks, modifying the training setting of GQA during the pre-training step (*e.g.*, decreasing the data scale of the GQA dataset or the training epochs of GQA) could make GQA a positive transfer to other tasks. These results may indicate that there is an optimal transfer point of knowledge transfer when the model learns just enough knowledge from the source task s but does not overfit. However, finding the optimal transfer point is a challenging task as the verification of the performance of model $m_{s \to t}$ needs another round of training and evaluation. We consider finding the optimal transfer point as one of our future work.

Explorations on large-scale pre-training models. Large-scale pre-training models, such as CLIP and BLIP-2, may contain rich and valuable knowledge that can bring more benefits to vision-and-language tasks. We are aware that it is important to involve the exploration based on these large-scale pre-training models. Actually, recent studies show that studies [95, 96] show that the knowledge from large-scale datasets is sensitive and easily suffers from catastrophic forgetting during the finetuning process. Since we focus more on the knowledge transferability between vision-and-language tasks, we would like to set the exploration of knowledge transferability from pre-text tasks or noisy but large-scale datasets as our future work.

2.8 Summary

We studied the knowledge transferability among 12 vision-and-language tasks. We confirmed that different tasks have different effects on each other, and the selection of tasks for knowledge transfer should be made carefully. Furthermore, we observed some interesting insights about knowledge transferability, *e.g.*, the tasks in the image retrieval and referring expressions groups tend to have a positive impact on other tasks, while the tasks in the visual question answering and multi-modal verification group give a negative contribution. The scale of datasets, training epochs, data domain similarity, and the difference in their goals may cause this divergence.

In general, this chapter sheds light on the knowledge transferability of vision-and-language tasks, including those factors such as data scale and training epoch that may affect the transferability. We hope our work can bring inspiration to the fields of knowledge transferability in vision-and-language tasks, especially for the topic of seeking knowledge from similar tasks for positive transfer.

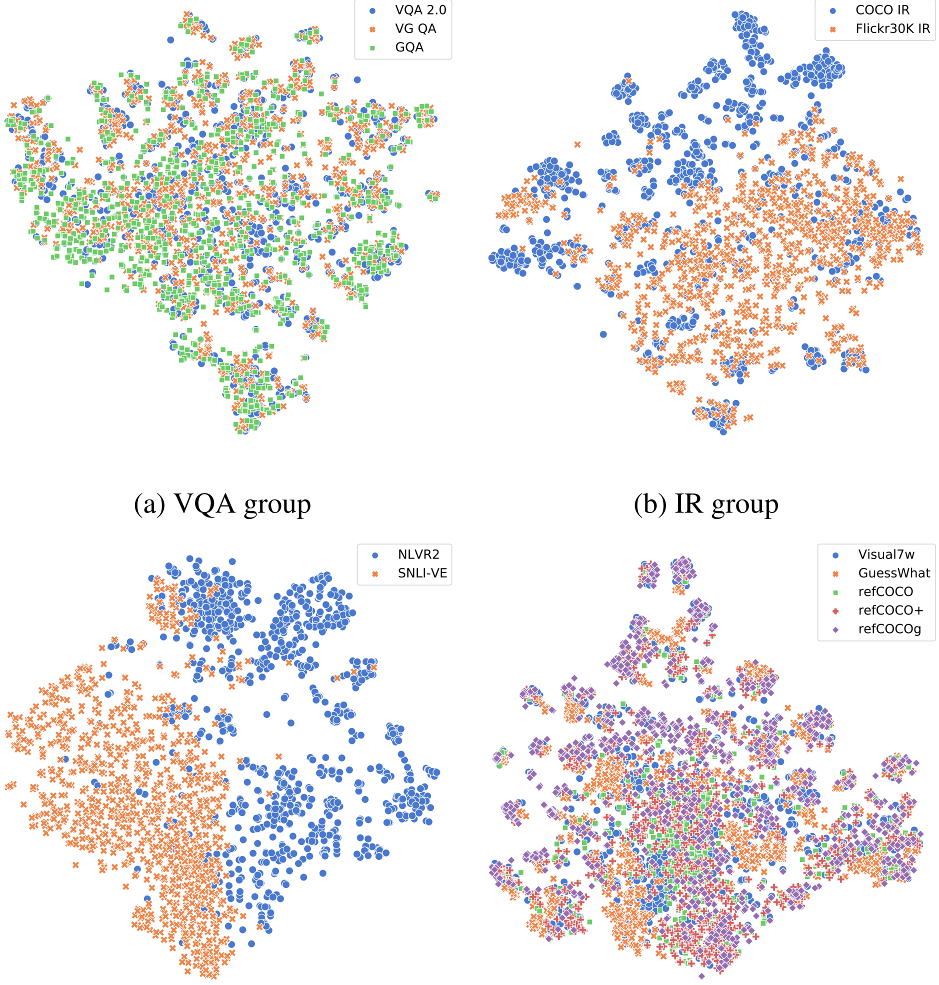


(a) Text feature

(b) Visual feature

(c) Vision-and-language feature

Figure 2.7: Appearance distribution of all 12 vision-and-language tasks. In this figure, different tasks get different colors while tasks within the same task group share the same shape of markers.



(a) VQA group

(b) IR group

(c) MV group

(d) RE group

Figure 2.8: Appearance distribution within four vision-and-language task groups. In this figure, different tasks get different colors and different shapes of markers.

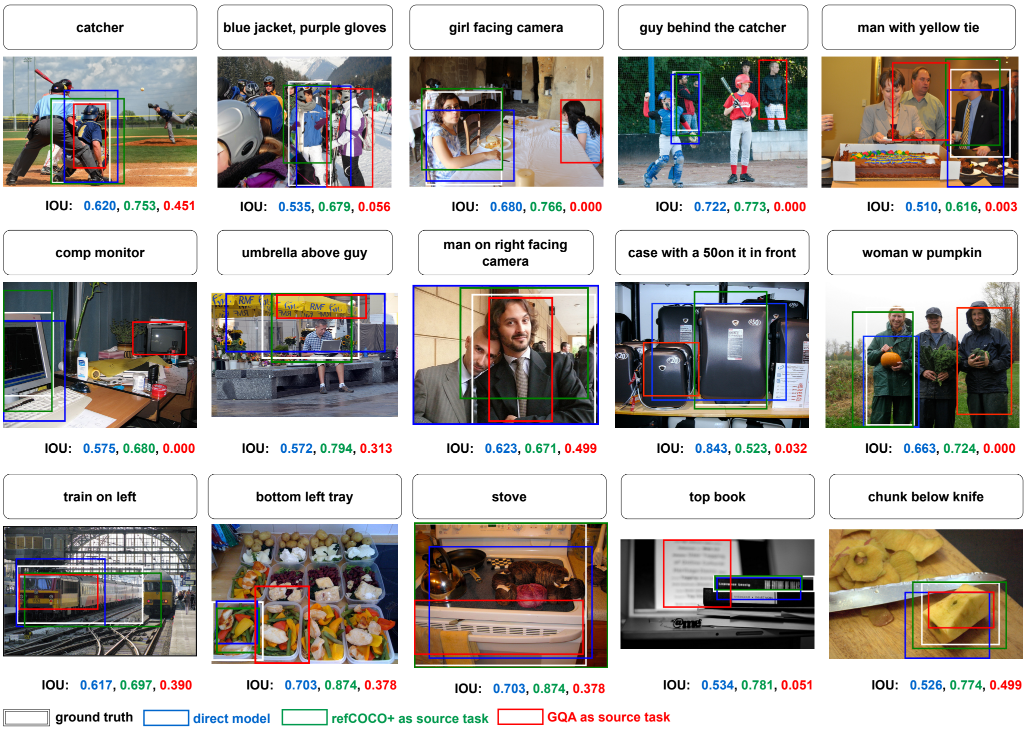


Figure 2.9: Example of the results on refCOCO. With the caption on top of the image, different models find different regions on the image. It is easy to see that refCOCO+ helps refCOCO to get a more accurate prediction, while GQA misleading refCOCO to some wrong regions.

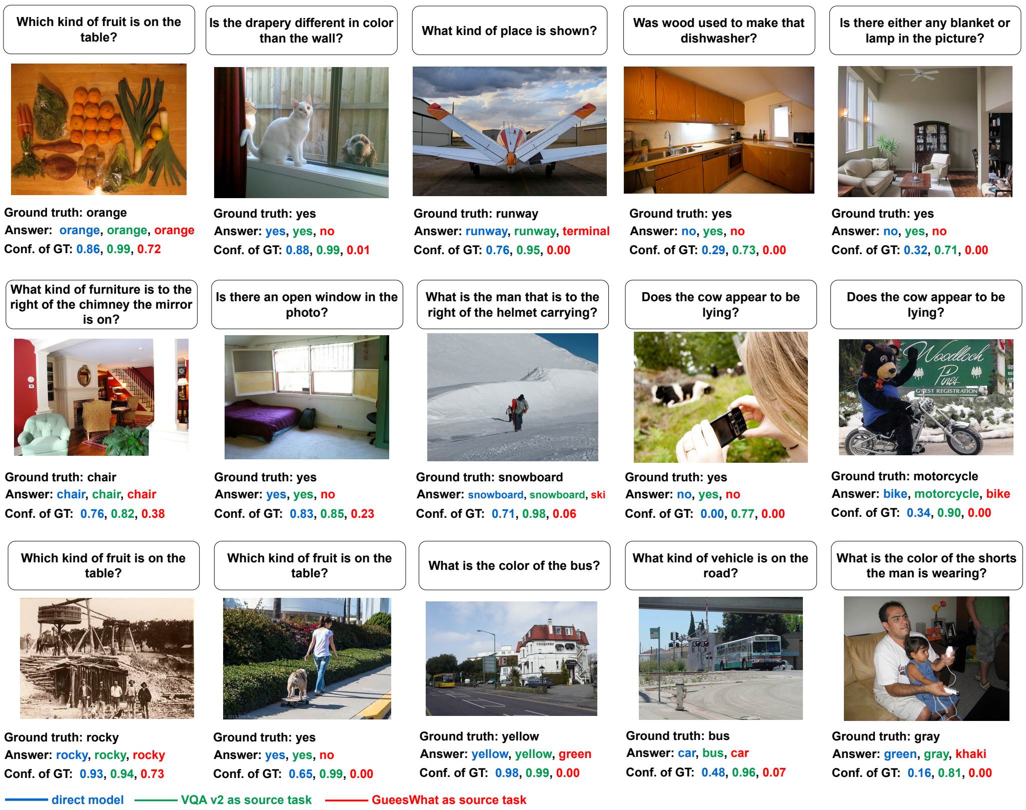

Which kind of fruit is on the table?
Ground truth: orange
Answer: orange, orange, orange
Conf. of GT: 0.86, 0.99, 0.72
Is the drapery different in color than the wall?
Ground truth: yes
Answer: yes, yes, no
Conf. of GT: 0.88, 0.99, 0.01
What kind of place is shown?
Ground truth: runway
Answer: runway, runway, terminal
Conf. of GT: 0.76, 0.95, 0.00
Was wood used to make that dishwasher?
Ground truth: yes
Answer: no, yes, no
Conf. of GT: 0.29, 0.73, 0.00
Is there either any blanket or lamp in the picture?
Ground truth: yes
Answer: no, yes, no
Conf. of GT: 0.32, 0.71, 0.00
What kind of furniture is to the right of the chimney the mirror is on?
Ground truth: chair
Answer: chair, chair, chair
Conf. of GT: 0.76, 0.82, 0.38
Is there an open window in the photo?
Ground truth: yes
Answer: yes, yes, no
Conf. of GT: 0.83, 0.85, 0.23
What is the man that is to the right of the helmet carrying?
Ground truth: snowboard
Answer: snowboard, snowboard, ski
Conf. of GT: 0.71, 0.98, 0.06
Does the cow appear to be lying?
Ground truth: yes
Answer: no, yes, no
Conf. of GT: 0.00, 0.77, 0.00
Does the cow appear to be lying?
Ground truth: motorcycle
Answer: bike, motorcycle, bike
Conf. of GT: 0.34, 0.90, 0.00
Which kind of fruit is on the table?
Ground truth: rocky
Answer: rocky, rocky, rocky
Conf. of GT: 0.93, 0.94, 0.73
Which kind of fruit is on the table?
Ground truth: yes
Answer: yes, yes, no
Conf. of GT: 0.65, 0.99, 0.00
What is the color of the bus?
Ground truth: yellow
Answer: yellow, yellow, green
Conf. of GT: 0.98, 0.99, 0.00
What kind of vehicle is on the road?
Ground truth: bus
Answer: car, bus, car
Conf. of GT: 0.48, 0.96, 0.07
What is the color of the shorts the man is wearing?
Ground truth: gray
Answer: green, gray, khaki
Conf. of GT: 0.16, 0.81, 0.00
direct model
VQA v2 as source task
GueesWhat as source task

Figure 2.10: Example of the results on GQA. With the question on top of the image, different models predict the answer based on the image. The predictions from m_{GQA}, $m_{\text{VQA v2}\rightarrow\text{GQA}}$, and $m_{\text{GuessWhat}\rightarrow\text{GQA}}$, as well as the confidence score of the ground truth class (Conf. of GT), are shown under the examples, respectively. It is easy to see that VQA v2 helps GQA to get a more accurate prediction, while GuessWhat misleads GQA to get a low confidence score in the ground truth class.

Chapter 3

Detecting Emotional Stimuli in Artworks

3.1 Overview

We introduce an emotional stimuli detection task that targets extracting emotional regions that evoke people's emotions (i.e., emotional stimuli) in artworks. This task offers new challenges to the community because of the diversity of artwork styles and the subjectivity of emotions, which can be a suitable testbed for benchmarking the capability of the current neural networks to deal with human emotion. For this task, we construct a dataset called APOLO for quantifying emotional stimuli detection performance in artworks by crowd-sourcing pixel-level annotation of emotional stimuli. APOLO contains 6,781 emotional stimuli in 4,718 artworks for validation and testing. We also evaluate eight baseline methods, including a dedicated one, to show the difficulties of the task and the limitations of the current techniques through qualitative and quantitative experiments. Our data and code are available in https://github.com/Tianwei3989/apolo.

Analyzing artworks in machine learning is a challenging task. Compared to photographs, artworks do not only depict real-world concepts, such as humans, animals, and natural scenes, but also represent humane contents, such as feelings, attitudes, and faiths. The richness in the representations and the diversity of styles make artworks the ideal testbed to study new challenges related to human emotion understanding in machine learning.

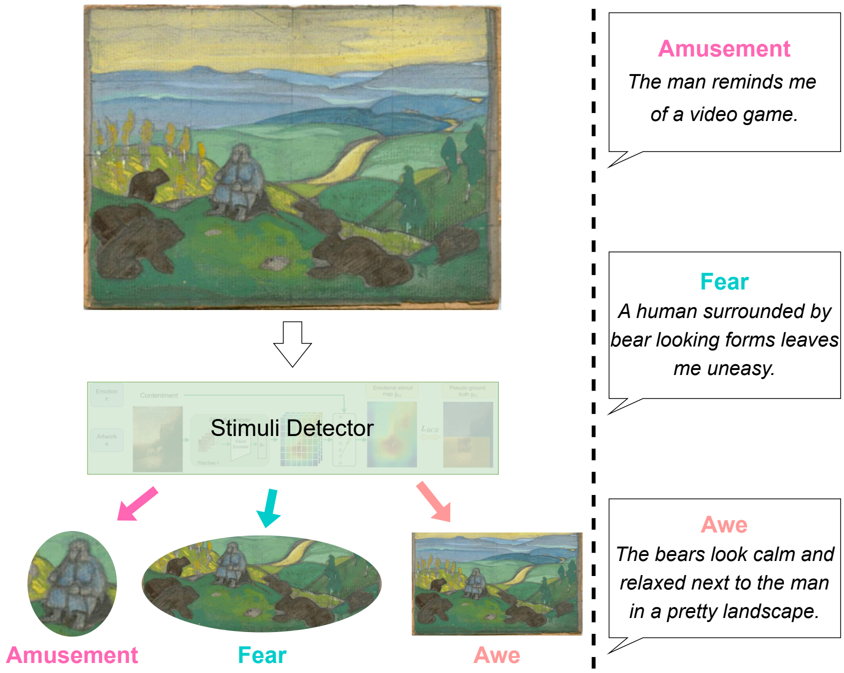

Stimuli Detector
Amusement
Fear
Awe
Amusement
The man reminds me of a video game.
Fear
A human surrounded by bear looking forms leaves me uneasy.
Awe
The bears look calm and relaxed next to the man in a pretty landscape.

Figure 3.1: A model detects emotional regions that evoke people's emotions (namely, emotional stimuli) from the given artwork. The utterances on the right side may be used as hints to spot emotional stimuli.

In recent years, many efforts have been paid to the field of artwork analysis [4, 87, 97–104], aiming to improve models' understanding of artworks and further extend models' capacity to support digital humanities, with tasks such as attribute identification [97, 101–104], object detection [99, 100] or artwork understanding through language [4, 87, 98]. Thanks to these studies, recent models have developed a reliable capacity to understand *objective* contents (*e.g.*, objects, attributes, descriptions) from the artworks. However, only a few studies [1, 105] are focusing on a more *subjective* and personal analysis, such as the relationship between artwork and emotions.

ArtEmis [1], as well as its extension ArtEmis V2.0 [105], are two datasets collected for studying of the relationship between artworks and emotions. The main focus is on the generation of emotional captions that can accurately capture the emotional influence of an artwork. However, a more in-depth analysis to uncover why and how emotions are evoked from the artworks is still not explored. In other words, ArtEmis and ArtEmis 2 show that models can generate emotional captions, but it is still unknown how the emotions are evoked from those artworks.

Artworks can easily elicit people's emotions, yet this elicitation process is complex and underexplored [106–109]. The appraisal theory toward artworks and emotions [106] says the emotion-evoking process is related to the viewer's analysis process: The emotions are evoked during the viewer's analysis process through the whole artwork. Thus, different analyses may lead to different emotions. For example, given the artwork in Figure 3.1, people may feel different emotions when the analysis concentrates on different visual concepts in the context of the artwork: if a viewer focuses on the distorted style of the person, a feeling of amusement may be evoked, while the bear-like brown figures may be linked with fear. Learning such processes could make models acquire knowledge about how human emotions are evoked and may improve models' capacity to utilize emotional stimuli. Such merits could be helpful for tasks related to emotions (*e.g.*, visual emotion recognition [110–114]) and tasks potentially involving emotion analysis (*e.g.*, image generation [20, 22]).

According to these observations, we propose a new task for emotional stimuli detection in artworks, in which a model is required to detect emotional stimuli from a given artwork, as shown in Figure 3.1. The task, which explores a machine's capacity to understand emotions and artworks, has two major challenges: First, differently from photorealistic images, *artworks are painted with a certain style*. For example, in Western art, Realism is one of the styles that may look more like a real photo, while Impressionism typically shows prominent brush strokes. Different styles lead to very different appearances of the same object. This style variation makes it harder to learn visual content (*e.g.*, objects) from artworks than photos [97, 100, 115]. Second, *emotions are subjective*. Different people may have different emotions evoked by the same artwork [1, 105]. This subjectivity makes the task unique, as an artwork can have multiple emotional stimuli for different emotions.

For this task, we construct a benchmark dataset, coined APOLO (**A**rtwork **P**rovoked em**O**tion Eva**L**uati**O**n), to evaluate models in both qualitative and quantitative ways. We build APOLO

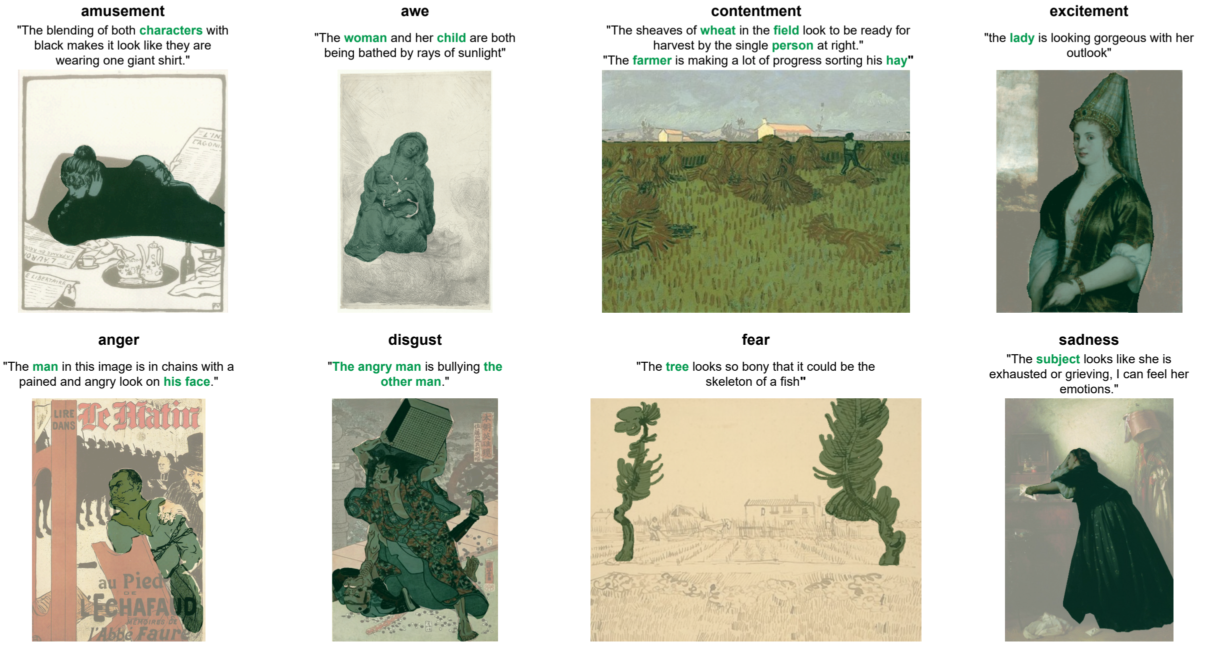

amusement
"The blending of both characters with black makes it look like they are wearing one giant shirt."
awe
"The woman and her child are both being bathed by rays of sunlight"
contentment
"The sheaves of wheat in the field look to be ready for harvest by the single person at right."
"The farmer is making a lot of progress sorting his hay"
excitement
"the lady is looking gorgeous with her outlook"
anger
"The man in this image is in chains with a pained and angry look on his face."
disgust
"The angry man is bullying the other man."
fear
"The tree looks so bony that it could be the skeleton of a fish"
sadness
"The subject looks like she is exhausted or grieving, I can feel her emotions."

Figure 3.2: Some samples in our dataset. The words and regions in green are the chosen noun phrases and the annotated emotional stimuli, respectively. If one artwork-emotion pair contains multiple utterances, the corresponding regions are then combined. We annotate regions for eight emotions from ArtEmis, except "something else."

on top of the ArtEmis dataset [1], which, for each artwork, includes emotion labels annotated by multiple annotators, and utterances (sentences) that explain why emotions are provoked. To further understand the stimuli that provoke emotions, APOLO includes pixel-level emotional stimuli annotations on the images of the test set. As a result, we collect $6,781$ emotional stimuli for $4,718$ artworks and 8 emotions. Our exhaustive control quality checks ensure the samples are balanced and reliable. To the best of our knowledge, this is the first dataset that offers pixel-level annotations of emotional stimuli in artworks.

Additionally, we explore multiple models for emotional stimuli detection, borrowed from related tasks, including object detection, referring expression, and saliency map detection. We also introduce a dedicated weakly supervised model as a baseline, which predicts emotional stimuli regions for each emotion without using region-level annotations for training. Our comprehensive experiments on APOLO show that the evaluated models can detect emotional stimuli even when not trained with region annotations. However, the emotional stimuli detection task is still challenging and with plenty of room for improvement. In addition, we explore how a text-to-image generative model, Stable Diffusion [22], handles emotions in the input prompts, observing that it fails to connect the emotional words in the input with the emotional stimuli in the generated images. We hope our work will help overcome this limitation in the future.

3.2 Related work

Visual emotion analysis Given an input image, visual emotion analysis aims to recognize emotions, analyze the emotional stimuli, and apply the recognized emotions to real-world applications (*e.g.*, psychological health [116, 117] and opinion mining [118, 119]) to improve the ability of emotional intelligence [120]. Most of the recent studies [110–114] use emotional stimuli to improve emotion recognition, but only a few efforts have been made to analyze how well the models detect such stimuli. To the best of our knowledge, only two datasets: EmotionROI [121] and EMOd [122] provide pixel-level annotations for evaluating emotional stimuli detection. However, they are both relatively small, offering $1,980$ and $1,019$ labeled images, respectively, consisting of social media images from the Internet.

Data scarcity is one of the main challenges in emotional stimuli detection. To overcome this problem, we propose two solutions: 1) transferring models from related tasks, and 2) designing a weakly supervised learning model that does not require costly pixel-level annotations for training. For evaluation, we collect a dataset with emotional stimuli annotations.

Artwork analysis Much effort has been dedicated to solving art-related problems with machine learning techniques, including style identification [97,123], object detection [97,99,100], instance-level recognition [124], or artwork description [87,98,125]. Concerning emotion analysis, some datasets [1, 97, 105, 126], including ArtEmis, contain labels with the emotion (*e.g.*, *amusement* and *fear*) that each artwork evokes. Nevertheless, the same artwork can evoke multiple emotions according to different regions of the image, a fact that has been unexplored in

Table 3.1: Summary of datasets for emotional stimuli detection. The source column indicates whether the images are from the social internet or artworks, while the ME (Multi-emotion) column indicates whether an image has annotations for multiple emotions.

	Samples	Images	Source	Emotions	ME
EmotionROI [121]	1,980	1,980	social	6	No
EMOd [122]	1,019	1,019	social	2+1	No
APOLO	**6,781**	**4,178**	**artwork**	**8**	**Yes**

current datasets. APOLO introduces a new challenge by investigating the connection between artworks and emotion at the pixel level.

3.3 Emotional stimuli detection

Our task aims to explore how a model can find the cues of the emotion elicitation process from the artwork, *i.e.*, the emotional stimuli. In general, we explore two separate scenarios: 1) emotional stimuli detection *without reference* (*i.e.*, utterances) and 2) emotional stimuli detection *with reference*. Ideally, a model should find emotional stimuli without reference, like humans. However, such models are rare since only a few studies are aimed at emotional stimuli detection. We thus also explore whether recent multimodal models can detect emotional stimuli by using the references.

Formally, let a, $e \in \mathcal{E}$, and u denote an artwork, its emotion label, and the utterance, which can be a set of sentences, in ArtEmis, where $\mathcal{E}$ is the set of the emotions. Then, $\mathcal{D}_\text{t}$ denote the training set of ArtEmis [1], where $\mathcal{D}_\text{t}$ contains triples (a, e, u). $\mathcal{D}_\text{v}$ and $\mathcal{D}_\text{e}$ denote the validation and test sets of APOLO, where both contain triples (a, e, u). As we presume that an artwork can evoke potentially *any* emotion depending on where the viewer focuses their attention, the emotional stimuli detection task can be formulated as a segmentation task given artwork a and emotion $e \in \mathcal{E}$, in which a model f_e predicts segment s that evoke emotion e as

$$\hat{s} = f_e(a), \tag{3.1}$$

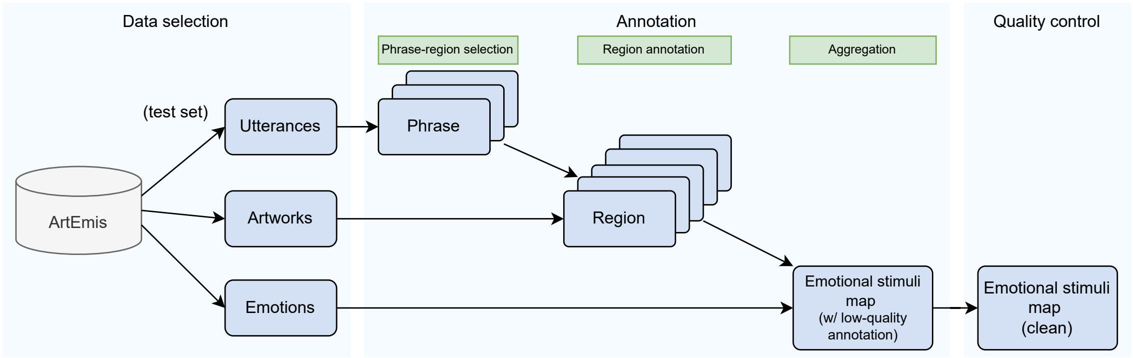


Figure 3.3: The workflow of APOLO dataset curation. In general, APOLO is extended from the ArtEmis dataset. We collect the annotation from Artemis's test set and further annotate the pixel-level emotional stimuli map from the artworks.

where $\hat{s}$ is the predicted segment.

This task can be extremely challenging as no cue is provided for specifying the regions that are involved given emotion e. We thus formulate a variant with reference by u, in which u is given to a model as an auxiliary cue for emotional stimuli detection, *i.e.*,

$$\hat{s} = f_e(a, u). \tag{3.2}$$

In both scenarios (emotional stimuli detection with and without reference), we can use $\mathcal{D}_t$ for training a model, but ground-truth segments are not available. $\mathcal{D}_\text{v}$ and $\mathcal{D}_\text{e}$ are solely used for validation and testing.

3.4 APOLO Dataset Curation

APOLO is a benchmark dataset for evaluating, both quantitatively and qualitatively, emotional stimuli detection in artworks. We utilize the test samples in ArtEmis [1], with $39,850$ explanatory utterances and emotional responses related to $8,003$ artworks from WikiArt.[1] ArtEmis is annotated with nine emotions: *amusement*, *anger*, *awe*, *contentment*, *disgust*, *excitement*, *fear*,

[1] https://www.wikiart.org/

sadness, and *something else*. As shown in Figure 3.2, the utterances are explanations of why a certain emotion is evoked by the artwork, which is usually related to its emotional stimuli. We observe that the utterances potentially align with the viewers' analysis processes and are related to a certain emotion that is specified by the emotion label and evoked by the artwork. Furthermore, the utterances tend to describe certain regions, which leads to a certain emotion, in the artwork. These features may help models to learn how humans perceive emotion from or associate emotion with such regions.

Toward this end, we construct a pixel-level emotional stimuli dataset, APOLO, by asking 91 annotators in Amazon Mechanical Turk [2] to identify the visual concepts that involve the utterances and to annotate the visual concepts at the pixel level, as shown in Figure 3.2. We show its details in Table 3.1.

We only collect validation and test sets by randomly sampling ArtEmis's annotation, considering 1) the evaluation could be applied to recent large models (*e.g.*, CLIP [18]) that are hard to train and 2) the cost of pixel-level annotation.

3.4.1 Data selection

To curate our annotations from ArtEmis, we annotate paintings from the first eight emotions from ArtEmis' nine emotions, *i.e.*, the emotions of *amusement*, *anger*, *awe*, *contentment*, *disgust*, *excitement*, *fear*, and *sadness*. We filter out samples with *something else* label, as we found from the associated utterance that their interpretation of the emotion is not trivial and annotators may not capture the clear ideas from them. For each of the other eight emotion labels, we randomly choose about $1,200$ artwork-utterance pairs from the ArtEmis test set, except for emotion *anger*, which only contains 672 artwork-utterance pairs. Overall, we select $9,599$ samples.

[2] https://www.mturk.com/

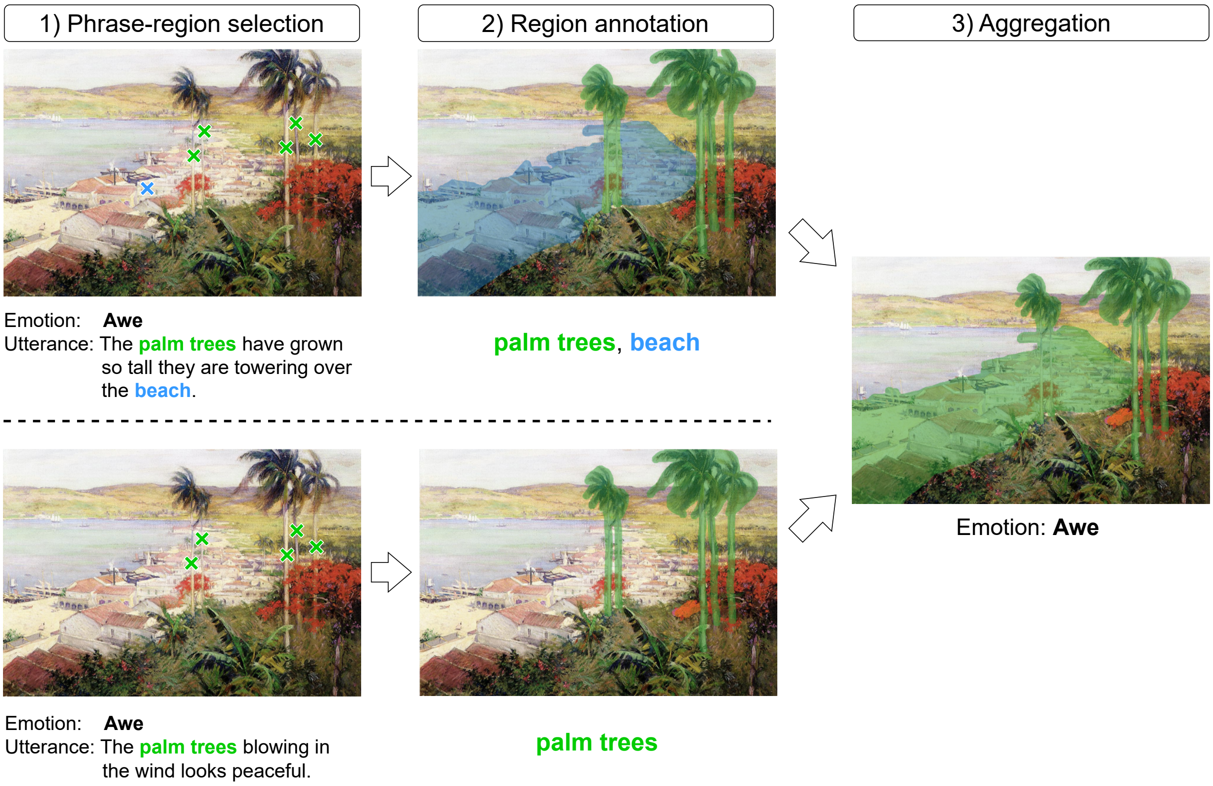
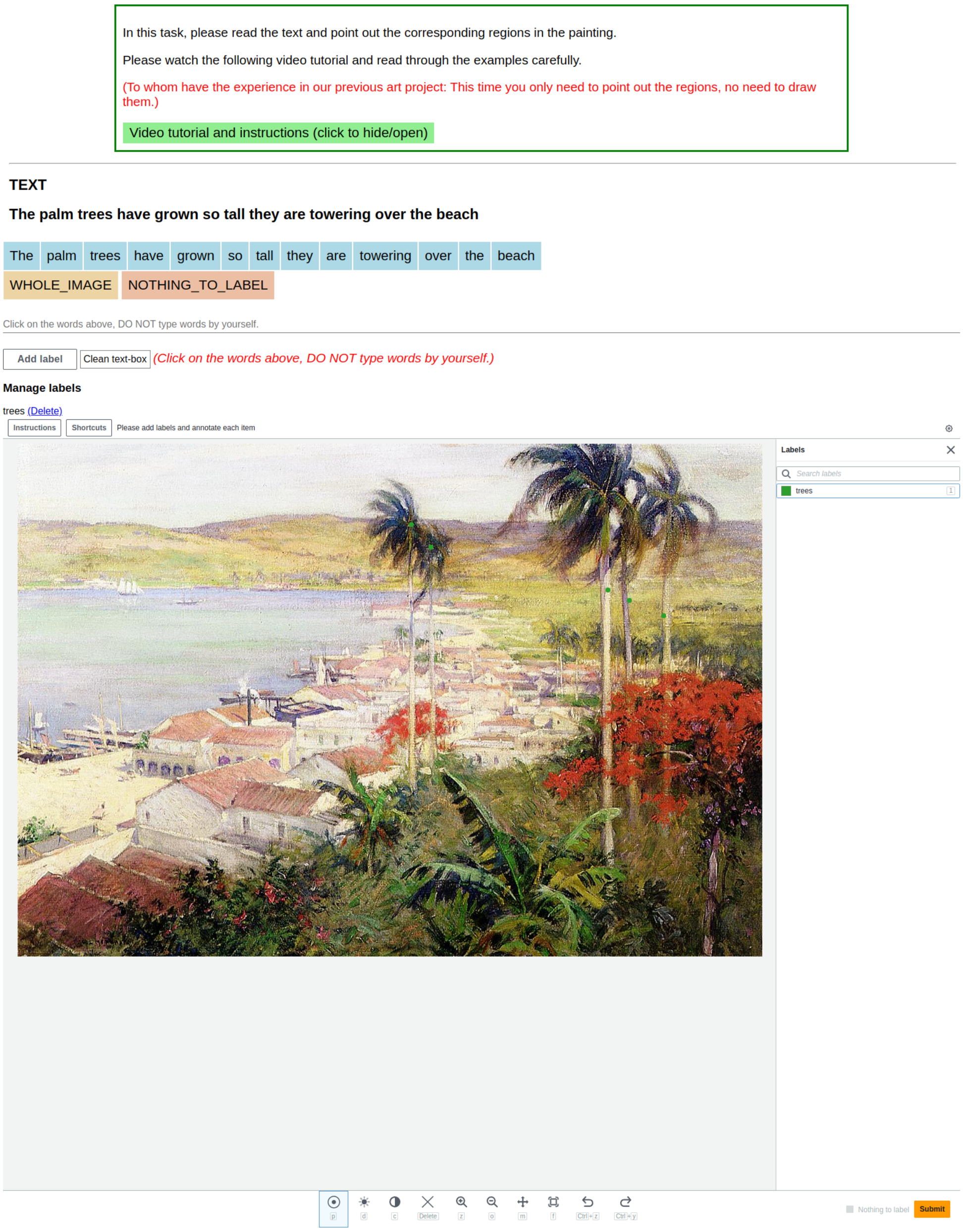

1) Phrase-region selection
2) Region annotation
3) Aggregation
Emotion: Awe
Utterance: The palm trees have grown so tall they are towering over the beach.
palm trees, beach
Emotion: Awe
Utterance: The palm trees blowing in the wind looks peaceful.
palm trees
Emotion: Awe

Figure 3.4: In our annotation process, workers should annotate the following three steps: 1) phrase-region selection, 2) region annotation, and 3) aggregation. We randomly check the submissions at every step to ensure the annotation quality.

3.4.2 Annotation process

As the aim is to annotate emotional stimuli, which are regions that can evoke a certain emotion, we design an annotation process focused on identifying the regions that correspond to specific phrases in the utterances, as these phrases are strongly tied to the emotions. The general annotation process is shown in Figure 3.4, and it consists of three steps: 1) phrase-region selection, 2) region annotation, and 3) aggregation.

Phrase-region selection The annotation interface is shown in Figure 3.5. In the first step, we aim to gather the cues of the emotion elicitation process from the utterances, *i.e.*, to collect phrases in the utterances that correspond to the emotional stimuli and their corresponding



In this task, please read the text and point out the corresponding regions in the painting.

Please watch the following video tutorial and read through the examples carefully.

(To whom have the experience in our previous art project: This time you only need to point out the regions, no need to draw them.)

Video tutorial and instructions (click to hide/open)

TEXT

The palm trees have grown so tall they are towering over the beach

The palm trees have grown so tall they are towering over the beach

WHOLE_IMAGE NOTHING_TO_LABEL

Click on the words above, DO NOT type words by yourself.

Add label | Clean text-box | (Click on the words above, DO NOT type words by yourself.)

Manage labels

trees (Delete)

Figure 3.5: Annotation interface of phrase-region selection. On this page, an annotator should first read the utterance and artwork and then point out the location of the region. The blue, yellow, and orange blocks in the "TEXT" section are the buttons for annotators to select.

location on the artwork.

To identify such phrases, we show annotators a single utterance u together with an artwork

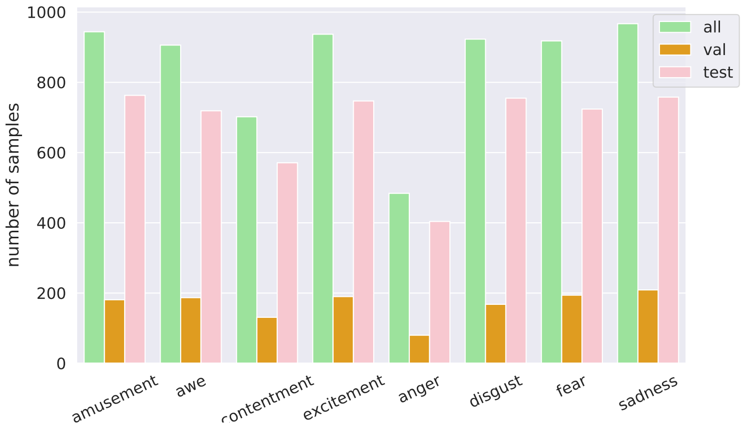


Figure 3.6: Emotion distribution of APOLO.

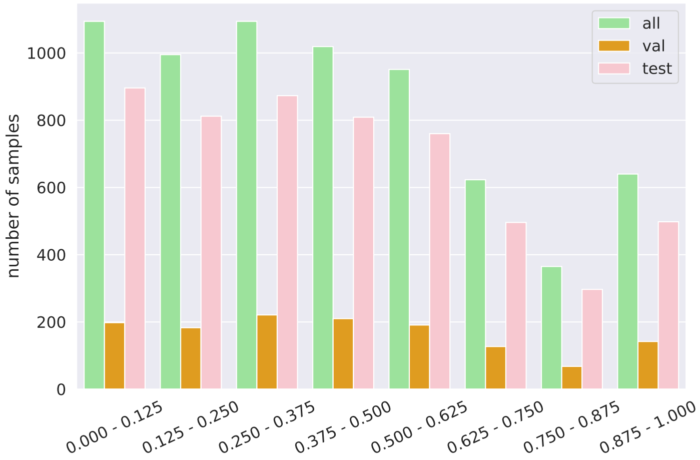


Figure 3.7: Stimuli occupation distribution of APOLO. The x-axis is the ratio of annotated pixels to the whole image, *i.e.*, the occupation of the stimuli.

a. Note that by design, ArtEmis utterances u explicitly describe the emotion generated by the artwork a. Then, we ask them to find all the noun phrases in u that explicitly mention visual concepts in a. We denote the set of identified phrases in u by W_u, where $w \in W_u$ is a phrase (*e.g.*, the "trees"). Specifically, we provide annotators two additional options, the *whole artwork* and the *nothing to label* (as the yellow and orange buttons in Figure 3.5), since some utterances may only talk about the whole image or nothing related to the artwork. If there is at least one phrase in the utterances that corresponds to the emotional stimuli, the annotators are then asked to locate the region in the artwork by spotting at least one point that lies in the region of the visual concepts by clicking on the artwork in our annotation interface. The set of points for phrase w is denoted by P_w, where $p \in P_w$ is in $\mathbb{R}^2$. To ensure that all phrases in the utterance and visual concepts are found, and also to reduce the subjectivity of annotation, we ask three annotators per triplet (a, e, u) and aggregate annotations by removing duplicates to form W_u and P_w for all $w \in W_u$. By this step, we collect two types of annotations: 1) the noun phrases that are related to both the artworks and the evoked emotions (the colored phrases in Figure 3.4), and 2) the locations of the region that the noun phrases (the colored ×'s in Figure 3.4).

Region annotation In this step, we aim to identify the regions related to the emotion elicitation process, *i.e.*, to draw pixel-level annotations according to the utterances. We collect these annotations based on the locations in the previous step. We show a, u, w, and P_w to an annotator and ask them to draw on top of a all pixels that fall into the visual concepts identified by $w \in W_u$ and P_w, obtaining a segment s_w, which is a set of pixels. By this step, we collect pixel-level annotations of the regions for each of the noun phrase (the colored regions in Figure 3.4).

Aggregation Next, we aggregate phrase-wise region annotations s_w belonging to the same a and emotion e. For all w that is associated with a and e, *i.e.*, $w \in W = \{w \in W_{u'} | (a', e', u') \in \mathcal{D}, a = a', e = e'\}$, we obtain the aggregated emotional stimulus s by

$$s = \bigcup_{w \in W} s_w. \tag{3.3}$$

In this step, we finally collect the region annotations for each emotion. Some examples of a, u and s are shown in Figure 3.2. As a result, we obtain $7,512$ emotional stimuli in $5,160$ artworks. The data structure is shown in Table 3.2.

3.4.3 Quality control

We apply quality controls both during and after the annotation process. During the annotation process, we randomly check 10% of the annotations in every round of submission and reject the

Table 3.2: Data structure of our evaluation set

painting	emotion	map_id
a.y.-jackson_indian-home-1927	sadness	000000
aaron-siskind_new-york-24-1988	anger	000001
abdullah-suriosubroto_bamboo-forest	contentment	000002
abdullah-suriosubroto_mountain-view	excitement	000003
abraham-manievich_moscow-iii	excitement	000004
abraham-manievich_moscow-iii	sadness	000005
...	...	...

dishonest ones (*e.g.*, phrase $w \in W$ is wrong, region s_w is wrong, *etc.*). After the annotation process, we manually check all the annotations with special attention to the following three cases: 1) when the *whole artwork* is annotated as a region, 2) when the annotation is *low-quality* (*e.g.*, only draw the contour) or wrong (*e.g.*, draw wrong regions), and 3) when no region (denoted *void*) is annotated in the artwork. We found $1,211$ of *whole artwork*, 33 *low-quality*, and 87 *void* annotations. We remove all of them from our dataset. Finally, to ensure that the dataset is balanced and the *whole artwork* annotations are not over-represented, we randomly remove 600 *whole artwork* annotations to form our APOLO dataset.

3.4.4 Evaluation dataset analysis

APOLO consists of $6,781$ emotional stimuli for $4,718$ artworks. We split it into validation and test sets with approximately 20% and 80% of the samples, respectively. The artworks in the validation and the test sets are disjoint. Figure 3.6 shows the distribution of emotion label e in APOLO. We remark that seven out of eight emotions have more than 500 samples, while the number of *anger* samples is smaller due to the fewer samples in the original ArtEmis dataset. The distributions of the validation and test sets are similar to that of the entire dataset.

We also calculate the distribution of the ratio of pixels in s over a, *i.e.*, $|s|/|a|$, where $|\cdot|$ gives the number of pixels in the region s or artwork a. Figure 3.7 shows the distribution. Many

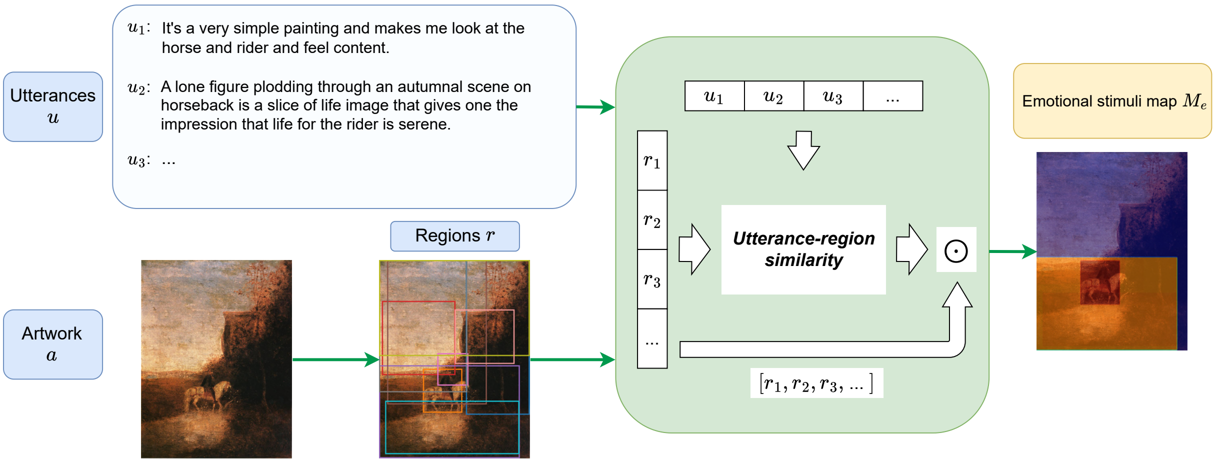

Utterances u
u_1: It's a very simple painting and makes me look at the horse and rider and feel content.
u_2: A lone figure plodding through an autumnal scene on horseback is a slice of life image that gives one the impression that life for the rider is serene.
u_3: ...
Regions r
Artwork a
u_1 u_2 u_3 ...
r_1 r_2 r_3 ...
Utterance-region similarity
$\odot$
$[r_1, r_2, r_3, ...]$
Emotional stimuli map M_e

Figure 3.8: The overview of emotional stimuli map generation for baselines with reference.

regions (46.94%) are small ($|s|/|a| \leq 0.375$), and less regions (24.01%) are large ($|s|/|a| > 0.625$). From this, our evaluation dataset tends to have regions that focus on local concepts. The distributions of $|s|/|a|$ for the validation and test sets are also similar to the entire dataset. Similar to ArtEmis, one artwork may contain a variated number (from one to eight) of emotions, and one artwork-emotion pair may contain a variated number of utterances.

3.5 Baselines

To better comprehend the challenges of the emotional stimuli detection task, we propose and evaluate several baselines.

3.5.1 Baselines with reference

In the with-reference variant, utterance u provides abundant information about what a model should look for, which reduces the task close to visual grounding, like refCOCO [12] and refCOCOg [13]. Our strategy is first to find regions relevant to u with utterance-region similarities and to weight the regions with the similarity to obtain a *emotional stimuli map* with pixel-level scores for e. This process is shown in Figure 3.8. Prediction $\hat{s}$ can be generated by thresholding the map.

We employ **VilBERT** [31] and **12-in-1** [35] as baselines, where 12-in-1 may have a variety of knowledge as it is trained over 12 vision-and-language tasks, while VilBERT is pre-trained on a large-scale dataset GCC [68]. To adapt to our task, VilBERT and 12-in-1 models are fine-tuned with refCOCO. These models give the probability of each region proposal given u, which can be interpreted as an utterance-region similarity score. **CLIP+VinVL** is a combination of CLIP [18] and VinVL [66]. CLIP [18] is renowned for its zero-shot capacity to solve vision-and-language tasks. We can first use VinVL to find region proposals and use CLIP to compute the utterance-region similarity with u.

Emotional stimuli map generation Let R denote the set of regions obtained from any of the above methods, and $sim(r, u)$ be the utterance-region similarity between $r \in R$ and u. We aggregate all regions in R to generate an *emotional stimuli map* M_u for u by

$$M_u = \sum_{r \in R} sim(r, u) m_r, \tag{3.4}$$

where m_r is a map that represents r by giving 1 if a pixel in m_r is in r and 0 otherwise. As an artwork a can be associated with multiple utterances for the same emotion, we aggregate all of them to obtain emotional stimuli map M_e for e as

$$M_e = \sum_{u \in U_{ae}} M_u, \tag{3.5}$$

where $U_{ae} = \{u' | (u', a', e') \in \mathcal{D}, a' = a, e' = e\}$. Thresholding is applied to M_e.

3.5.2 Baselines without reference

Object detection

One naïve idea for the without-reference task is to spot salient regions in some senses and give the regions as emotional stimuli regardless of given emotion e. Object detection can give such regions [127]. We adopt the region proposal networks in **FasterRCNN** [128] and **VinVL** [66]. VinVL's region proposal network may offer better performance as it can additionally detect

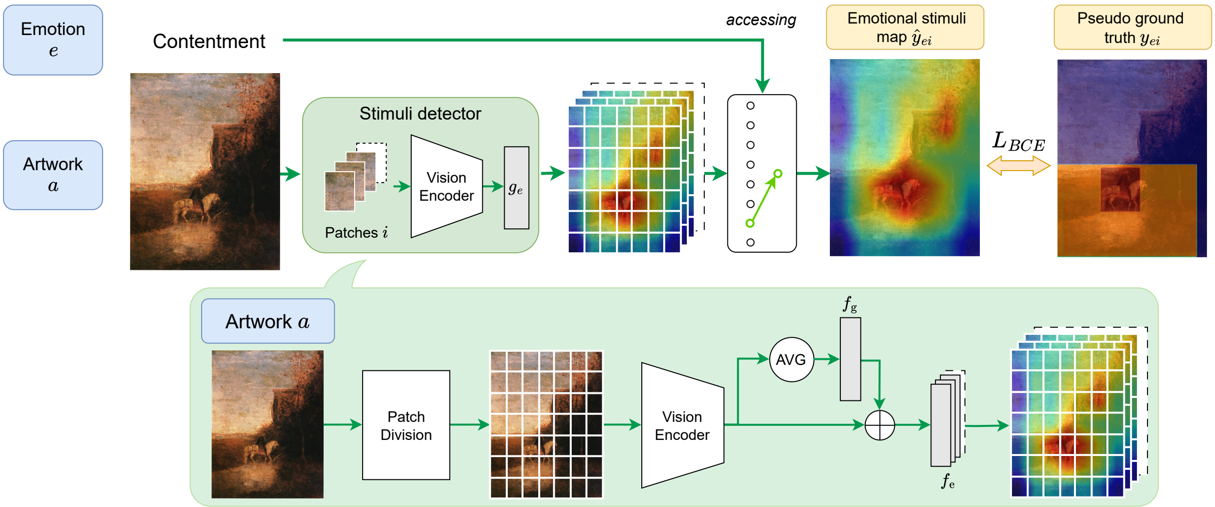


Figure 3.9: The overview of WESD. WESD predicts an emotional stimuli map for each emotion, and we access a certain map when the emotion (*e.g.*, *contentment*) is given. For training, we use pseudo ground truth from CLIP+VinVL since APOLO does not have training data.

some attributes (*e.g.*, *blue* and *calm*) that may exhibit stronger ties with some emotions. We aggregate proposals with top-K confidence to form $\hat{s}$ for any $e \in \mathcal{E}$ (*i.e.*, $f_e(a) = f_{e'}(a)$ even for $e \neq e'$). To obtain segment prediction $\hat{s}$, we follow the same procedure as emotional stimuli map generation in the previous section, but we use $1/|r|$ in place of $sim(r, u)$ as this task does not allow to use u, so we cannot compute $sim(r, u)$.

CASNet and CASNet II

CASNet [122] is a learning-based model for saliency detection, which generates a saliency map for a given image. The model is trained on a dataset called EMOd, which contains images that evoke some emotions and human fixations. With this dataset, CASNet learns to find regions that draw human attention. The work [122] showed, based on their analysis over EMOd, that humans tend to focus on *emotional* objects than *neutral* objects, where *emotional* and *neutral* objects are annotated by annotators. Therefore, CASNet also tends to focus on *emotional* objects. For our task, we apply thresholding to the saliency map to obtain $\hat{s}$. Again, prediction $\hat{s}$ is the same for all e. We also evaluate **CASNet II** [129], an extension of CASNet with atrous spatial pyramid pooling [130].

Weakly-supervised emotional stimuli detecter

As the baselines for the without-reference task so far are not designed for this task and are ignorant of emotion label e, we design a dedicated model, abbreviated as WESD (**W**eakly-supervised **E**motional **S**timuli **D**etection), using utterances in ArtEmis [1] for weakly-supervised training.

An overview of WESD is shown in Figure 3.9. It first uses a visual encoder, such as ResNet variants [131], that gives patch-wise visual features. The visual features of respective patches of artwork a are then fed into a binary classifier for each e to predict if the patch contains emotional stimuli for emotion e. Let v_i be a feature vector for patch $i \in K$, where K is the total number of patches in one artwork. Classifier g_e for emotion e predicts a score as

$$\hat{y}_{ei} = g_e(v_i) \in [0, 1]. \tag{3.6}$$

Specifically, g_e predict $\hat{y}_{ei}$ by the feature of both the certain patch and the whole artwork, as

$$g_e(v_i) = F_e\left(v_\text{i} + F_g\left(\frac{1}{K}\sum_k v_i\right)\right), \tag{3.7}$$

where $F_g(\cdot)$ is a fully-connected layer for embedding the whole artwork and $F_e(\cdot)$ is a fully-connected layer to predict $\hat{y}_{ei}$. WESD contains multiple $F_e(\cdot)$'s and each of $F_e(\cdot)$ is related to a certain emotion e (*e.g.*, *contentment*).

For training, ground-truth emotional stimulus s in APOLO can give direct supervision over y_{ei}; however, APOLO is only for validation and testing. We instead use a pseudo ground truth. We utilize CLIP+VinVL for the with-reference task to generate an emotional stimuli map M_e, which can be derived from the ArtEmis training set. This means that the predictions based on utterances are used to distill the knowledge about the emotional stimuli into f_e for the without-reference (without-utterance) task. Emotion label e is only for identifying the map that has pseudo ground truth.

For this, we first divide M_e into the same patches as v_i's and compute the mean within each patch to obtain a soft label y_{ei}. The binary cross-entropy loss BCE$(\cdot, \cdot)$ is used for training, *i.e.*,

$$L = \text{BCE}(\hat{y}_{ei}, y_{ei}). \tag{3.8}$$

For inference, WESD takes artwork a as the only input. The classifiers for all $e \in \mathcal{E}$ predict the score $\hat{y}_{ei}$, which is then summarized into $\hat{y}_e \in [0, 1]^{B_\text{w} \times B_\text{h}}$, where B_w and B_h are the numbers of patches in the horizontal and vertical axes, respectively. The map $\hat{y}_e$ is then resized to the same size as a to obtain predicted emotional stimuli map $\hat{Y}_e$. Predicted segment $\hat{s}$ can be obtained by thresholding over $\hat{Y}_e$.

3.6 Experiments

Metrics For evaluation, we borrow the ideas from previous works to employ *bounding box* [121] and *segmentation* [122] scenarios, where the former only requires to roughly locate emotional stimuli, while the latter requires their precise shapes. The *bounding box* [121] evaluation focuses on both stimuli and their background (*e.g.*, the emotion of *awe* in Figure 3.1), as it assumes that emotions are evoked not only by the stimuli but also by the background. While the *segmentation* [122] evaluation focuses on the stimuli (*e.g.*, the human and the bears in Figure 3.1) themselves, as it assumes that the stimuli are more important than other regions to evoke the certain emotions. We use both methods as both of them could be related to the emotion elicitation process. For both scenarios, we calculate the precision with intersection over union (IoU) threshold θ (Pr@θ), as in [12, 13, 31, 35, 132, 133]. We evaluate models with Pr@25 and Pr@50.

For baselines that output bounding boxes (*i.e.*, FasterRCNN and VinVL), we collectively treat them as a single region (though they can be disconnected) for evaluation in the segmentation scenario. In contrast, for baselines that give segments by thresholding, we generate a bounding box for each connected component for the bounding box scenario.

Implementation details Our baselines in most cases use the default setting in the original paper. As for CLIP, we use the ResNet-50 variant throughout our experiments. For WESD, we

Table 3.3: Results of the evaluation on eight baseline models as well as a lower-bound baseline (*i.e.*, the *Entire artwork*). From the left, *Task* tells that if a baseline model makes predictions with references (*i.e.*, emotion tags or utterances). *Region proposal* shows if and which region proposal network is used by the certain baseline model. *Text input* shows the text input for generating emotional stimuli maps. *Multiple maps* is ✓ when the model can output multiple emotional stimuli maps for different emotions. For results in both *Bounding box* and *Segmentation*, we bold the best score and take an underline to the second-best result.

	Task	Baseline	Region proposal	Input text	Multiple map	Bounding box		Segmentation	
						Pr@25	Pr@50	Pr@25	Pr@50
1	w/o reference	Entire artwork	-	-	-	82.37	63.81	68.61	37.70
2		FasterRCNN	FasterRCNN	-	-	84.03	66.40	74.43	43.67
3		VinVL	VinVL	-	-	84.43	**67.54**	75.10	43.94
4		CASNet	-	-	-	**84.84**	66.02	**76.40**	44.15
5		CASNet II	-	-	-	**84.84**	63.59	76.24	40.18
6		WESD	-	-	✓	84.30	66.66	75.89	**44.97**
7	w/ reference	VilBERT	FasterRCNN	emotion	✓	82.17	63.08	72.14	39.64
8		12-in-1	FasterRCNN	emotion	✓	72.51	50.71	63.90	31.87
9		CLIP + VinVL	VinVL	emotion	✓	81.97	63.00	71.29	40.05
10		VilBERT	FasterRCNN	utterance	✓	84.10	65.41	75.26	42.18
11		12-in-1	FasterRCNN	utterance	✓	80.52	59.16	72.52	37.99
12		CLIP + VinVL	VinVL	utterance	✓	83.31	64.99	72.58	40.08

resize artworks to 224×224 pixels. Bi-linear interpolation is used to resize $\hat{y}_e$ to $\hat{Y}_e$. We train the model for 20 epochs with batch size 128, learning rate 2×10^{-4}, and decay 0.01. The model is optimized with AdamW [134]. For VilBERT [31] and 12-in-1 [35], we follow the procedure in the respective papers to fine-tune the models on refCOCO [12]. All training processes were done on a Quadro RTX 8000 GPU, which took about 20 hours for WESD.

All baselines require a suitable threshold to obtain segment prediction $\hat{s}$. We use the APOLO validation set to find the best one on it with IoU@50 and apply it for evaluation.

Baseline variants For the baselines with reference, which take utterances as input, we can instead use the emotion label (*e.g.*, *excitement*), so that the models can find regions that can be

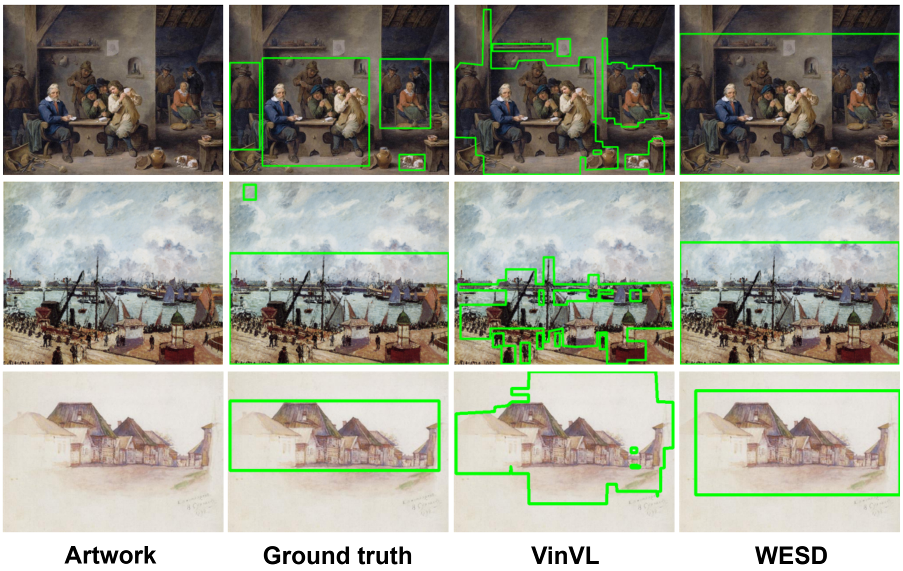


Figure 3.10: Examples of bounding box region detection. VinVL tends to distinguish objects exhaustively from the artwork, while WESD tends to find regions instead of certain objects.

associated with (or that the models learned to associate with) the word.

In addition to the baselines in Section 3.5, we evaluate the case where the entire artwork is predicted as $\hat{s}$.

3.6.1 Quantitative analysis

The scores of all baselines for both with-reference and without-reference tasks are summarized in Table 3.3. We list our findings as follows.

Artworks have something in common with natural images with respect to emotion. For the without-reference task, VinVL, CASNet, and WESD work well. CASNet is the best among these three models in terms of Pr@25 in both bounding box and segmentation scenarios. It also hits the second-best in Pr@50 of segmentation. Although marginal, the superiority of CASNet may imply that EMOd [122] used for training the model in a fully supervised manner has something in common with APOLO. This is intriguing as regions in images that seem to be in very different domains (*i.e.*, natural images and paintings in various artistic styles) share

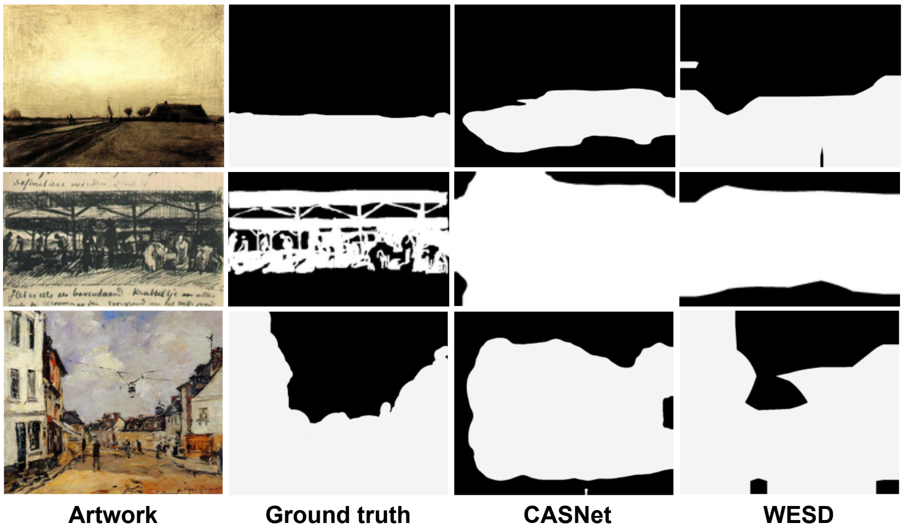


Figure 3.11: Examples of segmentation region detection. Compared to CASNet, WESD tends to find more related regions.

some characteristics. This insight may elicit further exploration of the connection between natural images and paintings, like studying the types of paintings for which a model learned from EMOd works.

Emotional stimuli are highly correlated with objects and attributes. The scores of region proposals by both FasterRCNN and VinVL are still comparable to CASNet and WESD. For the metrics that require precise localization (*i.e.*, Pr@50) and segmentation, the gap seems slightly larger. We would say that emotional stimuli highly coincide with some objects. This is reasonable because the utterances (*e.g.*, in Figure 3.1) mention some objects. A comparison between FasterRCNN and VinVL suggests the correlation between VinVL attributes [66] and emotion, which again makes much sense.

The domain of the utterances may be different from the text in typical vision-and-language tasks. Interestingly, the scores of the with-reference task are mostly lower than those of the without-reference task. This is counterintuitive as the utterances should give beneficial information to identify the emotional stimuli. One possible rationalization is the domain gap. The utterances in ArtEmis [1, 105] come with subjective statements (*e.g.*, "... *bear looking* *forms leaves me uneasy.*" in the second utterance in Figure 3.1), which is not likely in typical vision-and-language tasks. This observation can be supported by the fact that the use of the emotion labels, which is very different from typical text, as input to the vision-and-language models worsens the performance (Lines 7-9 versus Lines 10-12 in Table 3.3). Additionally, the worst scores of 12-in-1 also can also support this because the 12-in-1 model is fine-tuned to 12 vision-and-language tasks and may lose the generalization capability for unseen tasks.

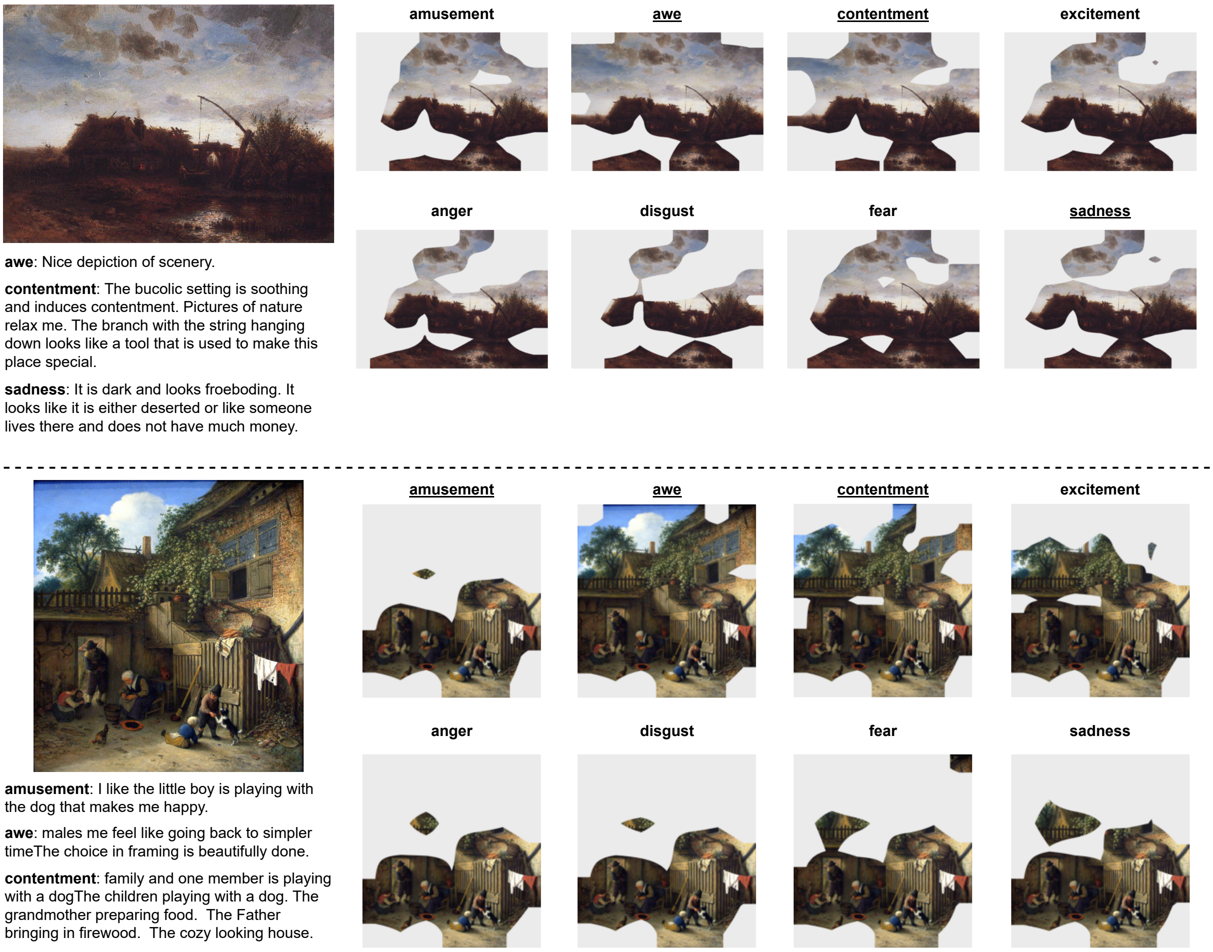
WESD achieved better performance than CLIP+VinVL. Regardless of the worse performance of CLIP+VinVL, WESD hit a higher performance than it, although WESD is trained from CLIP+VinVL. A possible explanation is that, despite the lower performance of CLIP+VinVL for individual artworks, there are some characteristics shared in the dataset, and WESD may capture them through training.

3.6.2 Qualitative analysis

Qualitative examples are shown in Figures 3.10 and 3.11 for the bounding box and segmentation scenarios, respectively. Figure 3.10 shows baselines with the top-2 scores, *i.e.*, VinVL and WESD, where VinVL's bounding boxes are merged when they overlap. Since WESD makes bounding boxes that contain each connected segment of emotional stimuli, it tends to cover a large area. VinVL generates many small bounding boxes around objects, which coincide with the ground-truth bounding boxes.

Figure 3.11 compares WESD against CASNet. We find that CASNet tends to predict regions near the center of the image as emotional stimuli. This tendency is not surprising as the model is supervised by fixations from eye trackers and the image center seems to have a salient component. Meanwhile, WESD tends to find more relevant regions than CASNet, at least for these examples (though because the difference in the scores between WESD and CASNet is small, the trend is not consistent for the APOLO test set.

In general, detecting emotional stimuli in artworks is still challenging as none of the three models perfectly spot the emotional stimuli in both Figure 3.10 and 3.11.



awe: Nice depiction of scenery.

contentment: The bucolic setting is soothing and induces contentment. Pictures of nature relax me. The branch with the string hanging down looks like a tool that is used to make this place special.

sadness: It is dark and looks froeboding. It looks like it is either deserted or like someone lives there and does not have much money.

amusement: I like the little boy is playing with the dog that makes me happy.

awe: males me feel like going back to simpler timeThe choice in framing is beautifully done.

contentment: family and one member is playing with a dogThe children playing with a dog. The grandmother preparing food. The Father bringing in firewood. The cozy looking house.

Figure 3.12: Examples of WESD's prediction on eight different emotions. The texts on the left are the utterances from ArtEmis [1], which are not used during the prediction. The regions on the right are the predicted regions that evoke the corresponding emotions. The emotion tag on the right has an underline if this emotion appears on the left, *i.e.*, has an annotation in ArtEmis.

3.6.3 Emotion-wise analysis on stimuli detector

In this section, we analyze how well WESD predicts emotional stimuli maps for each emotion in one artwork. Specifically, we use the artworks in the test set of ArtEmis for this experiment, and we ask WESD to predict the emotional stimuli map for all of the emotions. We only evaluate WESD as it is the only baseline model that is able to predict emotional stimuli maps for each emotion and does not need utterances as the reference. Some results are shown in Figure 3.12.

Through our experiments, we have the following observations:

- **Predictions focus on similar regions.** Although the WESD's predictions for each emotion are different, most of them focus on similar regions (*e.g.*, the house and the pool in the first example and the people in the second example) in one artwork. The results could be reasonable as some regions in the artwork may play an essential role in evoking multiple emotions. We observe that such regions are also involved in the utterances.

- ***Awe* and *contentment* tend to involve more regions.** Compared with other emotions, *awe* and *contentment* usually involve more regions, such as the whole sky in the first example, and the building and tree in the second example. These results may be related to the factor that the emotions of *awe* and *contentment* are usually evoked by wider sceneries in the artwork.

3.7 Emotional stimuli and deep generative models

We consider emotional stimuli detection a task that may benefit in training future deep generative models in generating emotional artworks (*e.g.*, set as on loss function.) In recent years, deep generative models, such as DALLE-2 [20] and Stable Diffusion [22], have demonstrated remarkable capabilities in producing high-quality images to users' requirements. Such capacities also make these models popular in the artwork field, such as artwork generation [135] and editing [136].

In this section, we explore how much a popular deep generative model, Stable Diffusion [22], can handle emotions when generating artwork, and if our task and models can help improve its performance. To explore this, we randomly select 20 artists with one of his/her artwork in APOLO dataset. Then, we make prompts by "The painting of *[artwork name]* by *[artist name]*, produce *[emotion]*." We use Stable Diffusion v1.5 [22] to generate artworks for all combinations of 20 artworks and eight emotions, resulting in 160 generated artworks. Recently, DAAM [137] found that the aggregation of the cross-attention maps from Stable Dif-

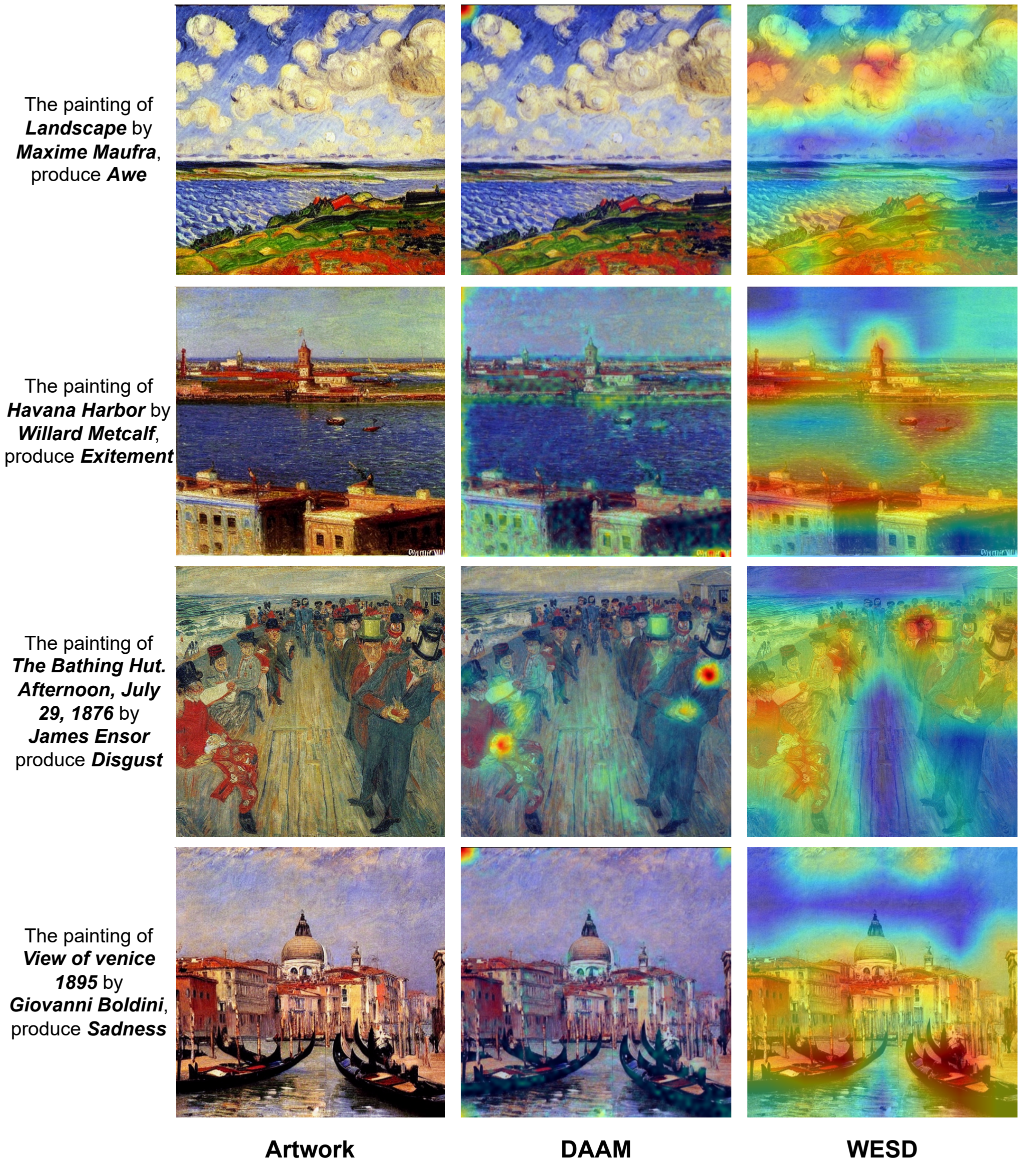

The painting of *Landscape* by *Maxime Maufra*, produce *Awe*
The painting of *Havana Harbor* by *Willard Metcalf*, produce *Exitement*
The painting of *The Bathing Hut. Afternoon, July 29, 1876* by *James Ensor* produce *Disgust*
The painting of *View of venice 1895* by *Giovanni Boldini*, produce *Sadness*
Artwork
DAAM
WESD

Figure 3.13: Examples of Stable Diffusion generated emotional artworks, the internal attention maps (DAAM), and WESD's predictions. The texts on the left are the prompts for Stable Diffusion. WESD could be useful for Stable Diffusion to generate emotional artwork by guiding the model to emphasize the emotional stimuli.

fusion can reveal the interpretation process of the model from prompts to images, *i.e.*, reveal which parts of the image are related to a word in the prompt. We use DAAM to extract the internal attention map of *[emotion]*, which may indicate how Stable Diffusion interprets the emotion to the generated artwork.

We show some results in Figure 3.13. From the generated images, we find that Stable Diffusion can somehow generate artworks that can evoke certain emotions. However, from the internal attention map, we find that attention maps related to *[emotion]* are seldom focused. Instead, we observe that the attention sometimes focuses on the four corners of the artwork. These observations may indicate that it is still hard for Stable Diffusion to handle the relation between emotions and emotional stimuli. Compared to Stable Diffusion, WESD shows more concentration on the regions that are more related to the given emotions. The results may show a potential application of our work and WESD, to work as a guide and benefit Stable Diffusion in focusing on the emotional stimuli and generating more emotional artworks.

3.8 Limitations and ethical concerns

Our task is based on the appraisal theory of artworks and emotions [106, 107]. Although this theory is reliable, it is continuously developing. We tried to remove inconsistent samples when constructing APOLO, as described in Section 3.4, but this may cause some domain gaps between our dataset and general artworks. Additionally, there are rising concerns about the ethical considerations of emotion recognition. As emotions are subjective and personal, trying to predict them with a machine learning model may be intrusive. We agree that emotion prediction could raise privacy issues and potential risks of model abuse. Being aware of this, we did our best to address these concerns proactively. In our experiments, we handled data responsibly and ensured that their use aligned with ethical standards. Additionally, we are planning to inform users of the inherent risks associated with our dataset and ensure they utilize it responsibly. Furthermore, we are prepared to take swift action, including freezing or deleting portions or the entirety of the dataset, if we identify any significant risks associated with its use. Through these measures, we hope to mitigate potential ethical risks and promote responsible usage of our research findings.

3.9 Summary

We introduced an emotional stimuli detection task that targets extracting regions from artworks that evoke emotions. For this task, we build a dedicated dataset, coined APOLO, with $6,781$ emotional stimuli in $4,718$ artworks for evaluation. We also provide with APOLO several baseline models to unveil the challenges in this task. Both qualitative and quantitative evaluations demonstrated that baseline models do not achieve a satisfactory performance, implying inherent difficulties in handling vague and abstract concepts of emotions. Furthermore, we explore how a deep generative model, Stable Diffusion, can handle emotions and emotional stimuli. We find that it is still hard for Stable Diffusion to understand and express emotions. We hope our work can bring inspiration to the fields of artwork analysis and visual emotion analysis.

Chapter 4

Would Deep Generative Models Amplify Bias in Future Models?

4.1 Overview

We investigate the impact of deep generative models on potential social biases in upcoming computer vision models. As the internet witnesses an increasing influx of AI-generated images, concerns arise regarding inherent biases that may accompany them, potentially leading to the dissemination of harmful content. This chapter explores whether a detrimental feedback loop, resulting in bias amplification, would occur if generated images were used as the training data for future models. We conduct simulations by progressively substituting original images in COCO and CC3M datasets with images generated through Stable Diffusion. The modified datasets are used to train OpenCLIP and image captioning models, which we evaluate in terms of quality and bias. Contrary to expectations, our findings indicate that introducing generated images during training does not uniformly amplify bias. Instead, instances of bias mitigation across specific tasks are observed. We further explore the factors that may influence these phenomena, such as artifacts in image generation (*e.g.*, blurry faces) or pre-existing biases in the original datasets.

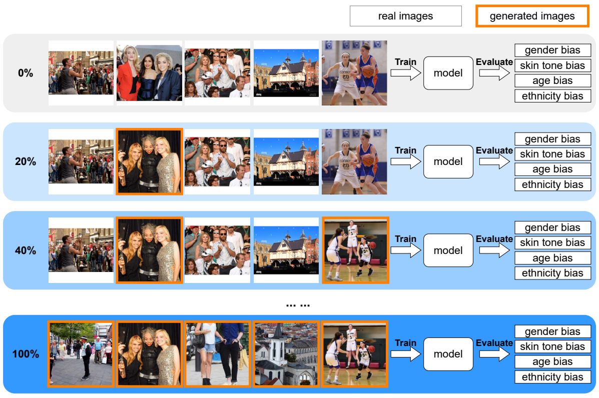


Figure 4.1: We investigate social biases in the training iterations of future models by simulating scenarios where generated images progressively replace real images in the training data.

Emerging deep generative models, such as DALL-E 2 [20], Imagen [21], or Stable Diffusion [22], have shown remarkable capabilities in producing high-quality images. Trained on extensive datasets gathered from the internet [25–27, 68], these models can generate visually compelling images based on user-customized text inputs or prompts, sparking a surge of enthusiasm for image generation across the online community. However, concerns regarding social biases have been systematically identified [138], including gender bias [38–40, 139–145], ethnicity bias [38, 39, 139, 144, 146], and geographical bias [38, 146–148]. In particular, previous work [38, 139, 140, 145] has highlighted the tendency of deep generative models to produce biased images even when prompted with ostensibly neutral inputs, uncovering unfair associations between specific social groups and certain attributes [39, 40, 141, 142]. A common example is the generation of images depicting occupations, such as doctors and nurses, which have been shown to be strongly tied to gender and race.

Issues with bias tend to be attributed to the composition of the training data. Training images are frequently scraped from the internet with minimal efforts to filter out problematic samples and address representational disparities. Moreover, in the current context, generated images are continuously shared online and mixed with real images, which means that future computer vision models may inadvertently incorporate large portions of synthetically generated images into their training processes. Coupled with the increasing concerns about the presence of social bias in deep generative models, this raises the following question: *What consequences might arise if images generated by biased models become increasingly involved in the training process of future models?*

To address this question, we conduct experiments focusing on vision-and-language (VL) tasks within a scenario where generated images are progressively integrated into the training data, as shown in Figure 4.1. Specifically, we generate new images for COCO [94] and CC3M[1] [68] datasets using Stable Diffusion [22], and we gradually replace the original images in the datasets with their generated counterparts. Our evaluation covers four types of demographic bias – gender, ethnicity, age, and skin tone – across two tasks: image-text pre-training and image captioning. For image-text pre-training, we evaluate the bias introduced by OpenCLIP [149] on two downstream tasks, *i.e.*image retrieval [150, 151] and face attribute recognition [152]. For image captioning, we evaluate the performance of ClipCap [77] and Transformer [153] using bias metrics such as leakage (LIC) [154] and gender misprediction (Error) [155, 156].

Our experiments show that the behaviors of the evaluated biases are inconsistent and vary as we gradually replace original images with generated ones. In some cases, biases increase, while in others, they decrease. To understand this phenomenon further, we hypothesize two potential causes: 1) as existing datasets inherently contain biases [150, 151], if the bias introduced by the generated images aligns with the pre-existing biases in the dataset, it may not aggravate the existing bias, and 2) artifacts in Stable Diffusion's generations, particularly concerning the generation of human faces (*e.g.*, blurred or poorly defined attributes), may lead models trained on such data to avoid learning demographic features. Overall, the key contributions of this chapter are:

1. We show that, under our experimental setup, generated images from current deep generative models do not consistently amplify bias. Our experiments reveal different levels of bias for gender, ethnicity, age, and skin tone on both the COCO and CC3M datasets when increasing the number of generated images.

[1] CC3M is also known as Google Conceptual Captions or GCC.

2. Through a set of follow-up experiments, we explore the underlying reasons behind these results, offering valuable insights into the dynamics between image generation models and existing datasets.

3. We propose recommendations for handling biased generated images in the training process of future models, contributing to the ongoing discourse on responsible and unbiased AI development.

While bias is not consistently amplified in our experiments, we find the presence of bias amplification in multiple instances concerning. Moreover, as our experiments are conducted on moderate-scale datasets with about 3 million images, representing about 130 times less data than the original CLIP [18], the impact of generated images on large-scale training remains uncertain. We believe that, as a community, addressing bias and ensuring models are safe for everyone should be a top priority. We hope our findings contribute to increased awareness of fairness in computer vision and inspire the creation of models with unbiased and equitable representations.

4.2 Related work

Bias in pre-trained vision-and-language models Pre-trained VL models are not only used in downstream tasks through fine-tuning [32,35,66] but also in guiding model training [22,157, 158] and serving as evaluation metrics [157, 159, 160]. With the proliferation of VL models, there is an increasing awareness about the inherent biases present in them [150,152,161–163]. For example, Wolfe *et al.* [152] evaluated the proximity of neutral text (*e.g.*, "a photo of a person") and an attributive text (*e.g.*, "a photo of a white person") in the CLIP embedding space [18]. The differences between demographic groups served as indicators of biases in the models. Chuang *et al.* [163] and Garcia *et al.* [150] explored performance gaps among demographic attributes (*e.g.*, `man` and `woman` for gender, and `lighter` and `darker` for skin tone) in downstream tasks, such as classification and image retrieval. Overall, previous work [150, 152, 163, 164] has provided methodologies for detecting and evaluating bias in pre-trained VL models, especially in relation to gender and ethnicity. We leverage these approaches to anticipate potential bias in forthcoming datasets, particularly in scenarios where generated images dominate a significant portion of the online image sources, which is a plausible but underexplored scenario.

Synthetic data and pre-trained models Synthetically generated data is increasingly influencing the pre-training and fine-tuning processes of VL models, whether intentionally or unintentionally. On the one hand, synthetic data is used as an additional training resource when the original dataset is insufficient [135, 165, 166] or unreliable [167]. On the other hand, the widespread dissemination of synthetic images on the internet can inadvertently contaminate datasets [138]. Taori *et al.* [168] explored the data feedback loop and found that incorporating generated data into subsequent model training rounds could exacerbate dataset biases. Furthermore, Hataya *et al.* [169] showed that models trained on large portions of synthetic data dropped their performance. Building upon these insights, we study the repercussions of synthetic data on social bias in VL models.

4.3 Dataset contamination process

VL models are trained on pairs of images and text. The process for collecting this type of data typically begins with scraping the internet to gather a set of images $\mathcal{X} = \{x\}$, where x is an image. For smaller or moderately sized datasets [5, 52, 94], textual descriptions y for each image x are manually generated by crowdsourcing or in-house annotators, resulting in the set $\mathcal{Y} = \{y\}$. However, for large-scale datasets [25–27, 68], where generating specific annotations is unfeasible, text accompanying the images in the original websites is used, often from the ALT[2] text. Subsequently, some form of filtering is applied to remove inappropriate content. Formally, let $p_{\mathcal{I}}(x)$ and $p_{\mathcal{T}}(y)$ represent the distributions of collected images and corresponding descriptions. All $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ can be seen as samples from $p_{\mathcal{I}}(x)$ and $p_{\mathcal{T}}(y)$, respectively.

[2] ALT text refers to the text in the ALT attribute of HTML tags.

The textual description y is derived from $x \sim p_{\mathcal{I}}(x)$ through a framing process $y = f(x)$, which determines what aspects of x to describe.

Biases in the dataset-creation process are introduced from three main sources [170]. Firstly, biases are inherited from the original population of images on the internet,[3] in which content from specific demographic groups and geographical regions is overrepresented. Secondly, additional biases are introduced by the image descriptions provided by annotators or website authors, reflecting their stereotypes. Lastly, the filtering process itself can introduce additional bias; for instance, in the CC3M dataset, entities appearing less than 100 times were filtered out, potentially removing content from minority groups.

We define dataset contamination with generated images (hereafter referred to as *dataset contamination*) as a dataset wherein part of its population is replaced with generated images. That is, someone uploads to the internet images $x' = g(y')$ generated by a generative model g with a prompt y'. In this process, we operate under two assumptions: (1) a mental image $\bar{x}$ that people aim to achieve with a generative model also conforms to the distribution $p_{\mathcal{I}}(x)$, and (2) the image description process from the mental image $\bar{x}$ to a prompt y' has the same framing and bias as f. Given these assumptions, we infer that y' adheres to the distribution $p_{\mathcal{T}}(y)$ as $y' = f(\bar{x})$ and $\bar{x} \sim p_{\mathcal{I}}(x)$. Therefore, the distribution $p_{\mathcal{G}}(x)$ of generated images is given by:

$$p_{\mathcal{G}}(x) = \sum_{y} p_{\mathcal{T}\rightarrow\mathcal{G}}(x|y) p_{\mathcal{T}}(y), \tag{4.1}$$

where $p_{\mathcal{T}\rightarrow\mathcal{G}}(x|y)$ corresponds to the generative process $g(y)$. This means that we can generate images from descriptions $y \in \mathcal{Y}$ as described in [169]. Eventually, we create a dataset $\mathcal{D}(\alpha)$ by sampling images x with a prior α from:

$$\mathcal{D}(\alpha) = \{x \sim (1-\alpha) p_{\mathcal{I}}(x) + \alpha p_{\mathcal{G}}(x)\}. \tag{4.2}$$

This process of dataset contamination allows us to evaluate the impact of the generative model while keeping the other sources of bias consistent with the original dataset.

[3] If the scraping is random sampling, the population is identical to $p_{\mathcal{I}}(x)$, but typically this is not the case because of filtering.

4.4 Bias evaluation tasks

The range of tasks in the scope of VL is extensive and diverse. For a survey, please refer to [50, 171]. In this work, we examine the effects of dataset contamination on two fundamental tasks: image-text pre-training and image captioning. Next, we outline bias evaluation in each of them.

4.4.1 Image-text pretraining

Image-text pertaining involves training a model to learn semantic correspondences between visual appearance and text, such as associating the word “rabbit" with and image of a rabbit. Models like CLIP [18] and its variants [19, 149, 172–174] are trained on large-scale image-text pairs sourced from the internet. CLIP-like models are reported to exhibit social biases, including gender [150, 152, 163, 164, 175], ethnicity [150, 152, 163], age [150, 152], and skin tone [150], and are susceptible to additional biases introduced by dataset contamination. We use OpenCLIP [149], an open-source variant, and assess its performance on text-to-image retrieval, self-similarity, and person preference.

Text-to-image retrieval Following Garcia *et al.* [150], where CLIP was shown to perform differently for different demographic attributes (*e.g.*images of men showed a higher recall at k (R@k) than images of women), we evaluate text-to-image retrieval performance. Text-to-image retrieval consists on finding the corresponding image given an input text. We compute R@k for different demographic attributes on PHASE [150] and COCO [94] datasets for OpenCLIP models trained on datasets $\mathcal{D}(\alpha)$.

Self-similarity Proposed by Wolfe *et al.* [152], *self-similarity* evaluates how images of an attribute group are distributed in the embedding space. The core idea is that if a CLIP-like model is trained on numerous images of a specific group with diverse descriptions in the contrastive training process, its encoders will attempt to distribute these images within a larger volume in the embedding space to differentiate them. Otherwise, images of an underrepresented group may occupy a smaller volume.

Formally, let $\mathcal{E}_a \subset \mathcal{E}$ denote the subset of the entire test set $\mathcal{E}$, containing only samples of a certain attribute group a. Self-similarity $\mathrm{SS}(\mathcal{E}_a)$ for group a is given by:

$$\mathrm{SS}(\mathcal{E}_a) = \frac{1}{|\mathcal{E}_a|^2 - |\mathcal{E}_a|} \sum_{x,x'} c(x, x'), \tag{4.3}$$

where $|\mathcal{E}_a|$ gives the number of samples in $\mathcal{E}_a$, $c(x, x')$ denotes the cosine similarity between x and x' in the embedding space,[4] and the summation is computed over all combinations of two samples x and x' in $\mathcal{E}_a$. A higher self-similarity means images in $\mathcal{E}_a$ are concentrated in the embedding space.

Different treatments of attribute groups appear in the difference of $\mathrm{SS}(\mathcal{E}_a)$'s among a in attribute $\mathcal{A}$.[5] Self-similarity is defined over the learned embedding space, and the samples in that space give different distributions for different datasets; therefore, self-similarity cannot be compared across models. As we are interested in how broad the distribution for $a \in \mathcal{A}$ are in comparison with others in $\mathcal{A}$, we normalize self-similarity scores as:

$$\bar{\mathrm{SS}}(\mathcal{E}_a) = \frac{\mathrm{SS}(\mathcal{E}_a)}{\sum_{a \in \mathcal{A}} \mathrm{SS}(\mathcal{E}_a)/|\mathcal{E}_a|} - 1. \tag{4.4}$$

Person preference Another possible reflection of bias in the embedding space is whether a neutral description of an image represents images of a specific attribute group, *i.e.*, if a certain group is well-represented in a dataset, a neutral description may cover the attribute group. *Person preference* [152] evaluates this skew by comparing the similarities among a neutral description (*e.g.*, "a photo of a person"), a description with a specific attribute group (*e.g.*, "a photo of a white person"), and images of the group. Formally, let t_{N} and t_a denote the neutral description and one attributed by a. The person preference score over $\mathcal{E}_a$ is given by:

$$\mathrm{PP}(\mathcal{E}_a) = \frac{1}{|\mathcal{E}_a|} \sum_{x \in \mathcal{E}_a} \mathbb{1}[c(x, t_{\mathrm{N}}) > c(x, t_a)] \tag{4.5}$$

[4] Letting e_{V} denote the CLIP visual encoder, $c(x, x')$ is defined as $c(x, x') = \cos(e_{\mathrm{V}}(x), e_{\mathrm{V}}(x'))$ where cos gives the cosine similarity.

[5] For instance, the binarized gender attribute in PHASE [150] is given by $\mathcal{A} = \{\texttt{male}, \texttt{female}\}$.

where $\mathbb{1}$ is the indicator function, and we abuse notation c to represent the cosine similarity between an image and a description, embedding them with appropriate encoders.

4.4.2 Image captioning

Image captioning is the task of generating descriptions for an input image. Descriptions generated by image captioning models [32, 77] have been found to reproduce bias, especially concerning gender and skin-tone [151, 154, 156]. We assess image captioning models trained on data contamination in terms on caption quality, LIC, and gender misprediction.

Caption quality Several automatic metrics have been proposed for evaluating captions quality, including BLEU [176], ROUGE [177], METEOR [178], CIDEr [179], and SPICE [180], which mainly involve a lexical comparison between the generated caption and the correspondent ground-truth caption. Alternatively, CLIPScore [159] evaluates the fidelity of a generated caption to the original image. In our experiments, we adopt BLEU-4, CIDEr, SPICE, and CLIPScore.

LIC To evaluate social bias amplification in image captioning models, Hirota *et al.* [154] proposed LIC. This metric evaluates whether the generated captions are more biased than the captions in the original trained dataset. For LIC, a set of captions is assumed to be biased if a protected attribute can be predicted without being explicitly mentioned. Specifically, an attribute classifier $h_a(y)$, which gives the likeliness of an attribute group a from a caption y, is trained on a training set $\mathcal{C}_\text{T} = \{(y, a)\}$, where a is the ground-truth attribute group. All attribute-specific words[6] in the caption y are masked so that the prediction is not trivial. Then, given a validation set $\mathcal{C}_\text{V}$, again with all attribute-specific words being masked, the model's leakage score is computed as:

$$\text{LIC}_\text{M} = \frac{1}{|\mathcal{C}_\text{E}|} \sum_{(y,a)\in\mathcal{C}_\text{E}} h_a(y)\mathbb{1}[\arg\max_{a'} h_{a'}(y) = a] \tag{4.6}$$

[6] We use the same list of attribute-specific words as [154].

$\mathrm{LIC_M}$ gives a higher value if the attribute group is correctly predicted with a higher confidence value even for the masked captions in $\mathcal{C}_\mathrm{E}$, suggesting that the attribute group can be easily predicted from captions.

The leakage score is also computed for the captions in the original dataset, *i.e.*, $\mathrm{LIC_D}$ for $\mathcal{Y}$. The final amplification metric LIC is defined as the difference between the dataset and the model leakage as:

$$\mathrm{LIC} = \mathrm{LIC_M} - \mathrm{LIC_D}. \tag{4.7}$$

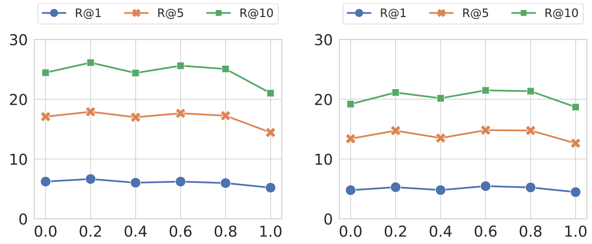
Gender misprediction Another bias evaluation metric for image captioning is the *Gender missprediction* or *Error* [155, 156], which measures gender mispredictions in the generated captions as:

$$\mathrm{Error} = \frac{N}{M}, \tag{4.8}$$

where M is the number of generated captions, and N is the number of captions among the M generated captions whose gender group is incorrectly predicted. Gender is considered incorrectly predicted if it contains any words in the attribute-specific word list for the gender opposite to the ground truth gender. For example, for the ground-truth group `man`, the gender in the generated caption is considered correct if there are no words from the `woman`-specific word list, such as *girl*.

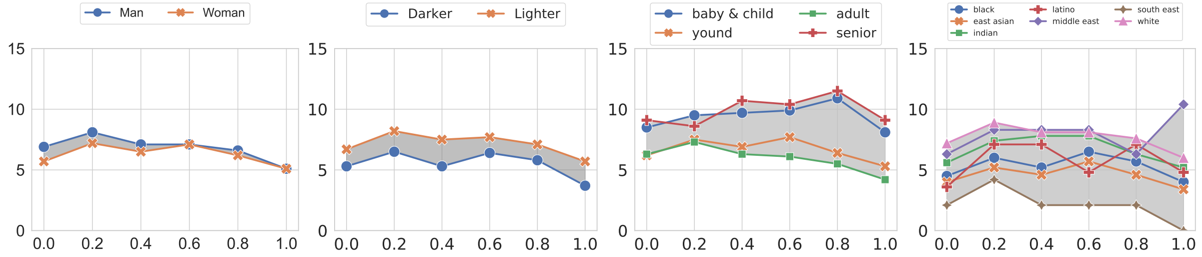
4.5 Results on OpenCLIP

We train OpenCLIP [149] using various versions of the CC3M [68] dataset, each with different levels of dataset contamination. For dataset contamination, we use Stable Diffusion v1.5 [22] to generate images using the original captions as prompts. Due to the nature of the CC3M dataset, where images are provided as URL links and many of these links have expired, we are only able to retrieve $2,772,289$ valid images for our training data. Consequently, we generate images solely for the prompts corresponding to the available images. We randomly replace



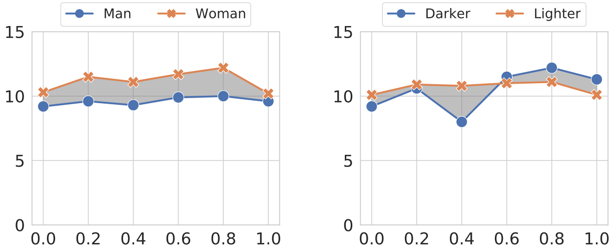
(a) COCO 2014 test set (b) Flickr30K test set

Figure 4.2: Image retrieval results on COCO 2014 test set and Flickr30k test set for different α. The performance of OpenCLIP remains consistent across different levels of dataset contamination.



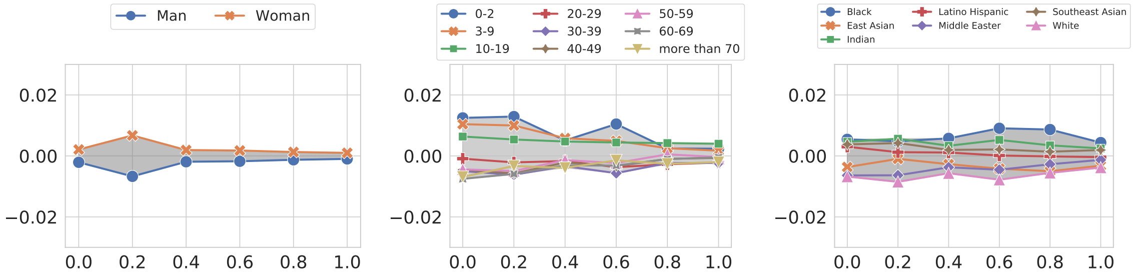
(a) gender (b) skin tone (c) age (d) ethnicity

Figure 4.3: $R@5$ on CC3M using PHASE annotations for different α. Bias is highlighted in gray as the difference between groups. We observe different trends: bias mitigation in Figure 4.3a, consistency in Figure 4.3b, amplification in Figure 4.3c, and no clear trend in Figure 4.3d.



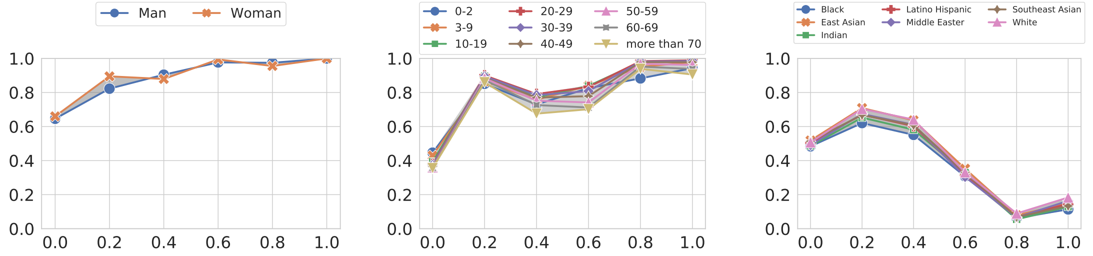
(a) gender (b) skin tone

Figure 4.4: $R@5$ on COCO 2014 test set for different α. Bias is highlighted in gray as the difference between groups. Both gender and skin tone bias show ambiguous trends.



(a) gender (b) age (c) ethnicity

Figure 4.5: Self-similarity score of each group in the FairFace dataset for different α. Bias is highlighted in gray.



(a) gender (b) age (c) ethnicity

Figure 4.6: Person preference score of each group in the FairFace dataset for different α. Bias is highlighted in gray. None of the three figures show a clear tendency. Besides, the changes in bias are relatively small compared with the person preference scores.

20%, 40%, 60%, 80%, and 100% of original images with the images we generate, *i.e.*$\mathcal{D}(\alpha)$ for $\alpha = 0.0$ (the original CC3M dataset), 0.2, 0.4, 0.6, 0.8, and 1. Evaluation is conducted on five datasets, two for performance evaluation and three for bias evaluation. For performance evaluation, we use the COCO 2014 1K test set [94] and the Flickr30k test set [53]. For bias evaluation, we use the CC3M validation set using PHASE demographic annotations [150], the COCO validation set using gender and skin-tone annotations [151], and the whole FairFace dataset [181]. We run all experiments three times with different random seeds and report the average.

4.5.1 OpenCLIP performance

We first evaluate the performance of OpenCLIP trained under our experimental settings on two standard datasets: the COCO 2014 test set and the Flickr30K test set. We report text-to-image retrieval performance as R@k with $k = 1, 5, 10$. Results are shown in Figure 4.2, from which we observe that:

- Image retrieval results remain relatively constant for all levels of dataset contamination, from $\mathcal{D}(0.0)$ to $\mathcal{D}(1.0)$, in both datasets and for R@1, R@5, and R@10.
- Our reported results on OpenCLIP are considerably lower than those of the original CLIP. We attribute this difference to the disparity in the size of the training set. While our training is conducted with less than 3 million image-text pairs, the original CLIP model is trained on about 400 million samples.

In summary, the use of generated images for training OpenCLIP on the CC3M dataset appears to have minimal influence on the retrieval performance of its encoders. Next, we proceed to evaluate the impact of dataset contamination on the bias metrics.

4.5.2 Bias in OpenCLIP

As described in Section 4.4.1, text-to-image pertaining bias is evaluated on three metrics: text-to-image retrieval, self-similarity, and person preference. For text-to-image retrieval, we report results on the CC3M validation set with age, gender, skin-tone and ethnicity annotations from PHASE [150] (Figure 4.3) and the COCO validation set with gender and skin-tone annotations from [151] (Figure 4.4). For self-similarity and person preference, we report results on the FairFace dataset (Figures 4.5 and 4.6). From these results, we find the following trends with respect to bias:

- **Consistent bias amplification**: We observe instances of consistent bias amplification, as illustrated in Figure 4.3c, where the text-to-image performance gap between the different age groups widens with increasing levels of dataset contamination.

- **Consistent bias mitigation**: In Figures 4.3a and 4.5a, we observe instances of consistent bias mitigation, where the gender gap is reduced for both text-to-image performance and self-similarity metrics. The gap in self-similarity for the age attribute is also consistently reduced, as shown in Figure 4.5b, indicating a bias mitigation effect with the increase of the dataset contamination parameter α.

- **Unaffected bias**: In some cases, bias remains unchanged. This is observed in Figure 4.3b, where the gap in text-to-image retrieval performance between lighter and darker-skin tone images remains constant for the different values of α from 0.0 to 1.0.

- **Ambiguous bias trends**: Across most instances, we do not discern a clear bias trend. In Figures 4.3a, 4.3d, 4.4b, 4.5c, 4.6a, and 4.6c, we find no consistent pattern of bias changes, representing half of our experimental results. Unlike *unaffected bias*, the bias in these six experiments fluctuates, showing alternating increases and decreases. For instance, in Figure 4.6a, both the woman and man groups intermittently achieve the highest person preference scores. This suggests that multiple factors contribute to bias changes: some amplify bias, while others mitigate it, making bias changes unstable.

Table 4.1: Captioning performance and bias metrics for ClipCap and Transformer.

	ClipCap							Transformer						
	Bias (↓)			Quality (↑)				Bias (↓)			Quality (↑)			
α	LIC-Gender	LIC-Skin	Error	BLEU-4	CIDEr	SPICE	CLIPScore	LIC-Gender	LIC-Skin	Error	BLEU-4	CIDEr	SPICE	CLIPScore
0	3.6	1.1	5.0	31.9	105.0	20.4	76.4	3.6	2.2	11.0	28.3	92.0	18.2	72.8
0.2	3.8	1.9	4.7	31.8	105.1	20.4	76.8	7.6	1.6	12.1	28.4	92.1	18.0	73.1
0.4	5.1	1.6	4.8	31.5	104.5	20.4	77.0	6.1	0.6	14.6	27.3	88.7	17.7	72.6
0.6	3.9	1.6	4.5	31.4	104.1	20.3	77.2	5.3	2.0	10.7	26.5	88.0	17.4	73.1
0.8	4.1	2.0	4.6	30.7	102.4	20.0	77.4	3.9	1.9	11.1	26.8	87.7	17.3	72.8
1.0	3.5	3.1	4.1	23.8	84.6	17.7	78.3	2.2	2.2	13.2	21.0	70.3	14.9	72.9

It is worth noting that the person preference scores show substantial variations in different experiments, surpassing 0.9 in gender and age (Figures 4.6a and 4.6b), while dropping to 0.2 for ethnicity (Figure 4.6c), despite the unclear trend of bias changes. This observation may be attributed to potential challenges associated with the generation of facial images with Stable Diffusion.

4.6 Results on image captioning

To analyze bias behavior in image captioning models trained with dataset contamination, we consider two models: Transformer[7] [153] and ClipCap [77]. Each model is trained on the COCO 2014 train set [94] with different levels of dataset contamination, ranging from $\mathcal{D}(0.0)$ to $\mathcal{D}(1.0)$. Evaluation is conducted in terms of caption quality and bias on the original COCO validation set using gender and skin-tone annotations from [151].

4.6.1 Image captioning performance

Image captioning results are presented in Table 4.1. Observing the image quality metrics (*i.e.*, BLEU-4, CIDEr, SPICE, and CLIPScore) we note the following:

[7] Transformer refers to a captioning model with a Transformer-based encoder-decoder where the encoder is ViT-B16 [45], and the decoder is BERT-base [182].

- All lexical similarity-based metrics (*i.e.*, BLUE-4, CIDEr, and SPICE) either experience a gradual decrease or remain relatively stable from $\alpha = 0$, the original dataset, to 0.8. However, there's a significant drop between 0.8 and 1.0, suggesting that even a small amount of real images is necessary to maintain captioning performance.

- In contrast, the semantic similarity-based metric (*i.e.*, CLIPScore) remains unaffected by variations in dataset contamination, particularly evident in the case of the Transformer model. While ClipCap slightly improves in CLIPScore, we hypothesize that it is because of the use of CLIP in both image generation and image captioning processes. That is, Stable Diffusion uses CLIP to obtain the text embedding for a caption, so the generated image is strongly tied to it. Therefore, the training set $\mathcal{D}(\alpha)$ with larger α gives image-caption pairs that are close to each other in the CLIP embedding space. ClipCap trained with such a dataset thus only needs to learn the inverse process of the CLIP text encoder, *i.e.*, from an embedding to a caption, for these pairs, which can be easier than learning to fill the gap between images to captions. Thus, ClipCap may easily generate captions that match well with the corresponding images in the CLIP embedding space, consequently increasing CLIPScore.

4.6.2 Bias metrics in image captioning

With regard to the bias metrics, which include LIC for gender (LIC-gender), LIC for skin-tone (LIC-skin), and gender mispredictions (error), the results are also presented in Table 4.1. We summarize our observations as follows:

- **No trend for gender bias**: LIC scores for gender show no noticeable trend across different values of α. In terms of gender mispredictions, similar to the LIC score, there is no clear tendency across the contamination ratios. Under our settings, we cannot draw any definitive conclusion about gender bias.

- **Skin-tone bias amplification**: While LIC for skin-tone on Transformer appears stable, on ClipCap it increases from 1.1 at $\alpha = 0$ to 3.1 at $\alpha = 1$. This trend could be attributed to Stable Diffusion accentuating the skin-tone bias present in the original dataset. For example, it has been found that, in the COCO dataset, indoor images tend to feature white people while black people tend to appear indoors [151]. Similar contextual biases have been observed in Stable Diffusion generations [38, 146].

4.7 Analysis

Through our experiments, we observe the existence of different trends in the biases as we progressively replace real images with generated ones. To comprehend the underlying reasons behind this phenomenon, we explore potential factors based on our observations. We primarily focus on two possible explanations: (1) the inherent biases present within the original training datasets, and (2) the limitations of current deep generative models.

Inherent biases in original datasets Even though Stable Diffusion is known to produce biased images [38–40, 139–141, 146, 147], the original datasets, CC3M and COCO datasets, have also been found to be strongly unbalanced [150, 151]. For example, the CC3M validation set shows large gaps in perceived skin tone, with $3,166$ images of lighter v.s. 318 images of darker skin-tone people, and perceived ethnicity, with $2,231$ images of White people v.s. 16 images of Middle Eastern people [150]. Similarly, the COCO validation set, has been annotated with $7,466$ images of man v.s. $3,314$ images of woman and $9,873$ images of lighter v.s. $1,096$ images of darker skin-tone people [151]. If the disparities in representation within the original datasets resemble the biases in the images generated by Stable Diffusion, it is plausible that the biases remain unchanged as real images are progressively replaced with generated ones.

Failure of generation in Stable Diffusion Deep generative models like Stable Diffusion present several limitations beyond bias concerns. One prominent issue is the tendency for faces to become blurred when generating multiple people. Moreover, Stable Diffusion has been shown to stereotype certain culturally-associated words [147]. When examining the generated

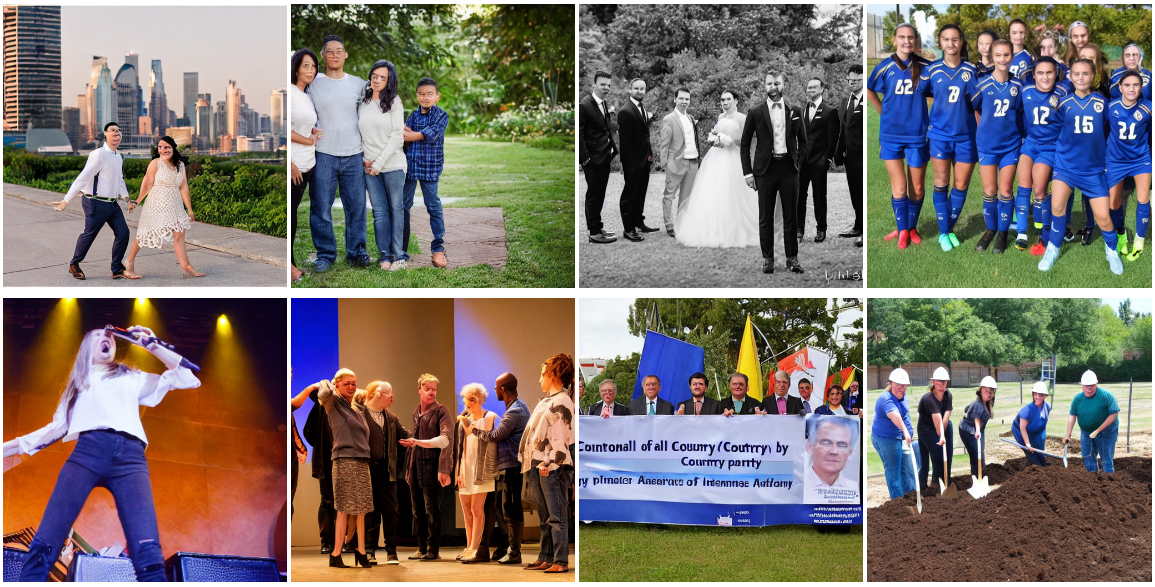


Figure 4.7: Blurry faces in the generated images. When this happens, the attributes (*e.g.*, gender and age) on the faces are hard to distinguish and further used in the model's training.

images in the training dataset, we find similar issues, as shown in Figures 4.7 and 4.8. These issues can impact bias: blurred faces may diminish gender or age biases, while stereotyping could potentially exacerbate ethnicity bias. This phenomenon could elucidate the gender bias mitigation observed in Figures 4.3a and 4.5a. Overall, due to the complexity of how bias originates and propagates across tasks, there is no one-size-fits-all solution to explain its causes and remedies.

4.8 Recommendations

From our experiments and analysis, we found that while images generated by Stable Diffusion exhibit bias across different demographic attributes, their use for training does not consistently amplify bias. This finding aligns with recent studies [19, 167, 183, 184] that use generated data from deep generative models for training. These studies highlight the diversity of effects that the generated data can have on model performance, potentially leading to performance

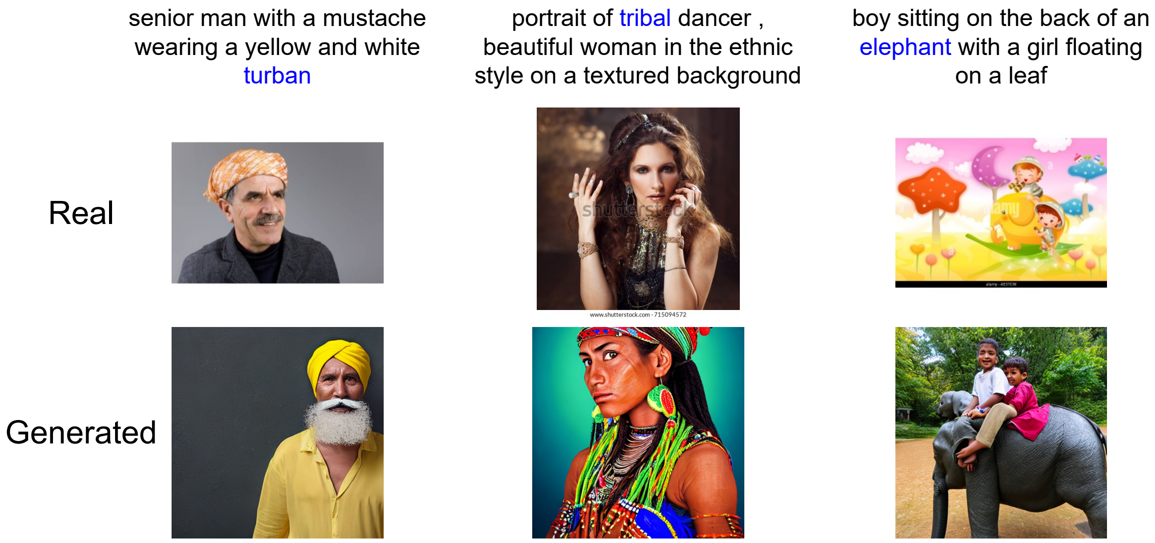

senior man with a mustache wearing a yellow and white turban
portrait of tribal dancer , beautiful woman in the ethnic style on a textured background
boy sitting on the back of an elephant with a girl floating on a leaf
Real
Generated

Figure 4.8: Stereotyping in the generated images. The words in blue may cause Stable Diffusion to generate stereotyped images.

improvements. Since the impact of generated data may depend on the original dataset and target task, we propose the following recommendations:

- **Bias-filtering preprocessing**: Considering the possibility that bias in the original dataset could be more pronounced than in deep generative models, we advocate for bias-filtering preprocessing during data collection from the internet, regardless of whether generated images are involved.

- **Caution with generation issues**: While generation issues like blurry faces may aid in bias mitigation in some tasks, they could potentially lead to bias amplification in others. Moreover, it is important not to regard generation issues as features, as they may be resolved in future iterations of generative models.

4.9 Limitations

- Due to the scale of current vision-and-language datasets like LAION-400M [26] and LAION-5B [27], our computational resources are insufficient for generating images and training models on such large datasets. Instead, our experiments are conducted using COCO and CC3M datasets, limiting the scope of insights to be drawn.

- The use of Stable Diffusion for image generation may overlook potential findings that could arise from other models with either more biased generations or better bias filtering capabilities.
- Our bias evaluation is focused on gender, age, ethnicity, and skin tone. The study does not explore all potential types of bias and leaves out the exploration of intersectional bias, leaving room for further investigation into additional dimensions of bias and fairness.

4.10 Summary

We investigated the impact of synthetic images generated by Stable Diffusion on bias in future models. We simulated a scenario where the generated images are progressively integrated into future datasets and evaluated bias in two downstream tasks: image-text pertaining with OpenCLIP and image captioning. Our findings revealed that the inclusion of generated images resulted in diverse effects on the downstream tasks, ranging from bias amplification to bias mitigation. Further visualization and analysis provided potential explanations underlying this phenomenon, including the inherent bias in the original datasets and the generation issues associated with Stable Diffusion.

Chapter 5

Conclusion

In this thesis, we explore the knowledge transferability in vision-and-language models and its applications, aiming to explore the limitations of the current knowledge transfer strategy, analyze the reason, and further improve the models' performance in solving vision-and-language tasks.

Through the experiments, we find that the current knowledge transfer strategy still has limitations in utilizing existing knowledge, as some knowledge transfer may not be helpful in solving certain tasks. Several factors, such as task similarity, training data scale, and training epochs, may affect the result. Besides, we find that the large-scale pre-trained models such as CLIP [18] still hard to solve the emotional stimuli detection task, which may indicate the knowledge edge of these models. Our experiments on evaluating Stable Diffusion [22]'s capacity in understand emotional stimuli also indicate that this model still has insufficient knowledge in understanding how the emotions are related to the regions of the artworks. Furthermore, we explore how harmful knowledge, such as social bias, in recent deep generative models can affect future vision-and-language models. Our results show the existence of social bias in both deep generative models and future vision-and-language models. However, we also find that the bias in the future vision-and-language models is not always amplified by the deep generative models. The results may show a possible way of utilizing the knowledge in deep generative models: if we can limit the transfer of biased knowledge from deep generative models, these models could be helpful in improving other tasks with their knowledge. We hope our work and our insights through the experiments can bring inspiration to the fields of knowledge transferability in vision-and-language tasks.

Acknowledgements

This work was completed under the supervision of Professor Hajime Nagahara, Professor Yuta Nakashima, and Associate Professor Noa Garcia of the Graduate School of Information Science and Technology, Osaka University.

The author would like to express his deepest appreciation to Dr. Mayu Otani and Associate Professor Chenhui Chu for their guidance and help. He is particularly grateful for the valuable suggestions and ideas from Associate Professor Noa Garcia during the research. Without her encouragement and assistance during his Ph.D., the dissertation would have been impossible.

The author would also like to thank the committee of the fellowship office at Osaka University for the funding and financial support of the 分野横断イノベーションを創造する情報人材育成フェローシップ , which helps the author to concentrate on his research topic.

Finally, the author would like to thank his family for their support and encouragement during his doctoral studies.

Reference

[1] Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J. Guibas. Artemis: Affective language for visual art. In *CVPR*, pp. 11564–11574, 2021.

[2] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Vqa: Visual question answering. *IJCV*, Vol. 123, pp. 4–31, 2015.

[3] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *CVPR*, pp. 6325–6334, 2017.

[4] Noa Garcia, Chentao Ye, Zihua Liu, Qingtao Hu, Mayu Otani, Chenhui Chu, Yuta Nakashima, and Teruko Mitamura. A dataset and baselines for visual question answering on art. In *ECCV Workshops*, Vol. 12536, pp. 92–108, 2020.

[5] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, Vol. 123, pp. 32–73, 2016.

[6] D. A. Hudson and C. D. Manning. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *CVPR*, pp. 6693–6702, 2019.

[7] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *ACL*, pp. 217–223, 2017.

[8] A. Suhr, S. Zhou, I. Zhang, H. Bai, and Y. Artzi. A Corpus for Reasoning about Natural Language Grounded in Photographs. *arXiv*, Vol. abs/1811.00491, , 2019.

[9] N. Xie, F. Lai, D. Doran, and A. Kadav. Visual Entailment Task for Visually-Grounded Language Learning. *arXiv*, Vol. abs/1811.10582, , 2018.

[10] Y. Zhu, O. Groth, M. S. Bernstein, and L. Fei-Fei. Visual7W: Grounded Question Answering in Images. In *CVPR*, pp. 4995–5004, 2016.

[11] H. d. Vries, F. Strub, A. P. S. Chandar, O. Pietquin, H. Larochelle, and A. C. Courville. GuessWhat?! Visual Object Discovery through Multi-modal Dialogue. In *CVPR*, pp. 4466–4475, 2017.

[12] S. Kazemzadeh, V. Ordonez, M. André Matten, and T. L. Berg. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *EMNLP*, 2014.

[13] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. P. Murphy. Generation and Comprehension of Unambiguous Object Descriptions. In *CVPR*, pp. 11–20, 2016.

[14] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, pp. 3608–3617, 2018.

[15] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. In *ECCV*, Vol. 12362, pp. 417–434, 2020.

[16] Zhihui Li, Lina Yao, Xiaoqin Zhang, Xianzhi Wang, Salil S. Kanhere, and Huaxiang Zhang. Zero-shot object detection with textual descriptions. In *AAAI*, pp. 8690–8697, 2019.

[17] Ruoyue Shen, Nakamasa Inoue, and Koichi Shinoda. Text-guided object detector for multi-modal video question answering. In *WACV*, pp. 1032–1042, 2023.

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[19] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. In *NeurIPS*, 2023.

[20] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, Vol. abs/2204.06125, , 2022.

[21] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022.

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

[23] Zhen Zhu, Zhiliang Xu, Ansheng You, and Xiang Bai. Semantically multi-modal image synthesis. In *CVPR*, pp. 5466–5475, 2020.

[24] Alberto Baldrati, Davide Morelli, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Multimodal garment designer: Human-centric latent diffusion models for fashion image editing. In *ICCV*, pp. 23336–23345, 2023.

[25] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021.

[26] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *ArXiv*, Vol. abs/2111.02114, , 2021.

[27] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa 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 *NeurIPS*, 2022.

[28] S. Ravi and H. Larochelle. Optimization as a Model for Few-Shot Learning. In *ICLR*, 2017.

[29] J. Snell, K. Swersky, and R. S. Zemel. Prototypical Networks for Few-shot Learning. In *NeurIPS*, 2017.

[30] D. Teney and A. van den Hengel. Visual Question Answering as a Meta Learning Task. In *ECCV*, 2018.

[31] J. Lu, D. Batra, D. Parikh, and S. Lee. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NeurIPS*, 2019.

[32] X. Li, X. Yin, C. Li, X. Hu, P. Zhang, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, and J. Gao. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *ECCV*, pp. 121–137. Springer, 2020.

[33] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 5583–5594, 2021.

[34] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

[35] J. Lu, V. Goswami, M. Rohrbach, D. Parikh, and S. Lee. 12-in-1: Multi-Task Vision and Language Representation Learning. In *CVPR*, pp. 10434–10443, 2020.

[36] A. R. Zamir, A. Sax, B. (William) Shen, L. J. Guibas, J. Malik, and S. Savarese. Taskonomy: Disentangling Task Transfer Learning. In *CVPR*, pp. 3712–3722, 2018.

[37] T. Mensink, J. R. R. Uijlings, A. Kuznetsova, M. Gygli, and V. Ferrari. Factors of Influence for Transfer Learning across Diverse Appearance Domains and Task Types. *TPAMI*, 2021.

[38] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Y. Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. 2023.

[39] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models. *ArXiv*, Vol. abs/2303.11408, , 2023.

[40] Yanzhe Zhang, Lucy Jiang, Greg Turk, and Diyi Yang. Auditing gender presentation differences in text-to-image models. *ArXiv*, Vol. abs/2302.03675, , 2023.

[41] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. J.

Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language Models are Few-Shot Learners. In *NeurIPS*, Vol. 33, pp. 1877–1901, 2020.

[42] S. Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barth-Maron, M. Gimenez, Y. Sulsky, J. Kay, J. T. Springenberg, T. Eccles, J. Bruce, A. Razavi, A. D. Edwards, N. M. O. Heess, Y. Chen, R. Hadsell, O. Vinyals, M. Bordbar, and N. d. Freitas. A Generalist Agent. *arXiv preprint arXiv:2205.06175*, 2022.

[43] J. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan. Flamingo: a Visual Language Model for Few-Shot Learning. *arXiv preprint arXiv:2204.14198*, 2022.

[44] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125*, 2022.

[45] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[46] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *FAccT*, pp. 610–623, 2021.

[47] V. U. Prabhu and A. Birhane. Large image datasets: A pyrrhic win for computer vision? In *WACV*, pp. 1536–1546, 2021.

[48] X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv*, Vol. abs/1504.00325, , 2015.

[49] F. Ferraro, N. Mostafazadeh, T. 'K.' Huang, L. Vanderwende, J. Devlin, Mi. Galley, and M. Mitchell. A Survey of Current Datasets for Vision and Language Research. In *EMNLP*, 2015.

[50] A. Mogadala, M. Kalimuthu, and D. Klakow. Trends in Integration of Vision and Language Research: A Survey of Tasks, Datasets, and Methods. *JAIR*, Vol. 71, pp. 1183–1317, 2021.

[51] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. S. Ni, D. N. Poland, D. Borth, and L. Li. YFCC100M: the new data in multimedia research. *Commun. ACM*, Vol. 59, pp. 64–73, 2016.

[52] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. *IJCV*, Vol. 123, pp. 74–93, 2015.

[53] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV*, Vol. 123, pp. 74–93, 2015.

[54] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby. Big Transfer (BiT): General Visual Representation Learning. In *ECCV*, 2020.

[55] G. Strezoski, N. v. Noord, and M. Worring. Many Task Learning With Task Routing. In *ICCV*, pp. 1375–1384, 2019.

[56] T. S. Standley, A. R. Zamir, D. Chen, L. J. Guibas, J. Malik, and S. Savarese. Which Tasks Should Be Learned Together in Multi-task Learning? In *ICML*, 2020.

[57] C. Finn, P. Abbeel, and S. Levine. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *ICML*, 2017.

[58] J. Kirkpatrick, R. Pascanu, N. C. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell. Overcoming catastrophic forgetting in neural networks. *PNAS*, Vol. 114, pp. 3521 – 3526, 2017.

[59] B. Thompson, J. Gwinnup, H. Khayrallah, K. Duh, and P. Koehn. Overcoming Catastrophic Forgetting During Domain Adaptation of Neural Machine Translation. In *NAACL*, 2019.

[60] H. H. Tan and M. Bansal. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. *arXiv*, Vol. abs/1908.07490, , 2019.

[61] L. H. Li, M. Yatskar, D. Yin, C. Hsieh, and K. Chang. VisualBERT: A Simple and Performant Baseline for Vision and Language. *arXiv*, Vol. abs/1908.03557, , 2019.

[62] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. *arXiv*, Vol. abs/1908.08530, , 2020.

[63] G. Li, N. Duan, Y. Fang, D. Jiang, and M. Zhou. Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training. In *AAAI*, 2020.

[64] Y. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu. UNITER: UNiversal Image-TExt Representation Learning. In *ECCV*, 2020.

[65] F. Yu, J. Tang, W. Yin, Y. Sun, H. Tian, H. Wu, and H. Wang. ERNIE-ViL: Knowledge Enhanced Vision-Language Representations Through Scene Graph. In *AAAI*, 2021.

[66] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*, pp. 5575–5584, 2021.

[67] Y. Sung, J. Cho, and M. Bansal. VL-Adapter: Parameter-Efficient Transfer Learning for Vision-and-Language Tasks. In *CVPR*, 2022.

[68] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pp. 2556–2565, 2018.

[69] T. Wu, N. Garcia, M. Otani, C. Chu, Y. Nakashima, and H. Takemura. Transferring Domain-Agnostic Knowledge in Video Question Answering. In *BMVC*, 2021.

[70] Haoyu Song, Li Dong, Weinan Zhang, Ting Liu, and Furu Wei. CLIP models are few-shot learners: Empirical studies on VQA and visual entailment. In *ACL*, pp. 6088–6100, 2022.

[71] Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In *NeurIPS*, 2021.

[72] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? In *ICLR*, 2021.

[73] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, pp. 9694–9705, 2021.

[74] S. Pramanik, P. Agrawal, and A. Hussain. OmniNet: A unified architecture for multi-modal multi-task learning. *arXiv*, Vol. abs/1907.07804, , 2019.

[75] D. Nguyen and T. Okatani. Multi-Task Learning of Hierarchical Vision-Language Representation. In *CVPR*, pp. 10484–10493, 2019.

[76] R. Hu and A. Singh. UniT: Multimodal Multitask Learning with a Unified Transformer. In *ICCV*, pp. 1419–1429, 2021.

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

[78] Yanyuan Qiao, Qi Chen, Chaorui Deng, Ning Ding, Yuankai Qi, Mingkui Tan, Xincheng Ren, and Qi Wu. R-gan: Exploring human-like way for reasonable text-to-image synthesis via generative adversarial networks. In *ACM MM*, 2021.

[79] Yanyuan Qiao, Yuankai Qi, Yicong Hong, Zheng Yu, Peifeng Wang, and Qi Wu. Hop+: History-enhanced and order-aware pre-training for vision-and-language navigation. *TPAMI*, Vol. 45, pp. 8524–8537, 2023.

[80] Wanrong Zhu, Yuankai Qi, P. Narayana, Kazoo Sone, Sugato Basu, Xin Eric Wang, Qi Wu, Miguel P. Eckstein, and William Yang Wang. Diagnosing vision-and-language navigation: What really matters. In *NAACL*, pp. 5981–5993, 2022.

[81] Han-Jia Ye, Guorong Li, Yuankai Qi, Shuhui Wang, Qingming Huang, and Ming-Hsuan Yang. Hierarchical modular network for video captioning. In *CVPR*, pp. 17918–17927, 2022.

[82] Weidong Chen, Dexiang Hong, Yuankai Qi, Zhenjun Han, Shuhui Wang, Laiyun Qing, Qingming Huang, and Guorong Li. Multi-attention network for compressed video referring object segmentation. In *ACM MM*, 2022.

[83] Alex Wang, Yixiao Ge, Rui Yan, Yuying Ge, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. All in one: Exploring unified video-language pre-training. In *CVPR*, pp. 6598–6608, 2023.

[84] Qi Chen, Yuanqing Li, Yuankai Qi, Jiaqiu Zhou, Mingkui Tan, and Qi Wu. V2c: Visual voice cloning. In *CVPR*, pp. 21210–21219, 2022.

[85] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NeurIPS*, pp. 1682–1690, 2014.

[86] Mengye Ren, Ryan Kiros, and Richard S. Zemel. Exploring models and data for image question answering. In *NeurIPS*, pp. 2953–2961, 2015.

[87] Noa Garcia and George Vogiatzis. How to read paintings: Semantic art understanding with multi-modal retrieval. In *ECCV Workshops*, Vol. 11130, pp. 676–691, 2018.

[88] Mustafa Shukor, Nicolas Thome, and Matthieu Cord. Vision and structured-language pretraining for cross-modal food retrieval. 2022.

[89] Terry Winograd. Understanding natural language. *Cognitive Psychology*, Vol. 3, No. 1, pp. 1–191, 1972.

[90] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, pp. 1988–1997, 2017.

[91] D. Picard. Torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision. *arXiv*, Vol. abs/2109.08203, , 2021.

[92] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, Vol. 115, pp. 211–252, 2015.

[93] Waldemar W. Koczkodaj, Tamar Kakiashvili, A. Szymanska, J. Montero-Marin, Ricardo Araya, Javier García-Campayo, K. Rutkowski, and Dominik Strzalka. How to reduce the number of rating scale items without predictability loss? *Scientometrics*, Vol. 111, No. 2, pp. 581–593, 2017.

[94] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[95] Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Y. Ma. Investigating the catastrophic forgetting in multimodal large language models. *ArXiv*, Vol. abs/2309.10313, , 2023.

[96] Yong Lin, Lu Tan, Hangyu Lin, Zeming Zheng, Renjie Pi, Jipeng Zhang, Shizhe Diao, Haoxiang Wang, Han Zhao, Yuan Yao, and T. Zhang. Speciality vs generality: An empirical study on catastrophic forgetting in fine-tuning foundation models. *ArXiv*, Vol. abs/2309.06256, , 2023.

[97] Michael J. Wilber, Chen Fang, Hailin Jin, Aaron Hertzmann, John P. Collomosse, and Serge J. Belongie. Bam! the behance artistic media dataset for recognition beyond photography. In *ICCV*, pp. 1211–1220, 2017.

[98] Zechen Bai, Yuta Nakashima, and Noa García. Explain me the painting: Multi-topic knowledgeable art description generation. In *ICCV*, pp. 5402–5412, 2021.

[99] Elliot J. Crowley and Andrew Zisserman. The state of the art: Object retrieval in paintings using discriminative regions. In *BMVC*, 2014.

[100] Nicolas Gonthier, Yann Gousseau, Saïd Ladjal, and Olivier Bonfait. Weakly supervised object detection in artworks. In *ECCV Workshops*, pp. 692–709, 2018.

[101] Thomas Mensink and Jan C. van Gemert. The rijksmuseum challenge: Museum-centered visual recognition. In *ICMR*, p. 451, 2014.

[102] Gjorgji Strezoski and Marcel Worring. Omniart: A large-scale artistic benchmark. *ACM Trans. Multim. Comput. Commun. Appl.*, Vol. 14, No. 4, pp. 88:1–88:21, 2018.

[103] Vincent Tonkes and Matthia Sabatelli. How well do vision transformers (vts) transfer to the non-natural image domain? an empirical study involving art classification. In *ECCV Workshop*, Vol. 13801, pp. 234–250, 2022.

[104] Artem Reshetnikov, Maria-Cristina V. Marinescu, and Joaquim Moré López. Deart: Dataset of european art. In *ECCV Workshop*, Vol. 13801, pp. 218–233, 2022.

[105] Youssef Mohamed, Faizan Farooq Khan, Kilichbek Haydarov, and Mohamed Elhoseiny. It is okay to not be okay: Overcoming emotional bias in affective image captioning by contrastive data collection. In *CVPR*, pp. 21231–21240, 2022.

[106] Paul J. Silvia. Emotional responses to art: From collation and arousal to cognition and emotion. *Review of General Psychology*, Vol. 9, pp. 342 – 357, 2005.

[107] Jessica M. Cooper and Paul J. Silvia. Opposing art: Rejection as an action tendency of hostile aesthetic emotions. *Empirical Studies of the Arts*, Vol. 27, pp. 109 – 126, 2009.

[108] Ioannis Xenakis, Argyris Arnellos, and John Darzentas. The functional role of emotions in aesthetic judgment. *New Ideas in Psychology*, Vol. 30, pp. 212–226, 2012.

[109] Matthew Pelowski and Fuminori Akiba. A model of art perception, evaluation and emotion in transformative aesthetic experience. *New Ideas in Psychology*, Vol. 29, pp. 80–97, 2011.

[110] Jingyuan Yang, Jie Li, Xiumei Wang, Yuxuan Ding, and Xinbo Gao. Stimuli-aware visual emotion analysis. *TIP*, Vol. 30, pp. 7432–7445, 2021.

[111] Jingyuan Yang, Xinbo Gao, Leida Li, Xiumei Wang, and Jinshan Ding. Solver: Scene-object interrelated visual emotion reasoning network. *TIP*, Vol. 30, pp. 8686–8701, 2021.

[112] Jufeng Yang, Dongyu She, Yu-Kun Lai, Paul L. Rosin, and Ming-Hsuan Yang. Weakly supervised coupled networks for visual sentiment analysis. In *CVPR*, pp. 7584–7592, 2018.

[113] Liwen Xu, Z. Wang, Bingwen Wu, and Simon S. Y. Lui. Mdan: Multi-level dependent attention network for visual emotion analysis. In *CVPR*, pp. 9469–9478, 2022.

[114] Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Emoset: A large-scale visual emotion dataset with rich attributes. In *ICCV*, pp. 20326–20337, 2023.

[115] XI Shen, Alexei A. Efros, and Mathieu Aubry. Discovering visual patterns in art collections with spatially-consistent feature learning. In *CVPR*, pp. 9270–9279, 2019.

[116] Huijie Lin, Jia Jia, Quan Guo, Yuanyuan Xue, Jie Huang, Lianhong Cai, and Ling Feng. Psychological stress detection from cross-media microblog data using deep sparse neural network. In *ICME*, pp. 1–6, 2014.

[117] Xin Wang, Huijun Zhang, Lei Cao, and Ling Feng. Leverage social media for personalized stress detection. In *ACM MM*, 2020.

[118] Quoc-Tuan Truong and Hady W. Lauw. Visual sentiment analysis for review images with item-oriented and user-oriented cnn. In *ACM MM*, 2017.

[119] Quoc-Tuan Truong and Hady W. Lauw. Vistanet: Visual aspect attention network for multimodal sentiment analysis. In *AAAI*, 2019.

[120] Sicheng Zhao, Xingxu Yao, Jufeng Yang, Guoli Jia, Guiguang Ding, Tat-Seng Chua, Björn W. Schuller, and Kurt Keutzer. Affective image content analysis: Two decades review and new perspectives. *TPAMI*, Vol. 44, pp. 6729–6751, 2022.

[121] Kuan-Chuan Peng, Amir Sadovnik, Andrew C. Gallagher, and Tsuhan Chen. Where do emotions come from? predicting the emotion stimuli map. In *ICIP*, pp. 614–618, 2016.

[122] Shaojing Fan, Zhiqi Shen, Ming Jiang, Bryan L. Koenig, Juan Xu, M. Kankanhalli, and Qi Zhao. Emotional attention: A study of image sentiment and visual attention. In *CVPR*, pp. 7521–7531, 2018.

[123] Gaowen Liu, Yan Yan, Elisa Ricci, Yi Yang, Yahong Han, Stefan Winkler, and N. Sebe. Inferring painting style with multi-task dictionary learning. In *IJCAI*, 2015.

[124] Nikolaos-Antonios Ypsilantis, Noa Garcia, Guangxing Han, Sarah Ibrahimi, Nanne Van Noord, and Giorgos Tolias. The met dataset: Instance-level recognition for artworks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[125] Dan Ruta, Andrew Gilbert, Pranav Aggarwal, Naveen Marri, Ajinkya Kale, John Briggs, Chris Speed, Hailin Jin, Baldo Faieta, Alex Filipkowski, Zhe Lin, and John P. Collomosse. Stylebabel: Artistic style tagging and captioning. In *ECCV*, 2022.

[126] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM MM*, 2010.

[127] 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 *CVPR*, pp. 6077–6086, 2018.

[128] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *TPAMI*, Vol. 39, pp. 1137–1149, 2015.

[129] Shaojing Fan, Zhiqi Shen, Ming Jiang, Bryan L. Koenig, Mohan S. Kankanhalli, and Qi Zhao. Emotional attention: From eye tracking to computational modeling. *TPAMI*, Vol. 45, pp. 1682–1699, 2022.

[130] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin P. Murphy, and Alan Loddon Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, Vol. 40, pp. 834–848, 2016.

[131] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2015.

[132] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip H. S. Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*, pp. 18134–18144, 2022.

[133] Li Xu, Mark He Huang, Xindi Shang, Zehuan Yuan, Ying Sun, and Jun Liu. Meta compositional referring expression segmentation. In *CVPR*, pp. 19478–19487, 2023.

[134] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2017.

[135] Yankun Wu, Yuta Nakashima, and Noa Garcia. Not only generative art: Stable diffusion for content-style disentanglement in art analysis. In *ICMR*, pp. 199–208, 2023.

[136] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, pp. 18392–18402, 2023.

[137] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: interpreting stable diffusion using cross attention. In *ACL*, pp. 5644–5659, 2023.

[138] Amelia Katirai, Noa García, Kazuki Ide, Yuta Nakashima, and Atsuo Kishimoto. Situating the social issues of image generation models in the model life cycle: a sociotechnical approach. *ArXiv*, Vol. abs/2311.18345, , 2023.

[139] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. 2022.

[140] Abhishek Mandal, Susan Leavy, and Suzanne Little. Multimodal composite association score: Measuring gender bias in generative multimodal models. *ArXiv*, Vol. abs/2304.13855, , 2023.

[141] Jialu Wang, Xinyue Liu, Zonglin Di, Y. Liu, and Xin Eric Wang. T2iat: Measuring valence and stereotypical biases in text-to-image generation. *ArXiv*, Vol. abs/2306.00905, , 2023.

[142] Preethi Seshadri, Sameer Singh, and Yanai Elazar. The bias amplification paradox in text-to-image generation. *ArXiv*, Vol. abs/2308.00755, , 2023.

[143] Felix Friedrich, Patrick Schramowski, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. *ArXiv*, Vol. abs/2302.10893, , 2023.

[144] Xudong Shen, Chao Du, Tianyu Pang, Min Lin, Yongkang Wong, and Mohan S. Kankanhalli. Finetuning text-to-image diffusion models for fairness. *ArXiv*, Vol. abs/2311.07604, , 2023.

[145] Yankun Wu, Yuta Nakashima, and Noa García. Stable diffusion exposed: Gender bias from prompt to image. *ArXiv*, Vol. abs/2312.03027, , 2023.

[146] Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 786–808, 2023.

[147] Lukas Struppek, Dominik Hintersdorf, Felix Friedrich, Manuel Brack, Patrick Schramowski, and Kristian Kersting. Exploiting cultural biases via homoglyphs in text-to-image synthesis. *ArXiv*, Vol. abs/2209.08891, , 2022.

[148] Aparna Basu, R. Venkatesh Babu, and Danish Pruthi. Inspecting the geographical representativeness of images from text-to-image models. In *ICCV*, pp. 5113–5124, 2023.

[149] 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 *CVPR*, pp. 2818–2829, 2023.

[150] Noa Garcia, Yusuke Hirota, Yankun Wu, and Yuta Nakashima. Uncurated image-text datasets: Shedding light on demographic bias. In *CVPR*, pp. 6957–6966, 2023.

[151] Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. In *ICCV*, pp. 14810–14820, 2021.

[152] Robert Wolfe and Aylin Caliskan. Markedness in visual semantic AI. In *FAccT*, pp. 1269–1279, 2022.

[153] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017.

[154] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. Quantifying societal bias amplification in image captioning. In *CVPR*, pp. 13440–13449, 2022.

[155] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, pp. 771–787, 2018.

[156] Ruixiang Tang, Mengnan Du, Yuening Li, Zirui Liu, Na Zou, and Xia Hu. Mitigating gender bias in captioning systems. In *WWW*, pp. 633–645, 2021.

[157] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *CVPR*, pp. 17886–17896, 2022.

[158] Jiefu Ou, Benno Krojer, and Daniel Fried. Pragmatic inference with a CLIP listener for contrastive captioning. In *ACL*, pp. 1904–1917, 2023.

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

[160] Dongsheng Xu, Wenye Zhao, Yi Cai, and Qingbao Huang. Zero-textcap: Zero-shot framework for text-based image captioning. In *ACM MM*, pp. 4949–4957, 2023.

[161] Tejas Srinivasan and Yonatan Bisk. Worst of both worlds: Biases compound in pre-trained vision-and-language models. *ArXiv*, Vol. abs/2104.08666, , 2021.

[162] Kankan Zhou, Eason Lai, and Jing Jiang. Vlstereoset: A study of stereotypical bias in pre-trained vision-language models. In *AACL/IJCNLP*, pp. 527–538, 2022.

[163] Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *ArXiv*, Vol. abs/2302.00070, , 2023.

[164] Candace Ross, Boris Katz, and Andrei Barbu. Measuring social biases in grounded vision and language embeddings. In *NAACL-HLT*, pp. 998–1008, 2021.

[165] Yifan Zhang, Daquan Zhou, Bryan Hooi, Kaixin Wang, and Jiashi Feng. Expanding small-scale datasets with guided imagination. *ArXiv*, Vol. abs/2211.13976, , 2022.

[166] David Junhao Zhang, Mutian Xu, Chuhui Xue, Wenqing Zhang, Xiaoguang Han, Song Bai, and Mike Zheng Shou. Free-atm: Exploring unsupervised learning on diffusion-generated images with free attention masks. *ArXiv*, Vol. abs/2308.06739, , 2023.

[167] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *ArXiv*, Vol. abs/2306.00984, , 2023.

[168] Rohan Taori and Tatsunori Hashimoto. Data feedback loops: Model-driven amplification of dataset biases. In *ICML*, Vol. 202, pp. 33883–33920, 2023.

[169] Ryuichiro Hataya, Han Bao, and Hiromi Arai. Will large-scale generative models corrupt future datasets? In *ICCV*, 2023.

[170] Simone Fabbrizzi, Symeon Papadopoulos, Eirini Ntoutsi, and Ioannis Kompatsiaris. A survey on bias in visual datasets. *Comput. Vis. Image Underst.*, Vol. 223, p. 103552, 2022.

[171] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *TPAMI*, 2024.

[172] Norman Mu, Alexander Kirillov, David A. Wagner, and Saining Xie. SLIP: self-supervision meets language-image pre-training. In *ECCV*, Vol. 13686, pp. 529–544, 2022.

[173] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *ICLR*, 2022.

[174] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, pp. 18102–18112, 2022.

[175] Melissa Hall, Laura Gustafson, Aaron B. Adcock, Ishan Misra, and Candace Ross. Vision-language models performing zero-shot tasks exhibit gender-based disparities. *ArXiv*, Vol. abs/2301.11100, , 2023.

[176] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

[177] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL*, pp. 74–81, 2004.

[178] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL*, pp. 65–72, 2005.

[179] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pp. 4566–4575, 2015.

[180] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, pp. 382–398. Springer, 2016.

[181] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *WACV*, pp. 1548–1558, 2021.

[182] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pp. 4171–4186, 2019.

[183] Mert Bülent Sariyildiz, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *CVPR*, pp. 8011–8021, 2023.

[184] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves imagenet classification. *ArXiv*, Vol. abs/2304.08466, , 2023.