University of Padova

Department of Mathematics

Master Thesis in Computer Science

Pushing The Limits of Visual Grounding: Pre-training on Large Synthetic Datasets

Supervisor
Prof. Lamberto Ballan
University of Padova

Co-supervisor
Luca Parolari
University of Padova

Master Candidate
Margarita Kosareva

Student ID
2041604

Academic Year

2023-2024

This thesis is dedicated to the people who supported me even in the hardest time. To my parents, to my friends, to my love.

Abstract

Visual Grounding is a crucial computer vision task requiring a deep understanding of data semantics. Leveraging the transformative trend of training controllable generative models, the research aims to demonstrate the substantial improvement of state-of-the-art visual grounding models through the use of massive, synthetically generated data. The study crafts a synthetic dataset using controllable generative models, offering a scalable solution to overcome challenges in traditional data collection processes. The study introduces a synthetic dataset, employing controllable generative models for scalability. Evaluating visual grounding model (TransVG) — on the synthetic dataset showcases promising results, with attributes contributing to a diverse dataset of 250,000 samples. The resulting datasets showcases the impact of synthetic data on visual grounding evolution, contributing to advancements in this dynamic field.

Contents

Listing of figures

Listing of tables

Listing of acronyms

AI	Artificial Intelligence
VG	Visual Grounding
NLP	Natural Language Processing
VDA	Visual Difference Attention
DiDa	Differentiable Difference Attention
GANs	Generative Adversarial Networks
VAEs	Variational Autoencoders
0Shot-TC	Zero-shot text classification
NLTK	Natural Language Toolkit
GLIGEN	Grounded-Language-to-Image Generation
BoxDiff	Box-Constrained Diffusion
TransVG	Transformers for Visual Grounding

1 Introduction

Visual Grounding (VG) stands as a pivotal computer vision task, demanding a profound understanding of data semantics. The essence of VG lies in aligning entity mentions in natural language queries with their corresponding portions in images. However, achieving this alignment necessitates copious annotations, a resource-intensive and challenging endeavor. Despite significant strides in VG techniques over the years, progress in this semantic-rich task remains somewhat constrained.

In the broader landscape of artificial intelligence, marked advancements, particularly in generative models, have emerged as a transformative trend. Notably, a recent trajectory involves training more conditionable generative models, where the output is guided not only by text but also by other conditions such as images, bounding boxes, textual entities, keypoints, and depth masks. This approach renders the output more controllable, opening avenues for enhanced model performance.

The primary objective of this research is to demonstrate that the availability of massive, synthetically generated data can substantially improve the performance of state-of-the-art VG models. The research further aims to harness controllable generative models for crafting a synthetic dataset — an area of increasing significance in AI research. Synthetic dataset generation offers a scalable solution, enabling models to be trained on extensive datasets without the challenges and expenses associated with conventional data collection processes.

The Related Work chapter embarks on an extensive exploration of the diverse visual grounding tasks, describing their varied types and methodologies for constructing effective models. Beyond the theoretical framework, the discussion extends to practical applications of well-solved visual grounding tasks. Furthermore, the chapter delves into prominent datasets widely utilized for pre-training and evaluating visual grounding models, offering insights into their sizes, structures, annotation systems, and inherent limitations. The chapter introduces the concept of synthetic data, delving into various generation techniques that serve as a precursor to the methods employed in this research.

The Method chapter serves as a detailed guide to the methodologies utilized in the creation of a synthetic dataset. It begins by elaborating on the selection process for the base dataset, providing an in-depth analysis of its core attributes and justifying its strategic importance. Throughout the chapter, the methodologies deployed are discussed, emphasizing their critical role in crafting a realistic and expansive dataset. From the initial steps of data collection to the intricacies of attribute-based sentence generation, each methodology is dissected to offer a clear understanding of its implementation and impact on the dataset synthesis process. Moreover, the chapter aims to underscore the significance of these methodologies in achieving the overarching goals of the project. By providing a thorough examination of each step in the dataset synthesis pipeline, it seeks to demonstrate how the combined application of these methodologies contributes to the creation of a high-quality, representative dataset capable of supporting robust research and development efforts.

The Results chapter delves into the chosen VG model — TransVG — evaluating its performance while pre-trained on the synthetic dataset. The comprehensive dataset, its creation methodology, and the experimental design are thoroughly examined. This includes trials with datasets comprising exclusively synthetic images, focusing on the synthetic dataset generated with the attribute semantic task. The chapter seeks to present a complex understanding of the impact of synthetic data on the chosen models, offering valuable insights into their adaptability and efficacy according to the larger scale of the data used for pre-training.

The Conclusion chapter provides a comprehensive summary of the work accomplished. It reflects on the key findings, acknowledges any limitations encountered, and proposes potential avenues for addressing them. Additionally, the chapter offers insights into future research directions, with a particular focus on exploring other semantic tasks such as size and spatial relations. It discusses methodologies that can be employed for these tasks, highlighting the potential for leveraging existing datasets and techniques for dataset augmentation. Preliminary results from initial experiments in these areas may also be discussed, providing a glimpse into the potential impact and significance of future research.

This thesis aims to contribute to the evolving landscape of Visual Grounding by showcasing the potential of synthetic data and controllable generative models in advancing the state of the art.

2
Related work

This chapter provides an in-depth exploration of the visual grounding task, encompassing its diverse types and methodologies for constructing effective models. The discussion extends to the practical applications of well-solved visual grounding tasks. Additionally, the chapter delves into prominent datasets widely utilized for pre-training and evaluating visual grounding models, offering insights into their sizes, structures, annotation systems, and inherent limitations. The definition of synthetic data is introduced, accompanied by a comprehensive overview of techniques for its generation, exploring the various types.

2.1 Visual Grounding

Visual grounding is a concept in computer vision and natural language processing that refers to the ability to connect or ”ground” language descriptions to corresponding visual elements in a scene. The example of visual grounding is presented at Figure 2.1.

The goal is to establish a meaningful link between textual descriptions and the corresponding objects, regions, or entities in an image or video. This area of study has garnered growing interest due to its significant potential in closing the divide between language expressions and visual comprehension.

Visual grounding is characterized by its heightened precision and flexibility compared to image captioning, object detection, object recognition, and instance segmentation tasks. In the context of visual grounding, the indicated object is typically identified through various details

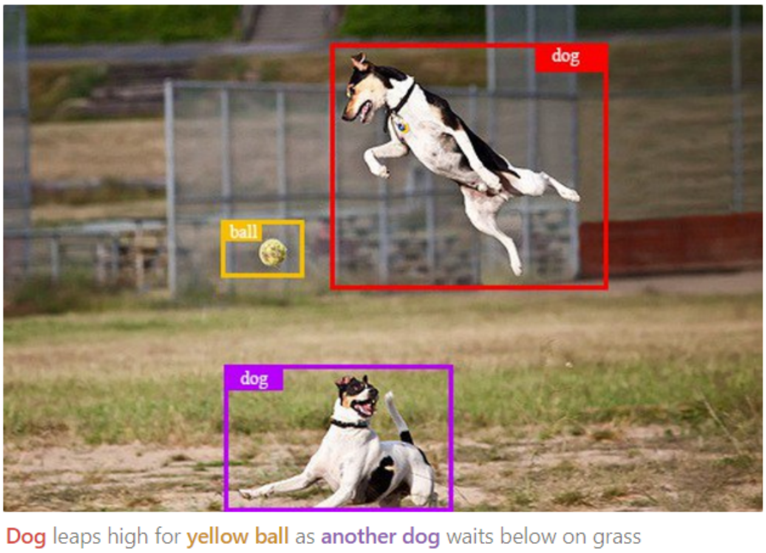

dog
ball
dog
Dog leaps high for yellow ball as another dog waits below on grass

Figure 2.1: Visual Grounding

provided in the language expression. These details may encompass object categories, visual attributes, relational contexts with other objects, attributes, spatial relations (such as relative or absolute positions, size, shape), and more. Consequently, to enhance the precision of reasoning and mitigate ambiguity, it is essential to thoroughly utilize textual information and incorporate discriminative visual features for effective visual grounding.

2.1.1 VG types

The different types of visual grounding task can be categorized based on the level of supervision involved in the training process. Visual grounding can be fully supervised and weakly supervised. The comparison of the models is presented at Figure 2.2.

1. **Fully supervised VG** In fully supervised visual grounding, the model is trained using pairs of annotated data, where each input is paired with a corresponding ground truth annotation that specifies the regions or objects referred to by the given language description.

   The training dataset consists of pairs of examples, each containing an input and a corresponding textual description. For each example, there are detailed annotations that precisely identify the regions or objects in the visual data that are being referred to in the text. These annotations are often provided in the form of bounding boxes, pixel-level segmentation masks, or other region-specific information.

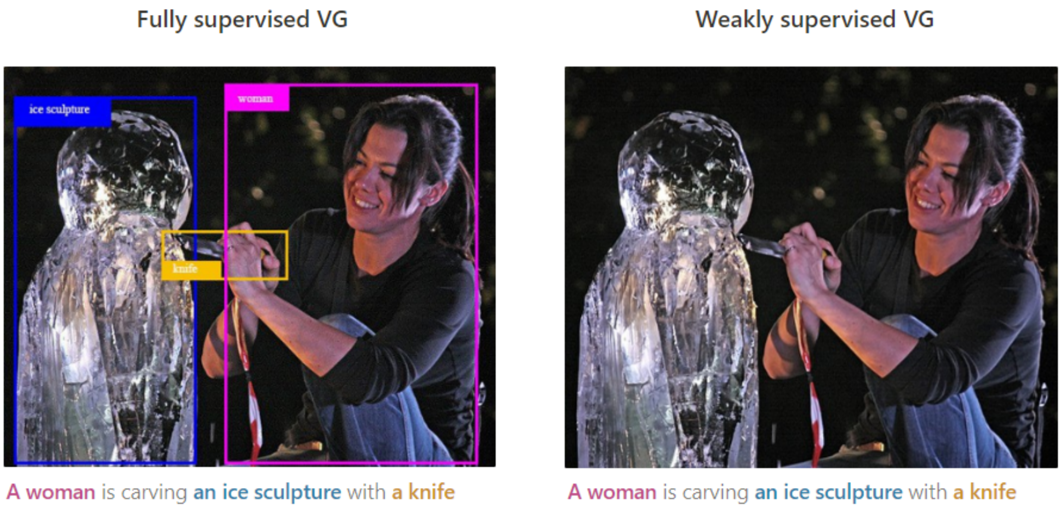

Fully supervised VG
Weakly supervised VG
ice sculpture
woman
knife
A woman is carving an ice sculpture with a knife
A woman is carving an ice sculpture with a knife

Figure 2.2: Comparison of fully supervised VG and weakly supervised VG models

The ground truth annotations serve as the supervision signal during training. The model learns to associate the language descriptions with the correct visual elements by minimizing the discrepancy between its predictions and the annotated regions.

The objective function used during training is typically a loss function that measures the dissimilarity between the predicted regions by the model and the ground truth regions specified in the annotations. Common loss functions for this task include localization-based losses such as mean squared error for bounding box regression or pixel-wise cross-entropy for segmentation tasks.

In a fully supervised visual grounding task, given an image and a textual description like ”a red car in the center”, the model should be trained to predict a bounding box or segmentation mask that precisely outlines the region corresponding to the red car in the center of the image.

Example of fully supervised VG model is VLTVG [1]. This is a transformer-based visual grounding framework that has been developed to directly retrieve the feature representation of the target object for localization. The framework establishes discriminative feature representations through visual-linguistic verification and context aggregation. Subsequently, it employs multi-stage reasoning to identify the referred object, enhancing the precision of object localization in the visual data.

2. **Weakly supervised VG** Weakly supervised visual grounding involves training models with less precise or weaker forms of supervision compared to fully supervised methods. The annotations provided during training are less specific and may not precisely specify the regions or objects in the visual data. This type of learning is often used when

obtaining detailed annotations for large datasets is impractical. For example, instead of bounding boxes, weak supervision might involve image-level labels indicating the presence of certain objects in the image.

The training dataset may contain images paired with textual descriptions, but the annotations are less detailed than in fully supervised scenarios. Instead of precise bounding boxes or pixel-level masks, weak supervision might involve image-level labels indicating the presence of certain objects or concepts in the image. During the training phase the model learns to associate the provided weak annotations with the corresponding visual elements and not to learn the patterns given with the fully annotated files.

The objective function used during training is adapted to accommodate the weaker supervision. For example, if the annotations are image-level labels, the loss function might be designed to encourage the model to focus on relevant regions associated with the labeled concepts without requiring precise localization.

In a weakly supervised visual grounding task with image-level labels, given an image and a textual description like ”a beach scene”, the model is trained to understand that the language description is associated with the general concept of a beach but without precise information about the location of objects within the image.

Weakly supervised visual grounding is more challenging than fully supervised methods because the model needs to infer the relevant regions without explicit localization information. The ambiguity in weak supervision may lead to less accurate localization, and the model needs to rely on contextual information to make predictions.

The advantage of weakly supervised visual grounding is that it allows for training on larger datasets with less manual annotation effort. However, the trade-off is that the model's ability to precisely locate objects in the visual data is limited by the quality and granularity of the weak annotations provided during training.

The example of weakly supervised learning is Confidence-aware Pseudo-label Learning (CPL) where the proposal selection is conducted based on the is determined by cross-modal (region-textual) analysis, involving the direct computation of matching scores between the proposal and the query.

3. **Self-supervised VG** Self-supervised visual grounding refers to a paradigm where a model is trained for the task of associating language descriptions with specific regions or objects in visual data without relying on external annotations. Instead, the model generates its own supervision signal from the input data.

   Self-supervised learning involves designing pretext tasks that do not require external annotations but are still relevant to the target task of visual grounding. These pretext tasks are typically constructed based on inherent properties of the visual and textual data.

> The model learns representations that capture semantic relationships between visual elements and corresponding language descriptions during the pretext tasks. These learned representations are then utilized for the primary task of visual grounding.
>
> Self-supervised visual grounding is advantageous in scenarios where obtaining labeled data is challenging or expensive. By leveraging the intrinsic structure of the data, the model can still learn meaningful representations that benefit subsequent tasks.
>
> The example of self-supervised visual grounding model is the model presented by Agarwal et al [2]. They proposed Visual Difference Attention (VDA) as a differentiable operation and Differentiable Difference Attention (DiDA) loss as a new learning objective. VDA utilizes attention maps by computing the difference in feature vectors between an original image and a version where the salient region is masked out. DiDA, instead, results in attention maps of higher quality and brings about quantitative enhancements in tasks such as classification, detection, and segmentation.

This project will exclusively focus on discussing fully supervised visual grounding models and will present the technology needed to generate sufficiently synthetic datasets that are suitable for visual grounding tasks.

2.1.2 VG approaches

Current existing approaches for building VG models can be divided into three groups: one-stage methodology, two-stage methodology and transformer-based methodology.

Two-stage methodologies are characterized by a dual-phase process, involving the generation of region proposals in the initial stage and subsequently utilizing language expressions to identify the most suitable region in the second stage. Typically, these region proposals are generated through either unsupervised methods or by employing a pre-trained object detector. In the second stage, training loss, manifested as either binary classification or maximum-margin ranking, is applied to optimize the similarity between positive object-query pairs.

Recent advancements in the two-stage methodology involve refining object relationships, incorporating correspondence learning, and leveraging phrase co-occurrences to enhance performance. These improvements represent a concerted effort to bolster the efficacy and adaptability of two-stage approaches in the context of visual grounding tasks.

The two-stage approach offers advantages in terms of flexibility, optimized training, effective integration of language and visual information, modular design, adaptability to complex relationships, and demonstrated success in previous studies.

In contrast to two-stage approaches, one-stage methods streamline the visual grounding process by eliminating the computationally intensive steps of object proposal generation and region feature extraction. Instead, these methods densely integrate linguistic context with visual features, utilizing language-attended feature maps for bounding box prediction in a sliding-window fashion.

Pioneering work in this category includes FAOA [3], which encodes text expressions into a language vector and seamlessly integrates this vector into the YOLOv3 detector. The integration aims to effectively ground the referred instance without the need for explicit object proposal generation.

Another notable one-stage approach is RCCF, which formulates the visual grounding problem as a correlation filtering process. This method selects the peak value of the correlation heatmap as the center for target object localization, offering a computationally efficient alternative.

Addressing the limitations of FAOA in handling complex queries, the recent advancement known as ReSC introduces a recursive sub-query construction module. This module enhances the model's capability to effectively ground complex queries, contributing to improved performance in scenarios where nuanced linguistic expressions are involved.

One-stage and two-stage methodologies have predominantly depended on either extensively pre-trained object detectors or proposal-free frameworks that enhance off-the-shelf one-stage detectors through the integration of textual embeddings. In contrast, the transformer-based approach is constructed upon a transformer encoder-decoder architecture and operates autonomously of any pre-trained detectors or word embedding models.

Two instances of the transformer-based approach are TransVG [4] and VGTR [5]. In the case of TransVG, it leverages transformers for multi-modal correspondence and, through empirical evidence, demonstrates that intricate fusion modules - such as modular attention networks, dynamic graphs, and multi-modal trees - can be effectively replaced by a simpler stack of transformer encoder layers, resulting in improved performance. VGTR is designed to capture global visual and linguistic contexts without relying on the generation of object proposals. It redefines visual grounding as a task of regressing object bounding box coordinates conditioned on the input query sentence. VGTR employs the potent capabilities of transformers to comprehend natural language descriptions, seeking to acquire more discerning visual evidence to mitigate semantic ambiguities.

2.1.3 VG applications

Visual grounding applications leverage the relationship between language and visual content to enhance various tasks.

VG is helpful in robotics while aiding robots and autonomous systems in understanding and interpreting their surroundings [6]. This is particularly important for tasks like navigation, where the system needs to recognize and interact with objects based on visual input.

VG can be helpful in Question-Answering Systems and in particular in Visual Question Answering [7]. It is employed in VQA systems to comprehend and respond to questions about images. It involves linking linguistic queries with specific visual elements, requiring a nuanced understanding of the visual context.

VG is essential in Augmented Reality applications [8] where digital information or objects need to be overlaid onto the real-world scene. This involves accurately mapping and aligning virtual elements with corresponding entities in the visual field. Same stands for Virtual Reality. VG contributes to creating realistic and immersive virtual environments. This includes mapping virtual objects to real-world counterparts and ensuring a coherent user experience.

Overall, visual grounding is crucial for various applications, including image and video captioning, human-robot interaction, and augmented reality. Visual grounding enables machines to understand and interpret natural language descriptions in the context of visual information, enhancing their ability to interact with the visual world.

2.2 Overview of existing VG datasets

This part describes several well-known datasets that are employed for the visual grounding task, serving both as pre-training and fine-tuning resources, as well as benchmarks for model evaluation. However, the constrained size of these datasets poses limitations on achieving optimal model performance. The annotation process, which relies on human resources, is cost-intensive, making it impractical to replicate extensive datasets inexpensively.

2.2.1 Flickr30k

The Flickr30k dataset [9] is a widely used benchmark dataset in the field of visual grounding. It was created for the task of associating textual descriptions with specific regions or objects in images, making it relevant for visual grounding and image-captioning research.

The dataset consists of 31,000 images collected from the Flickr website. Each image is associated with five different textual descriptions, providing diverse linguistic expressions for the visual content. The annotations in Flickr30k are provided in the form of sentence descriptions, where each description is a human-generated sentence describing the content of the image. Annotations are considered at the sentence level, and each image has multiple associated sentences.

2.2.2 ReferItGame

The ReferItGame dataset is a dataset designed specifically for the task of VG. The ReferItGame dataset is created to address this task and provide a resource for evaluating models on visual grounding.

The dataset consists of 130,525 expressions, referring to 96,654 distinct objects, in 19,894 photographs. The images collected from the ReferIt game, an online game where users refer to specific objects or regions in images using natural language expressions. It includes a diverse set of images containing various scenes, objects, and backgrounds. The annotations in ReferItGame are in the form of referring expressions, which are sentences or phrases that refer to a particular object or region in an image. Each image may have multiple referring expressions, capturing the diversity of ways people might describe the same visual content.

2.2.3 MS COCO

MS COCO dataset [10] is a widely used and comprehensive dataset for various computer vision tasks, including visual grounding.

MS COCO is a large-scale dataset that contains images covering a diverse range of scenes and everyday activities. It includes over 200,000 images, each annotated with object instance segmentation masks, bounding boxes, and captions. Annotations in MS COCO are rich and include multiple types of information. Each image has annotations for object instances, providing bounding boxes and segmentation masks for various objects. In addition to object annotations, MS COCO includes five captions for each image, capturing different ways to describe the visual content.

2.2.4 RefCOCO

The RefCOCO dataset [11] is another dataset designed for the task of visual grounding. ”RefCOCO” stands for ReferIt Game Referring to COCO, indicating its connection to the MS

COCO (Common Objects in Context) dataset. RefCOCO specifically addresses the task of referring expression comprehension in the context of the MS COCO dataset.

The RefCOCO dataset is created by selecting a subset of images from MS COCO and providing referring expressions for specific objects in those images. Annotations in RefCOCO consist of natural language sentences or phrases that refer to particular objects in the images. Each image in RefCOCO has corresponding annotations capturing how people might linguistically refer to specific objects within that image.

2.2.5 Visual Genome

The Visual Genome dataset [12] is also a large-scale dataset designed for various computer vision tasks, including visual grounding.

Visual Genome is a rich and extensive dataset, containing over 100,000 images that cover a diverse range of scenes and objects. Each image in the dataset is densely annotated with object instances, object relationships, and scene graphs. Annotations in Visual Genome go beyond simple object detection. Each image is annotated with detailed information about object instances, their relationships, and spatial arrangements in the form of scene graphs. Scene graphs represent relationships between objects and provide a structured representation of the visual content.

2.3 Synthetic data

Synthetic data generation involves the creation of artificial data to augment or supplement real-world datasets for training machine learning models. This process aims to overcome limitations such as data scarcity, annotation costs, or privacy concerns associated with using exclusively real-world data. Here are key aspects of synthetic data generation.

Synthetic data is often generated to supplement existing datasets, providing additional diverse examples to enhance the generalization and robustness of machine learning models. In certain applications where real data may contain sensitive information, synthetic data can be used to create privacy-preserving datasets for model development without revealing actual individual details.

The use of synthetic data offers several advantages, particularly in scenarios involving regulated or sensitive data. It alleviates constraints associated with privacy concerns and regulatory compliance by allowing the generation of artificial data that mimics certain characteristics of the original data. Synthetic data is especially beneficial for creating datasets tailored to specific requirements that may be challenging to achieve with authentic data. It is commonly employed in quality assurance and software testing processes.

However, there are notable disadvantages to using synthetic data. One primary challenge is the difficulty in replicating the complexity present in the original data accurately. The synthetic data may not fully capture the intricate patterns and variations found in real-world datasets, leading to inconsistencies. Additionally, synthetic data cannot straightforwardly replace authentic data because models trained solely on synthetic data might not generalize well to real-world scenarios. Despite the advantages, accurate and authentic data remains crucial for producing meaningful and reliable results in various applications.

2.3.1 Synthetic data genetation techniques

1. **Rule-based Generation**

   Rule-based generation [13] refers to the process of creating synthetic data by applying explicit rules and predefined heuristics to generate samples that adhere to specific patterns or characteristics. This approach does not rely on complex learning algorithms or generative models but rather involves the formulation of explicit rules that guide the creation of artificial data.

   Rule-based generation involves the formulation of explicit rules and conditions that govern the generation of synthetic data. These rules are often based on domain knowledge or an understanding of the desired characteristics of the data. Heuristics, which are practical problem-solving methods or rules of thumb, may be incorporated into the rule-based generation process to handle specific situations or to introduce variability in the synthetic data.

   Rule-based generation is often tailored to a specific domain or application. The rules are designed to capture the essential features and patterns relevant to the task. The generation process is deterministic, meaning that given the same set of rules and initial conditions, the synthetic data will be reproduced identically.

   The effectiveness of rule-based generation depends on the accuracy and completeness of the formulated rules. If the rules do not capture the nuances of the real-world data, the synthetic data may lack realism.

   A disadvantage of the approach is that while rule-based generation provides control over the generated data, it may lack the flexibility to capture complex relationships present in real-world datasets.

2. **Generative Models**

Generative models are an advanced approach for creating synthetic datasets by leveraging sophisticated algorithms to learn and simulate the underlying data distribution. Unlike rule-based methods, generative models learn from existing data and generate new samples that share similar statistical properties. Two prominent types of generative models used for synthetic data generation are Generative Adversarial Networks (GANs) [14] and Variational Autoencoders (VAEs) [15].

Generative models learn the underlying data distribution by analyzing real-world examples. GANs and VAEs, for instance, use neural networks to capture complex patterns and relationships in the training data

Generative models are capable of producing synthetic data that exhibits both realism and diversity. The generated samples capture the statistical characteristics of the training data, offering a more complex representation of the underlying distribution.

2.3.2 Types of synthetic data

1. **Image Synthesis**

   Image synthesis is the process of generating artificial images through computational methods, often driven by algorithms, models, or neural networks. The goal of image synthesis is to create visually realistic or meaningful images that share characteristics with real-world photographs.

   Image synthesis can refer to data augmentation, style transferring, or super-resolution. By generating variations of existing images, synthetic data can be added to the training set, enhancing model robustness and generalization. Style transfer techniques use image synthesis to apply artistic styles to photographs. These methods aim to create visually appealing images by combining the content of one image with the style of another. In super-resolution tasks, image synthesis is applied when low-resolution images are transformed into high-resolution counterparts.

   Some image synthesis methods allow for conditional generation, where specific attributes or features can be controlled during the synthesis process.

   Image synthesis is a dynamic and evolving field with applications across various industries. As algorithms and models continue to advance, the quality and realism of synthesized images are expected to improve, opening up new possibilities and challenges in the realm of computer-generated visual content.

2. **Text Generation**

Text generation is the process of automatically creating coherent and contextually relevant text based on certain input or conditions. This can involve various techniques, from rule-based methods to advanced machine learning models.

Simple rule-based methods involve using predefined templates, grammatical rules, or heuristics to generate text.

Statistical language models, such as n-grams or Hidden Markov Models (HMMs), use statistical patterns in the training data to predict the likelihood of words or sequences of words.

Advanced machine learning models, particularly recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and transformers, have demonstrated significant success in text generation tasks.

Sequence-to-sequence models, often based on encoder-decoder architectures, are used for tasks like machine translation and text summarization.

Attention mechanisms, commonly employed in transformer models, allow models to focus on different parts of the input sequence when generating each element of the output sequence.

In this research project, a hybrid approach is employed, incorporating both rule-based generation and generative model techniques for the purpose of generating an extensive synthetic dataset tailored to the visual grounding task. This comprehensive strategy involves applying generation methods not only to images but also to textual data, ensuring a diverse and varied dataset to enhance the training and evaluation of visual grounding models. Example of **??** reference.

3
Method

This chapter outlines the methodologies employed in creating a comprehensive synthetic dataset. The initial section introduces the selected base dataset chosen for subsequent modifications and augmentations, providing a detailed overview of its main features and the rationale behind its selection. Following this, the core semantic tasks crucial for generating a high-quality dataset are enumerated. These tasks serve as the foundation for every step in the generation process.

The subsequent sections delve into the specific techniques applied for sentence generation and image synthesis. Each technique is thoroughly explained, elucidating the reasoning behind its utilization and how it contributes to the overall dataset quality.

To conclude, the chapter discusses the adaptability of these methods for scaling to other original visual grounding datasets. This exploration broadens the applicability and relevance of the developed methodologies beyond the initial dataset, highlighting the potential for widespread use and impact.

3.1 Dataset

The selected dataset for subsequent augmentation is derived from the Flickr30k Entities, an expanded iteration of the original Flickr30k dataset. This dataset encompasses a substantial corpus featuring 244,000 coreference chains, establishing connections between references to identical entities across diverse captions associated with a given image. Additionally, it incorporates 276,000 meticulously annotated bounding boxes, providing spatial delineations for the entities of interest. This enriched dataset not only amplifies the quantity of available data but also enhances the semantic depth and structural understanding through the incorporation of coreference chains and bounding box annotations.

The Flickr30k Entities dataset comprises a total of 31,783 images, consistent with the quantity found in the original Flickr30k dataset. Each image, on average, features 8.7 objects, thereby contributing to a rich and diverse visual context. The dataset encompasses a comprehensive spectrum of 44,518 distinct categories, with an average of 6.2 objects per category, reflecting a notable diversity and granularity in object representation.

The dataset maintains a linguistic dimension by providing five distinct sentences to describe each image, aligning with the structure observed in the original Flickr30k dataset.

Each sentence within the dataset is annotated with entities falling under distinct macro categories, including people, other, notvisual, scene, body parts, clothing, animals, instruments, and vehicles. For instance, an entity mention could be classified as belonging to categories such as people, animals, or scene, reflecting the diverse semantic elements present in the annotations.

Moreover, the dataset accounts for the potential ambiguity in entity categorization through the introduction of merged categories, such as vehicles/scene, animals/people, clothing/people, bodyparts/people, people/scene, animals/scene, clothing/scene, bodyparts/scene, clothing/vehicles, among others.

The annotation process assumes that any noun-phrase (NP) chunk within the sentences can potentially represent an entity mention. These NP chunks, characterized by their brevity (with an average of 2.35 words) and non-recursive nature, capture succinct and meaningful segments within the sentences. The diversity of entity mentions within the dataset is notable, encompassing references to single entities (e.g., "a dog"), regions of "stuff" (e.g., "grass"), multiple distinct entities (e.g., "two men", "flags", "football players"), groups of entities that may not readily be identified as individuals (e.g., "a crowd", "a pile of oranges"), or even the entirety of a scene (e.g., "the park").

The process of generating bounding boxes within the dataset follows specific guidelines. Notably, for entities categorized under "scene", the creation of a bounding box is deemed unnecessary. In cases where individual entities can be distinctly identified, such as with "a man", "a dog", individual bounding boxes are employed to encapsulate each entity. Conversely, when the individual elements within a group, such as "a crowd of people", cannot be readily distinguished from one another, a singular bounding box is utilized to encompass the entire group. This approach acknowledges the variability in entity representation, ensuring that bounding boxes appropriately reflect the perceptual and structural distinctions among entities, whether

Category	Number of appearances	Example
people	144,931	**[/EN#262852/people A man]** wearing [/EN#262855/clothing a blue wrestling suit] with [/EN#262860/other an US emblem] , is wrestling **[/EN#262852/people another person]** in [/EN#262856/other a competition] setting .
animals	15,916	**[/EN#116768/animals A black and white dog]** is playing with [/EN#116769/other an orange ball] in [/EN#116770/scene the snow] .
clothing	52,179	Here are [/EN#247722/people ten people] skating and [/EN#247727/people one guy] without **[/EN#247729/clothing a shirt]** dancing on [/EN#247725/other concrete] in [/EN#247724/scene a tree-lined park] .
instruments	4,485	[/EN#9319/people A man] plays **[/EN#9321/instruments saxophone]** next to [/EN#9320/other a yellow fire hydrant] .
vehicles	12,615	[/EN#139627/people A boy] does [/EN#0/notvisual tricks] on **[/EN#139628/vehicles a bicycle]** while in [/EN#139631/scene a park] .
body parts	15,535	[/EN#43116/people A young boy] with **[/EN#43129/bodyparts close-cropped hair]** , wearing [/EN#43121/clothing a red robe] , is holding [/EN#43122/other a black kettle] as [/EN#43125/people someone] is about to pour [/EN#43126/other something] in [/EN#0/notvisual it] .
scene	73,409	[/EN#0/notvisual There] are [/EN#206621/people men] playing [/EN#206623/instruments the drums] , while walking along **[/EN#206622/scene the street]** .
not visual	26,371	[/EN#97320/people An adult] holds [/EN#97319/people a small child] **[/EN#0/notvisual who]** sits on [/EN#97322/other a table] in [/EN#97321/scene a mall food court] .
other	99,050	[/EN#0/notvisual I] won **[/EN#193444/other the trophy]** at [/EN#193447/people the parade] !

Table 3.1: Description of categories in F30K Entities dataset

individual or grouped, within the dataset.

The selection of this dataset as the foundation for constructing a synthetic augmented dataset stems not only from its inherent diversity but also from the efficiency of its annotation system. Flickr30k Entities is evident in its expansive range of categories and diverse scenes. well-organized entity labeling system contributes significantly to the synthesis of diverse and contextually relevant sentences, enhancing the naturalness and coherence of generated textual content. The provision of precise coordinates facilitates efficient manipulation and correct positioning of entities in newly generated images. This not only streamlines the synthesis process but also ensures the spatial accuracy of generated scenes, bolstering the overall quality and realism of the synthetic visual content.

3.2 Semantic tasks

In the context of VG, various semantic tasks can be employed to generate sentences that effectively convey information about the depicted scene. The following categories encompass key strategies.

1. **Attributes**
   Queries in this category focus on detailing the descriptive characteristics of the object. Attributes may encompass a range of features, such as color, material, ethnicity, and more. In this work only the color attributes are used.

2. **Spatial Relations**
   Spatial relations involve describing the location of objects either in absolute terms or relative to other elements in the scene. Absolute location queries specify the precise placement of the object (e.g., ”woman on the left”), while relative location queries establish the object’s position in relation to another entity (e.g., ”girl standing under the bridge”).

3. **Size**
   Size considerations can be expressed in terms of absolute or relative dimensions. Absolute size queries describe the object’s size outright (e.g., ”big dog”), while relative size queries indicate the object’s size in comparison to another element or other elements in the scene (e.g., ”the smallest bowl”).

These semantic tasks can be manipulated within sentences to achieve diverse expressions and meanings, thereby enhancing the effectiveness of VG in understanding and describing visual content.

This thesis focuses exclusively on the attributes semantic task for dataset generation, with a specific emphasis on utilizing color as the selected attribute for generating new sentences. Color is chosen due to its straightforward nature and ease of manipulation. While the primary discussion revolves around sentence generation with variations in color, brief descriptions of techniques applicable to other semantic tasks will also be provided.

3.3 Sentence generation

The effective execution of most semantic tasks necessitates a profound comprehension of the provided query to generate a sentence with analogous structure but incorporating specific alterations. To achieve this, preprocessing of the original query is essential to enhance overall understanding.

In the context of the research, the original database includes sentences with annotated entities. However, it's essential to design a generation process that can be easily scaled to other datasets lacking such entity annotations. Thus, a cost-effective source for obtaining these annotations needs to be identified.

The challenge involves text classification without the availability of annotated training data. Text classification is the task of mapping text to labels based on textual descriptions, and the aim is to accomplish this without the aid of annotated training instances.

There are two possible approaches to solve this problem.

First strategy is a similarity-based approach. This approach generates semantic embeddings for both the texts and label descriptions. The matching process involves measuring the similarity between texts and labels using metrics like cosine similarity. Among various similarity-based approaches, the recently introduced Lbl2Vec [16] method has demonstrated superior performance. Lbl2Vec embeds word, document, and label representations jointly. Initially, word and document representations are learned using Doc2Vec. Subsequently, the average of label keyword representations is employed to identify the most similar candidate document representations via cosine similarity. The average of these candidate document representations forms the label vector for each class. During classification, documents are assigned to the class with the highest cosine similarity between the label vector and the document vector.

Another avenue explored is zero-shot classification. This involves leveraging labeled training instances from known classes to train a classifier capable of predicting instances from unseen classes. Notably, zero-shot learning techniques utilize annotated data for training but do not use labels to inform the target classes, relying on knowledge from previously seen classes to classify instances from entirely new classes.

Zero-shot text classification (0Shot-TC) [17] is an approach that enables a text classification model to categorize instances into classes it has never seen during training. This is achieved through the use of pre-trained language models and embeddings, which capture semantic relationships and contextual information about words and phrases. The foundation of zero-shot text classification is the use of pre-trained word embeddings or contextual embeddings. Pre-trained embeddings enable the model to understand the semantic relationships between words and phrases. The embeddings encode contextual information, capturing the meaning of words in various contexts.

The zero-shot text classification model is typically built on top of a pre-trained language model. This model could be a transformer-based architecture like BERT or GPT, where the pre-trained embeddings serve as the foundation. In zero-shot learning scenarios, the model is provided with additional information about the classes it has not seen during training. This information can come in the form of class descriptions, attributes, or any relevant metadata that characterizes the unseen classes.

The demonstrated preprocessing plays an important role in enhancing comprehension of semantic tasks and refining algorithms for sentence generation.

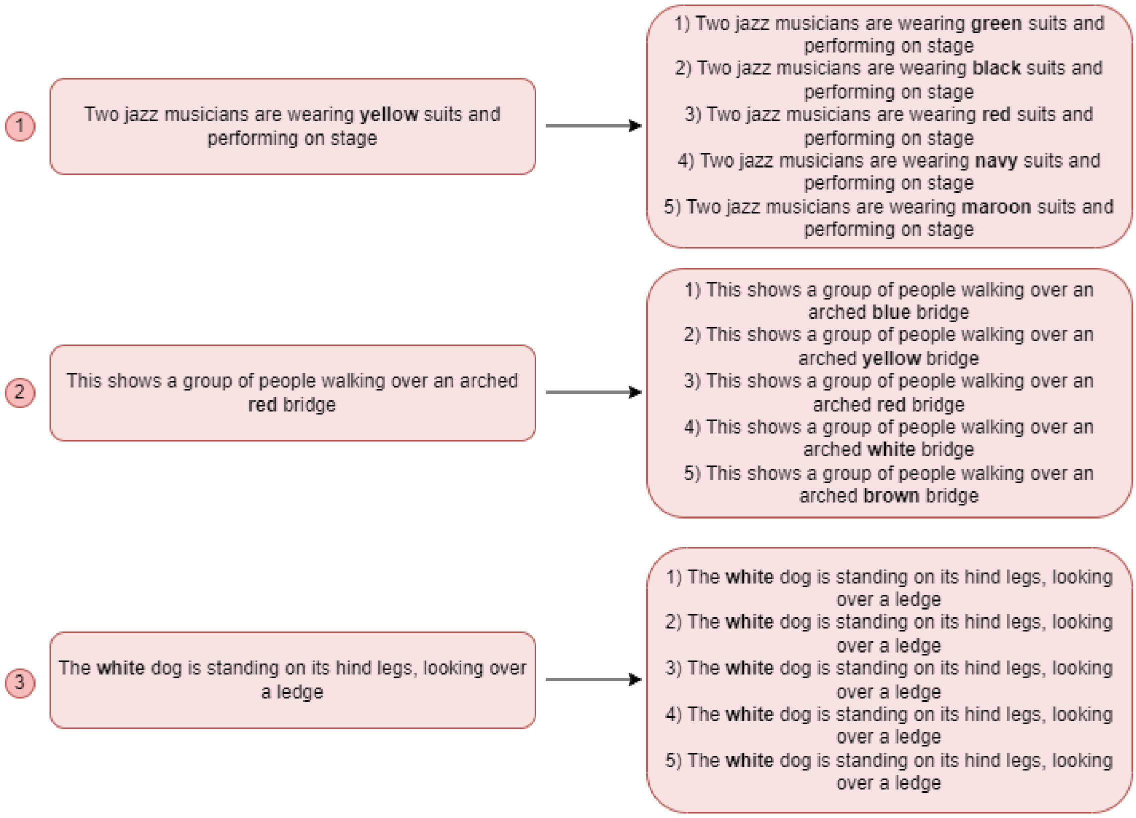
3.3.1 Query transformation with attributes

The sentence transformation technique employed in this project is attribute-based query transformation, a category that encompasses a broad range of object characteristics, surpassing the scope of spatial relations, and size. Attributes can manifest as diverse features such as color, shape, texture, and more, contributing to the richness of entity descriptions.

For simplicity of implementation, only the color attribute was utilized in this project. The rationale behind this choice stems from the ease of understanding and clarity associated with color. Unlike attributes like material diversity or animal race, which pose challenges for the model in terms of distinction, color is more intuitively comprehensible. Hence, it was chosen to assess the visual grounding models' ability to handle this semantic task.

To generate new sentences incorporating the color attribute, only sentences containing a color token were selected from the original dataset. This selection process does not require preprocessing, as colors are uniquely identifiable without ambiguity in the sentence.

Expanding upon the process of generating sentences based on attributes, particularly focus-



Two jazz musicians are wearing **yellow** suits and performing on stage

1) Two jazz musicians are wearing **green** suits and performing on stage
2) Two jazz musicians are wearing **black** suits and performing on stage
3) Two jazz musicians are wearing **red** suits and performing on stage
4) Two jazz musicians are wearing **navy** suits and performing on stage
5) Two jazz musicians are wearing **maroon** suits and performing on stage

This shows a group of people walking over an arched **red** bridge

1) This shows a group of people walking over an arched **blue** bridge
2) This shows a group of people walking over an arched **yellow** bridge
3) This shows a group of people walking over an arched **red** bridge
4) This shows a group of people walking over an arched **white** bridge
5) This shows a group of people walking over an arched **brown** bridge

The **white** dog is standing on its hind legs, looking over a ledge

1) The **white** dog is standing on its hind legs, looking over a ledge
2) The **white** dog is standing on its hind legs, looking over a ledge
3) The **white** dog is standing on its hind legs, looking over a ledge
4) The **white** dog is standing on its hind legs, looking over a ledge
5) The **white** dog is standing on its hind legs, looking over a ledge

Figure 3.1: Example of sentences generated with randomized colors

ing on color, involves a multi-step approach aimed at enriching the dataset with diverse and contextually relevant examples.

To begin with, the identification of sentences containing explicit color tokens within the original dataset is crucial. This step ensures that only sentences with clear references to color attributes are selected, streamlining the subsequent transformation process and minimizing ambiguity.

Once identified, these sentences serve as the foundation for generating new sentences with color variations. The core strategy involves substituting the specified entity's color attribute with a different color while retaining the entity's root and other attributes intact. This meticulous replacement process ensures that the semantic context of the sentence remains consistent while introducing variability in color attributes.

The randomness in assigning new colors adds an element of unpredictability and diversity to the generated sentences, effectively expanding the dataset's breadth and enabling more comprehensive training of visual grounding models. Iterating this process multiple times further enhances dataset diversity, capturing a wide spectrum of color attribute variations and facilitating robust model learning.

The introduction of randomness in assigning colors can result in particularly challenging samples for both image generation models and visual grounding (VG) models. Examples like ”blue banana” or ”green dog” present instances where the assigned color deviates significantly from the expected or typical color associated with the object, posing difficulties for model interpretation and generation.

The generation of new sentences involved replacing the entity with the color with another entity sharing the same root and possessing the same range of attributes, but with a different randomly assigned color. This procedure was repeated multiple times to introduce greater diversity into the dataset.

Furthermore, the iterative nature of attribute-based sentence generation fosters continuous refinement and optimization of the dataset, ensuring its adaptability to evolving research needs and challenges.

In summary, the process of generating sentences based on attributes, particularly focusing on color, involves a systematic approach aimed at maximizing dataset diversity, semantic coherence, and model adaptability. By leveraging advanced techniques and incorporating feedback-driven refinement, attribute-generated sentences serve as invaluable resources for advancing research in visual grounding and related fields.

3.3.2 Query transformation with other tasks

Spatial relations serve as a critical element in characterizing the location of objects within a given context, encompassing both absolute and relative positional distinctions. Manipulating these relations, particularly by substituting them with their opposites, becomes instrumental in inducing semantic variations within sentences.

Absolute locations offer a direct indication of an object's position in the image, devoid of comparative references to other objects. Expressions such as ”bottom”, ”top”, and ”right” precisely denote the object's spatial orientation as perceived by the observer. On the other hand, relative locations delineate the interconnections among two or more objects within the image, introducing concepts of spatial hierarchy such as being ”lower” or ”the lowest”, or positioning ”to the right of” another entity.

In the context of dataset augmentation, a deliberate strategy involves locating spatial relation tokens and substituting them with their opposites. This deliberate alteration serves to not only modify the sentence's meaning but also contributes to the diversification of the generated dataset.

To introduce modifications related to spatial relations in bounding boxes, information from the sentence generation step is crucial. The first key consideration is whether the spatial relation is indicative of absolute or relative location.

For absolute location, adjustments to the bounding box coordinates are necessary, aligning with the entire space of the image. In this scenario, the relevant bounding box pertains solely to the entity. Given the change in spatial relation tokens to their opposites, adjustments along either the horizontal or vertical axes are required. For instance, when transitioning from "a woman on the left" to "a woman on the right", the bounding box shifts horizontally to the right.

In the case of relative location, coordinates are interlinked with multiple objects, typically involving a base case of two objects. Here, the coordinates of both objects necessitate movement to opposite sides. Importantly, the size of the bounding box should remain constant to preserve semantics, preventing unintended alterations, such as making a "little dog" entity significantly larger than "the owner".

Another crucial consideration is the application of similar adjustments to bounding boxes whose coordinates intersect with the original one. This ensures consistency in characterizing intersections, particularly relevant for entities in categories like "instruments", "clothing", "vehicles", and "body parts". These categories are closely tied to the original entities, minimizing the risk of inconsistencies, such as having a "woman wearing a yellow dress" with the woman on the right and the dress on the left.

This straightforward modification to bounding boxes contributes to improved comprehension of spatial relations by visual grounding models. It enhances their ability to accurately identify and align spatial relations in both image and text contexts.

Query transformation involving size introduces the concept of size as either absolute or relative, offering diverse ways to characterize objects within a given context.

Absolute size pertains to the inherent description of an object, resembling an attribute that captures a specific characteristic of the entity. Expressions like "small", "wide", or "huge" fall under this category, encapsulating the determined features of the object.

On the other hand, relative size is employed to compare the object in question with another entity or entities. This comparison can highlight extreme characteristics relative to others (e.g., "the elephant was much bigger than a mouse") or single out an item from a set of similar entities (e.g., "the largest bowl on the table").

Overall, incorporating query generation with size adds an additional layer of diversity to the newly generated synthetic dataset, enriching the range of language patterns within the dataset.

To incorporate alterations related to size changes into bounding boxes, insights from the sentence generation process become pivotal. It is paramount to discern whether the size change is absolute or relative in nature.

For absolute size changes, the bounding box coordinates should be adjusted while considering the entire image space. The focus here is on the bounding box corresponding to the entity undergoing the size change. For example, transitioning from ”a small dog” to ”a large dog” could involve an increase in size, but the bounding box remains centered.

In cases of relative size changes, where the size is compared with another entity, adjustments to bounding box coordinates are contingent upon both entities involved. Size comparisons are often made between two objects, and in such instances, the bounding box coordinates for both entities need to be modified accordingly. It is crucial to maintain the original size relationships to prevent semantic distortion. For instance, changing the size of ”a small cup next to a large mug” requires coordinated adjustments to both bounding boxes.

Similarly, analogous modifications should be applied to bounding boxes whose coordinates intersect with the original one. This ensures consistency in visual representation, particularly for entities in categories closely associated with one another, such as ”instruments”, ”clothing”, ”vehicles”, and ”body parts”.

By implementing these straightforward adjustments to bounding boxes, the visual grounding models can better interpret and comprehend size changes in both image and text, facilitating more accurate alignment between the two modalities.s

3.4 Image synthesis

Having successfully generated sentences and bounding boxes in preceding steps, the next crucial phase involves synthesizing the final components of a robust visual grounding dataset — images. This step aims to represent all modifications obtained during sentence generation effectively capturing and expressing the nuances of the chosen semantic task.

In this section of the chapter, the methodologies employed for image synthesis are comprehensively detailed. These techniques are selected to ensure that the resulting images faithfully reflect the diverse modifications achieved throughout the dataset creation process, thereby encapsulating the intricacies of spatial relations, size, and attribute representation. The synthesis techniques employed are essential for producing a well-rounded and effective visual grounding

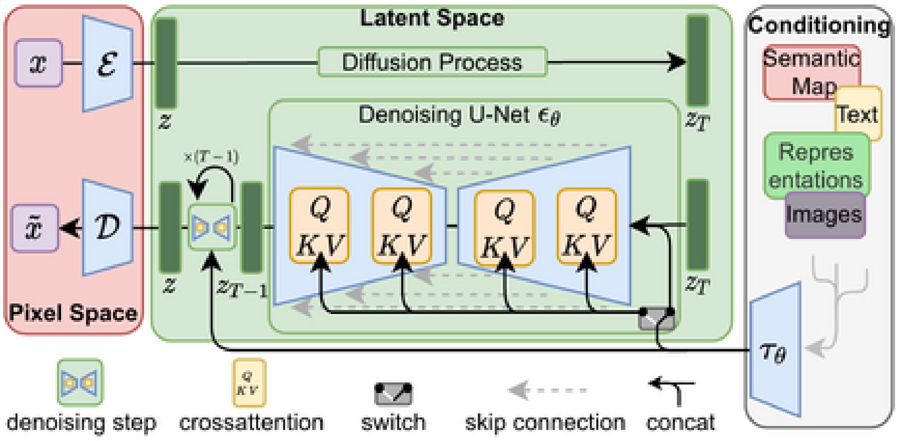


Figure 3.2: Stable diffusion architecture

dataset.

3.4.1 Stable diffusion

Stable diffusion [18] refers to a stochastic process characterized by the stability of its distribution of increments over time. In the context of image generation, stable diffusion offers a novel approach where images are synthesized through a series of diffusion steps, each introducing controlled noise to the image. Unlike traditional generative models that directly learn a mapping from latent variables to images, stable diffusion models leverage the dynamics of diffusion to generate images progressively.

The generation process begins with an initial image, typically a noise sample or a low-resolution image. The image undergoes a series of diffusion steps, where noise is added to the image in a controlled manner. Each diffusion step increases the entropy of the image, progressively blurring it while preserving global structures. The intensity of the noise added at each diffusion step follows an annealing schedule, which gradually decreases over time. This schedule plays a crucial role in shaping the distribution of pixel values in the generated images. At each diffusion step, samples are drawn from the conditional distribution of pixel values given the previous state of the image and the added noise. This sampling process ensures that the generated images follow the desired distribution. After completing all diffusion steps, the final noisy image is reconstructed by reversing the diffusion process. This reconstruction yields a high-quality, photorealistic image with fine details and textures.

Stable diffusion offers several advantages for image generation tasks:

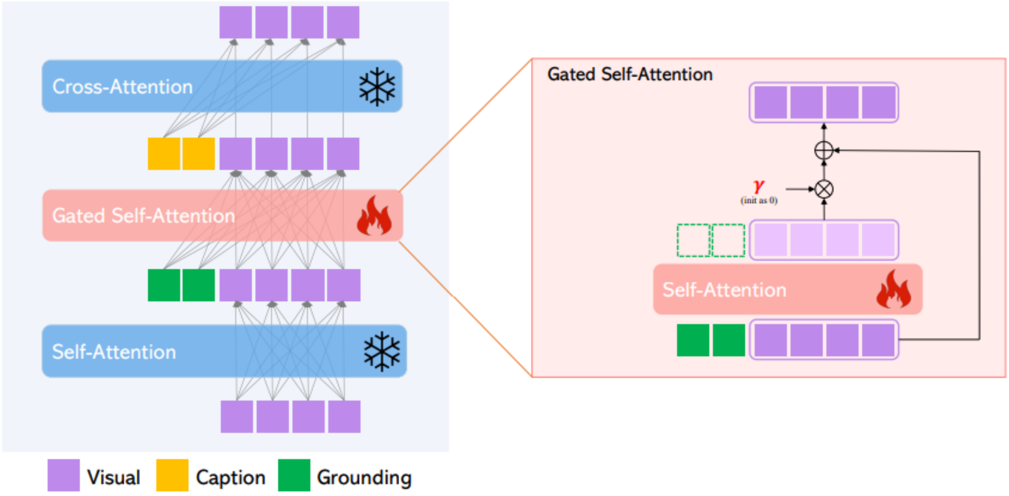


Figure 3.3: GLIGEN architecture

1. **Fine-Grained Control**
   By manipulating the annealing schedule and diffusion parameters, practitioners can control the level of detail and diversity in the generated images.

2. **Robustness to Noise**
   Stable diffusion models are inherently robust to noise, making them suitable for generating images in noisy environments or from low-quality inputs.

3. **Scalability**
   The progressive nature of stable diffusion allows for the generation of high-resolution images with minimal memory requirements, facilitating scalability to large image sizes.

4. **Diversity**
   The stochastic nature of stable diffusion enables the generation of diverse images from the same initial state, offering a rich source of variability in the generated samples.

Stable diffusion offers a promising framework for image generation, leveraging the principles of stochastic processes to produce high-quality, diverse, and photorealistic images.

3.4.2 GLIGEN

Grounded-Language-to-Image Generation (GLIGEN) [19] is an innovative approach that extends existing pre-trained text-to-image diffusion models by incorporating grounding inputs,

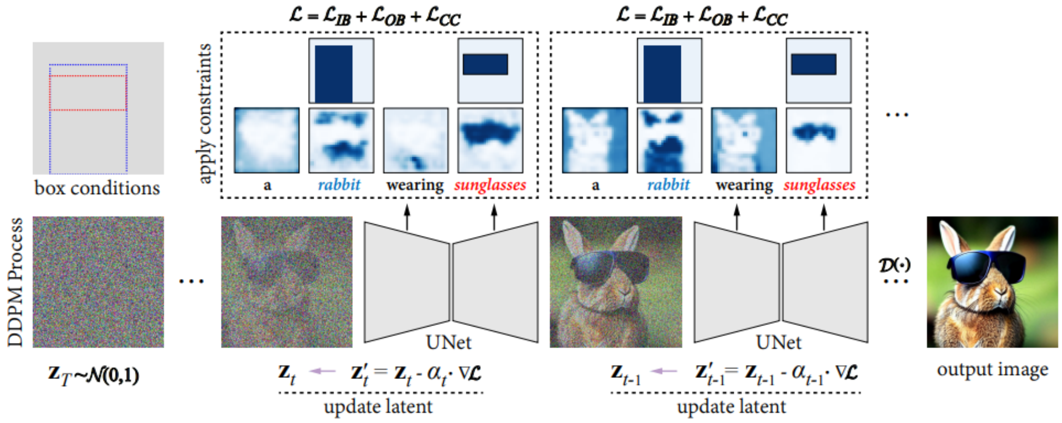

$\mathcal{L} = \mathcal{L}_{IB} + \mathcal{L}_{OB} + \mathcal{L}_{CC}$
$\mathcal{L} = \mathcal{L}_{IB} + \mathcal{L}_{OB} + \mathcal{L}_{CC}$
apply constraints
box conditions
a
rabbit
wearing
sunglasses
a
rabbit
wearing
sunglasses
DDPM Process
$\mathbf{z}_T \sim \mathcal{N}(0,1)$
UNet
UNet
$\mathcal{D}(\cdot)$
output image
$\mathbf{z}_t \leftarrow \mathbf{z}'_t = \mathbf{z}_t - \alpha_t \cdot \nabla \mathcal{L}$
update latent
$\mathbf{z}_{t-1} \leftarrow \mathbf{z}'_{t-1} = \mathbf{z}_{t-1} - \alpha_{t-1} \cdot \nabla \mathcal{L}$
update latent

Figure 3.4: BoxDiff architecture

such as bounding boxes for grounding concepts, reference images, and part keypoints. This novel methodology addresses the challenge of integrating new grounding information while preserving the extensive concept knowledge encoded in the pre-trained models.

The original Transformer block of Latent Diffusion Models consists of two attention layers: self-attention from visual tokens and cross-attention from caption tokens. GLIGEN's primary contribution involves freezing the training weights of the original model and integrating Gated Self Attention layers between the model's attention layers.

This modification facilitates spatial grounding capabilities by directing model attention to the concatenation of visual and grounding tokens. Importantly, this grounding truth injection has no impact on the original model's understanding or pre-trained concept knowledge. The adapted block allows for enhanced user influence by specifying regions to modulate with the novel feature, resulting in a substantial reduction in the cost of tuning the model to a specific concept.

During inference, the model dynamically decides whether to utilize grounding tokens (by adding the new layer) or the original diffusion model (by removing the new layer), known as Scheduled Sampling. This innovation significantly enhances output visual quality by leveraging rough concept locations and outlines in the early steps of re-noising, followed by the incorporation of fine-grained details in later steps.

A noteworthy aspect of this project is the emphasis on spatial control through bounding boxes, providing a significant advantage for aligning semantic tasks: spatial relations, size.

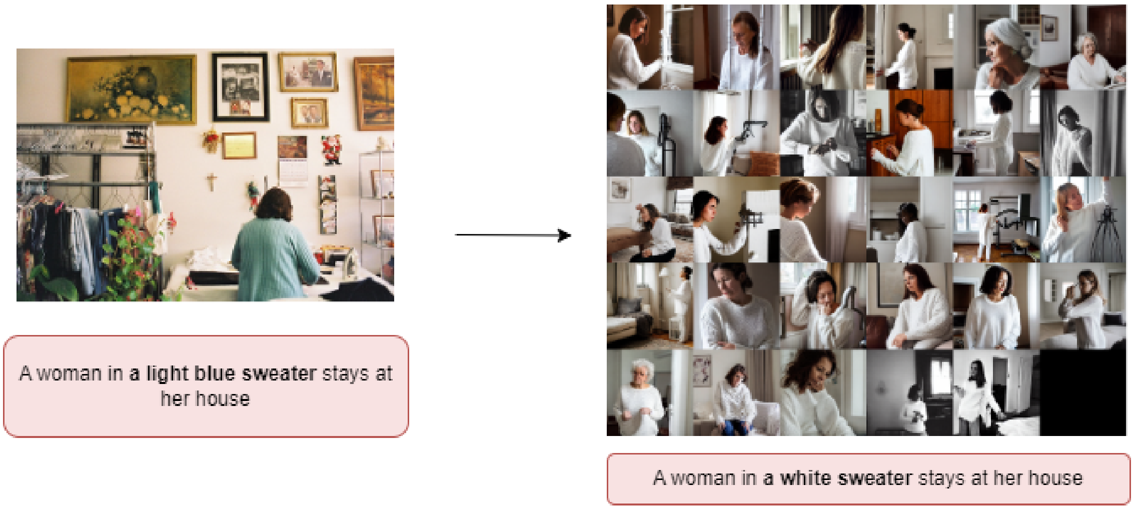

A woman in a light blue sweater stays at her house
A woman in a white sweater stays at her house

Figure 3.5: BoxDiff generation results with provided bounding boxes

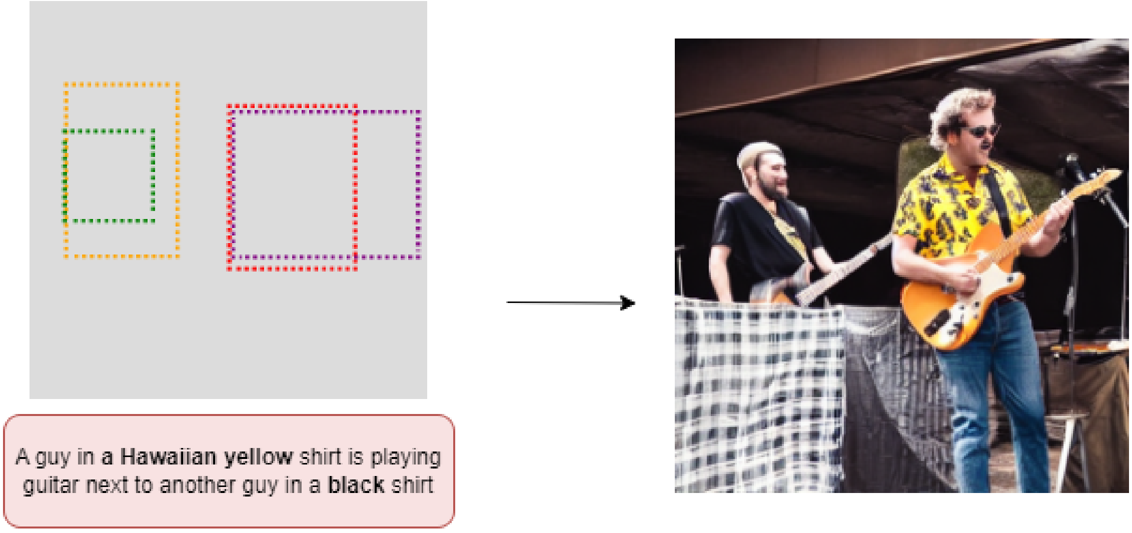
3.4.3 BoxDiff

Box-Constrained Diffusion (BoxDiff) [20] emerges as a noteworthy conditional image synthesis method, leveraging the simplest spatial constraints, such as boxes or scribbles, provided by users. These constraints seamlessly guide object and context synthesis within the denoising step of Stable Diffusion models, eliminating the need for additional model training with extensive paired layout-image data.

Stable Diffusion models incorporate explicit cross-attentions between a given text prompt and intermediate features of the denoiser. This enables the extraction of specific spatial attention maps corresponding to objects or contexts mentioned in the text.

BoxDiff stands out as a training-free approach, enhancing synthesis by introducing three spatial constraints: Inner-Box, Outer-Box, and Corner Constraints. These constraints influence the update directions of the noised latent vector, gradually aligning synthesized objects or contexts with the specified spatial conditions. To address potential fidelity issues arising from strong constraints during the denoising step, a representative sampling approach is explored to mitigate such challenges.

Despite its assertion of controlling spatial relations, empirical tests reveal discrepancies between object positioning and the specified bounding boxes. This misalignment results in challenges for spatial relations, size. However, BoxDiff demonstrates commendable performance in accurately capturing attributes from textual descriptions. Notably, color consistency is maintained as per the textual descriptions, and even intricate structures and patterns in clothing are



A guy in a Hawaiian yellow shirt is playing guitar next to another guy in a black shirt

Figure 3.6: Results of image synthesis using the hybrid GLIGEN + BoxDiff approach

faithfully reproduced.

In summary, while facing limitations in spatial tasks, BoxDiff excels in faithfully rendering attributes outlined in text, contributing to its efficacy in synthesizing detailed and accurate visual content based on textual input.

3.4.4 GLIGEN + BoxDiff

Given the strengths of both the GLIGEN and BoxDiff models, a strategic decision was made to integrate these models for achieving the desired outcomes in the final synthetic dataset. GLIGEN excels in accurately representing bounding boxes, ensuring consistent and precise object positioning, while BoxDiff is adept at expressing attributes effectively.

The versatility of BoxDiff as a plug-and-play component in various diffusion models makes it an ideal candidate for integration with GLIGEN. This collaborative approach leverages the spatial control provided by GLIGEN for accurate bounding box representation and combines it with the attribute expression prowess of BoxDiff.

The final results, as illustrated in the figure, showcase meaningful synthesis. The positioning of men remains consistent, guided by the bounding boxes generated in the previous step through GLIGEN. Simultaneously, the attribute representation is notable, reflecting the successful collaboration of the models.

This integration strategy not only harnesses the strengths of each model but also demonstrates the potential for synergistic effects when combining specialized components for comprehensive and accurate synthetic dataset generation.

4
Results

This chapter provides an in-depth exploration of the chosen visual grounding model selected for testing on the synthetic dataset — TransVG, and the VG dataset chosen for tests. The chapter also offers a comprehensive overview of the created dataset, detailing the methodology outlined in the preceding chapter. Additionally, the chapter delves into the conducted experiments on the candidate model, encompassing trials with datasets comprising exclusively synthetic images, assessments on attribute semantic task to ascertain optimal model performance.

4.1 Color attributes dataset

The color attributes dataset comprises a comprehensive collection of structured data focusing specifically on color attributes associated with various objects. Each entry in the dataset is curated to include detailed information about the object and its corresponding color attribute. Each dataset sample includes a sentence (textual description of object attributes, with a focus on color attributes) and corresponding image sample synthesized using the image generation model.The dataset consists of 250,000 sentence-image pairs that can be used for VG tasks.

The dataset encompasses a wide range of objects from diverse categories, including but not limited to animals, clothing, furniture, and vehicles. For each object, multiple color attributes are utilized, including the ones that are unlikely to appear in real world scenario.

The image samples in the dataset were generated using the Stable diffusion model, a state-of-the-art technique for producing high-quality, realistic images based on textual descriptions.

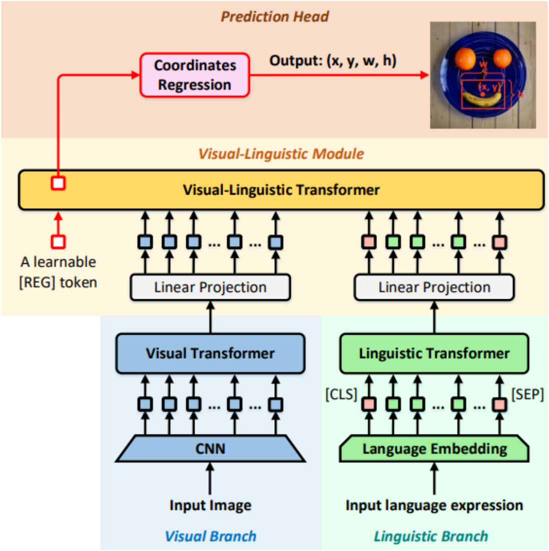


Figure 4.1: TransVG architecture

This approach ensured the creation of visually coherent and contextually relevant image samples that closely corresponded to the attributes described in the accompanying sentences.

In conclusion, the dataset created for this project represents a valuable resource for advancing research in VG field. By combining attribute-based sentence generation with image synthesis, the dataset offers a rich and diverse collection of samples for training and evaluating visual grounding models.

4.2 Training pipeline

This chapter outlines the training pipeline employed in this study to validate the efficacy of the synthetic datasets generated. The pipeline encompasses the selection of a state-of-the-art visual grounding model, pre-training the chosen model on the synthetic dataset, and subsequently fine-tuning it on established visual grounding datasets.

4.2.1 Candidate model

The Transformers for Visual Grounding (TransVG) model [5] is a transformer-based fully-supervised framework for the VG task. TransVG follows a two-staged approach. Notably, the transformer architecture is applied to both the visual and linguistic branches, contributing to a comprehensive understanding of both modalities.

TransVG starts by taking an image and a language expression as inputs. These inputs are then separated into two sibling branches: a visual branch and a linguistic branch. This separation is intended to independently generate visual and linguistic feature embeddings.

The visual and linguistic feature embeddings are combined to form a multimodal feature embedding. To facilitate this combination, a learnable token [REG] is appended to the feature embeddings. This [REG] token plays a crucial role in the subsequent visual-linguistic fusion modules.

The model employs a visual-linguistic transformer that homogeneously embeds the input tokens from different modalities into a common semantic space. This is achieved by utilizing transformer encoder layers, which are adept at capturing both intra-modality and inter-modality context through the self-attention mechanism.

The final output state of the [REG] token is leveraged to directly predict the 4-dimensional coordinates of the referred object in the prediction head. This implies that the model learns to predict the bounding box coordinates (e.g., top-left and bottom-right) of the object mentioned in the language expression.

The model is trained to align the visual and linguistic representations, enabling it to understand the correspondence between objects in the image and their linguistic descriptions. A loss function, likely related to the accuracy of bounding box coordinates predicted by the model, is employed during training to guide the learning process.

Notably, TransVG introduces a novel approach by directly predicting the coordinates of the referred object. This can simplify the training process and potentially lead to more accurate visual grounding results.

4.2.2 Pre-training strategy

The construction of the new synthetic dataset involved executing the primary semantic task of attributes, with a specific emphasis on color variation.

The dataset generation followed a strategic process:

1. **Sentence Generation**
   Novel sentences were derived from existing ones in the Flickr30k Entities [9] dataset using specific rules. Each sentence was deemed meaningful and designed to represent the attribute semantic task, in particular showing the variety of colors.

2. **Conditional Text-to-Image Generation**
   Conditional text-to-image generators were employed to produce new images correspond-

ing to the generated queries. The positioning adhered to the bounding boxes, reflecting the attributes provided in the text.

The strategy involved pre-training the TransVG model on the 250,000 samples included in the attribute-based dataset. The diverse range of color descriptions present in the dataset aimed to train the model to effectively distinguish between attributes and improve its ability to recognize objects based on color attributes. The substantial size of the dataset served as a significant advantage, providing training data to facilitate robust model learning and generalization.

Overall, the strategic approach employed in dataset construction aimed to create a comprehensive and representative dataset for training visual grounding models, with a particular focus on attribute-based tasks and color variation. Through meticulous sentence generation and conditional text-to-image generation, coupled with extensive model pre-training, the dataset aimed to provide a valuable resource for advancing research in VG field.

4.2.3 Fine-tuning strategy

In this thesis, the evaluation and analysis were conducted using the RefCOCO dataset [11] and its variations, namely RefCOCO+ and RefCOCOg. These datasets were chosen due to their ability to showcase diverse semantic tasks. RefCOCO+ primarily emphasizes samples containing attributes, while RefCOCOg is specifically designed for recognizing aspects related to positioning.

The RefCOCO dataset comprises 19,994 images, encompassing 50,000 referred objects with a total of 142,210 referring expressions. Each object is associated with multiple referring expressions. The dataset is officially divided into a training set with 120,624 expressions, a validation set with 10,834 expressions, and a test set with 5,657 expressions.

Similarly, RefCOCO+ consists of 19,992 images featuring 49,856 referred objects and a total of 141,564 referring expressions. The dataset is officially split into a training set with 120,191 expressions, a validation set with 10,758 expressions, and a test set with 5,726 expressions.

RefCOCOg comprises 25,799 images and 49,856 referred objects, with expressions tailored to recognize various positioning aspects.

These datasets serve as comprehensive benchmarks, enabling the assessment of models across different semantic tasks. The official splits into training, validation, and test sets provide a standardized framework for the evaluation of algorithms and models, contributing to a nuanced understanding of their performance under various conditions.

RefCOCO is the original subset of RefCOCO dataset. It contains images with referring expressions generated by players in the ReferIt Game. RefCoco focuses on basic instances of referring expressions without additional complexities. Notably, RefCOCO presents ambiguous expressions. This makes it a suitable benchmark for algorithms that aim to disambiguate and accurately interpret referring expressions in visual contexts.

RefCOCO+ is an extension of the original RefCOCO dataset. It includes additional images with corresponding referring expressions, collected in a similar manner to the original dataset. However, the dataset includes mostly the samples with attributes variety. The goal is to provide a larger and more diverse set of data for training and evaluation.

RefCocoG stands for ”Referring Expressions in COCO Games”. This subset is designed to introduce more abstract and complex referring expressions. RefCocoG images typically involve scenes with more intricate relationships and abstract descriptions, posing a greater challenge for models.

Each subset serves a specific purpose in evaluating the performance of models on different aspects of referring expression comprehension. These differences in subsets help ensure that the RefCoco dataset covers a broad range of scenarios, making it a valuable resource for advancing the visual grounding models, and, therefore, for the evaluation of a new synthetic dataset.

4.3 Experiments

During the experiments the TransVG model underwent pre-training on the synthetic dataset composed by generating samples with different colors.

The attributes dataset was specifically curated to encompass a broader spectrum of descriptions for entities.

The pre-trained model underwent additional fine-tuning on the RefCOCO, RefCOCO+, and RefCOCOg datasets. Subsequently, the performance of these fine-tuned models was compared with that of the original TransVG model, which was trained directly on RefCOCO, RefCOCO+, and RefCOCOg from scratch, respectively.

This comparative analysis aimed to assess the impact of pre-training on synthetic dataset on the overall performance of the TransVG model when applied to a specific visual grounding task - attribute-based with a focus on colors.

The TransVG model was pre-trained on the attribute dataset, comprising approximately 250,000 generated images and 1,000,000 queries. Following pre-training, the model under-

Model	**Training**	**RefCOCO**	**RefCOCO+**	**RefCOCOg**
TransVG	from scratch	69.05	69.02	63.22
TransVG	pre-training on attributes dataset + fine-tuning	**70.66**	**72.01**	64.33

Table 4.1: Comparative performance of TransVG model trained from scratch and pre-trained on synthetic image dataset

went further fine-tuning on the RefCOCO, RefCOCO+, and RefCOCOg datasets, subsequently being evaluated on the respective test splits of each dataset.

The comparative analysis, as presented in the table 4.1, reveals that the model pre-trained on the attribute dataset exhibits superior performance across all subsets of the RefCOCO dataset. In specific terms, the model demonstrated a performance increase of 1.61% on RefCOCO, 2.99% on RefCOCO+, and 1.11% on RefCOCOg.

Notably, the most significant performance growth was observed on the RefCOCO+ dataset, which aligns with expectations given that this dataset emphasizes attribute variety, a focus of the attribute dataset's pre-training. Importantly, the enhanced performance is not limited to RefCOCO+; rather, it extends to all other datasets as well.

Furthermore, it is crucial to highlight that the attribute dataset is substantially larger, with a size 12 times greater (250,000 images) compared to the individual sizes of RefCOCO (20,000 images), RefCOCO+ (20,000 images), and RefCOCOg (25,000 images). This significant increase in dataset size likely contributes to the model's improved performance, indicating the effectiveness of leveraging large synthetic datasets.

5
Conclusion

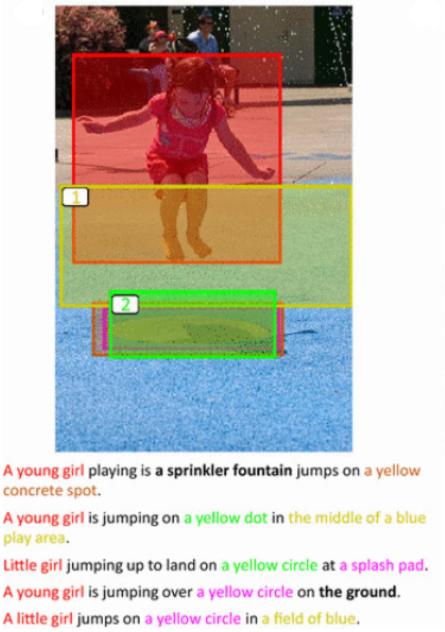
5.1 Discussion

In summary, the construction of the new synthetic dataset involved a meticulous process encompassing the attribute semantic task with great variety of color descriptions, sentence generation, and image synthesis. This strategic approach aimed to create a diverse and comprehensive dataset for training and evaluating models in the visual grounding task.

The experiments involved pre-training the TransVG model on attribute synthetic dataset. Subsequently, pre-trained model underwent fine-tuning on RefCOCO, RefCOCO+, and RefCOCOg, and their performances were evaluated on the respective test splits of each dataset. The results were compared to the TransVG model trained on RefCOCO, RefCOCO+, and RefCOCOg from scratch.

The pre-training on attribute dataset achieved superior performance on RefCOCO, RefCOCO+, and RefCOCOg compared to training from scratch demonstrating performance increases of 1.61%, 2.99%, and 1.11% on RefCOCO, RefCOCO+, and RefCOCOg respectively. It had particularly strong performance growth on RefCOCO+ due to the dataset's emphasis on attribute variety. Moreover, it proved that the larger scale of the synthetic dataset with relevant samples contributes to improved results.

In conclusion, the pre-training strategy contributed to improved performance across RefCOCO datasets The results underscore the importance of dataset characteristics and scale in



A young girl playing is **a sprinkler fountain** jumps on a yellow concrete spot.

A young girl is jumping on a yellow dot in the middle of a blue play area.

Little girl jumping up to land on a yellow circle at a splash pad.

A young girl is jumping over a yellow circle on **the ground**.

A little girl jumps on a yellow circle in a field of blue.

Figure 5.1: Caption error in Flickr30K dataset

shaping the model's ability in the visual grounding task.

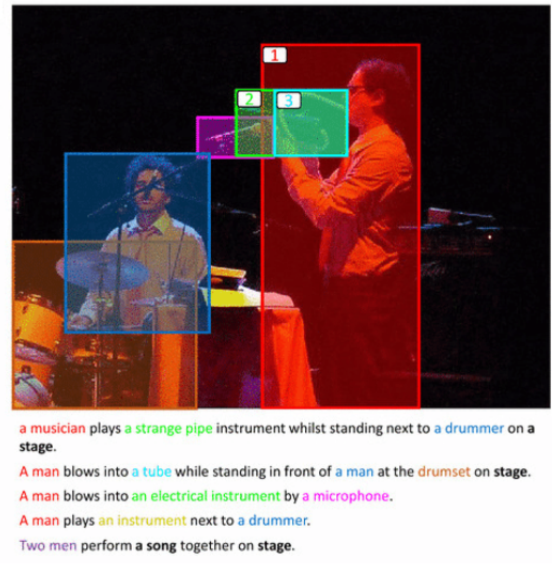
5.2 Limitations and suggestions

5.2.1 Original Flickr30K errors

The utilization of Flickr30K to generate the new synthetic dataset introduces inherent challenges due to errors present in the original dataset. These errors can propagate and amplify in the generated datasets, adversely affecting the performance of visual grounding models.

Two types of errors exist: errors in captions and errors in matching captions with bounding boxes. While the former primarily pertains to language understanding, the latter influences both language comprehension and object detection.

An illustrative example of the first type of error is depicted in the figure. The second caption in the image contains a complex construction error, leading to ambiguity in coreference links. The phrase ”the middle of a blue play area” should be chunked as ”[the middle] of [a blue play area],” indicating that ”the middle” specifically refers to the region containing the yellow dot. The existing coreference link between ”the middle of a blue play area” and ”a field of blue” is invalid, introducing uncertainty about the correct interpretation of the corresponding tan box



a musician plays a strange pipe instrument whilst standing next to a drummer on **a stage**.

A man blows into a tube while standing in front of a man at the drumset on **stage**.

A man blows into an electrical instrument by a microphone.

A man plays an instrument next to a drummer.

Two men perform **a song** together on **stage**.

Figure 5.2: Caption and dismatching error in Flickr30K dataset

(labeled 1). Additionally, the entity mentions, including ”a yellow dot,” ”a yellow circle,” ”a splash pad,” and ”a yellow concrete spot,” are fragmented into three chains with three distinct bounding boxes (labeled 2).

The second type of error is shown in figure. Coreferent entity mentions, such as ”a strange pipe,” ”a tube,” ”an electrical instrument,” and ”an instrument,” are fragmented into three chains. Notably, the phrase ”an instrument” in the fourth sentence is erroneously linked to both boxes 1 and 2 when it should be associated with box 2 alone. Moreover, box 3, corresponding to ”a tube,” is too small, preventing its merger with box 2. An additional example illustrates a mismatch between bounding boxes and captions. In this case, a woman in a store is described as holding an item in her left hand. However, the bounding box corresponds to the item she is holding, leading to a misalignment with the textual description, which refers to the left hand itself. The bounding box for the item in this case is selected for the item that the woman is holding in her right hand.

These examples underscore the critical need for meticulous error analysis and correction in the original dataset, as errors in captions and bounding box associations can adversely impact the performance of visual grounding models. Addressing these issues is paramount for ensuring accurate training and evaluation of models on the synthetic dataset.

Indeed, one potential solution to mitigate the errors stemming from existing datasets is to generate entirely new sentences without relying on pre-existing datasets that might introduce

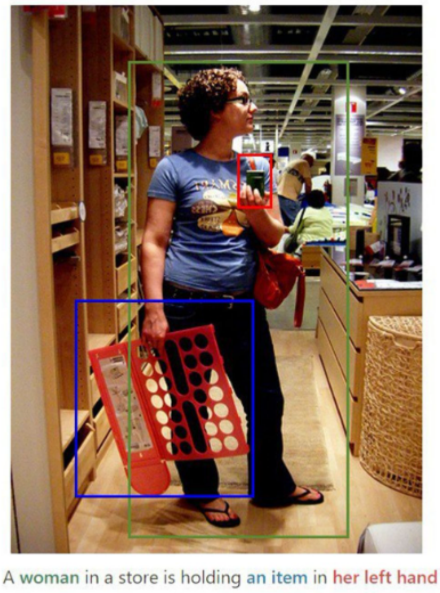


Figure 5.3: Dismatching error in Flickr30K dataset

patterns or errors.

5.2.2 Errors in the generated dataset

A challenge emerged during the image generation process for the attributes semantic task. While the GLIGEN model exhibited impressive results in generating images based on provided bounding boxes and contextual cues, it struggled with sentences that were intricate and detailed, particularly failing in accurately capturing the specified colors of objects, as evidenced in the case of ”<color> hard hat”.

This issue may arise from two potential sources. Firstly, complex sentences with numerous details could overwhelm the generation model, making it difficult to faithfully follow the given instructions. Secondly, the problem might be attributed to a scarcity of relevant examples in the training dataset, such as instances involving a ”purple hard hat.”

Although the number of instances exhibiting these shortcomings is limited, it has the potential to impact the training of VG models negatively, leading to suboptimal performance on the attributes semantic task.

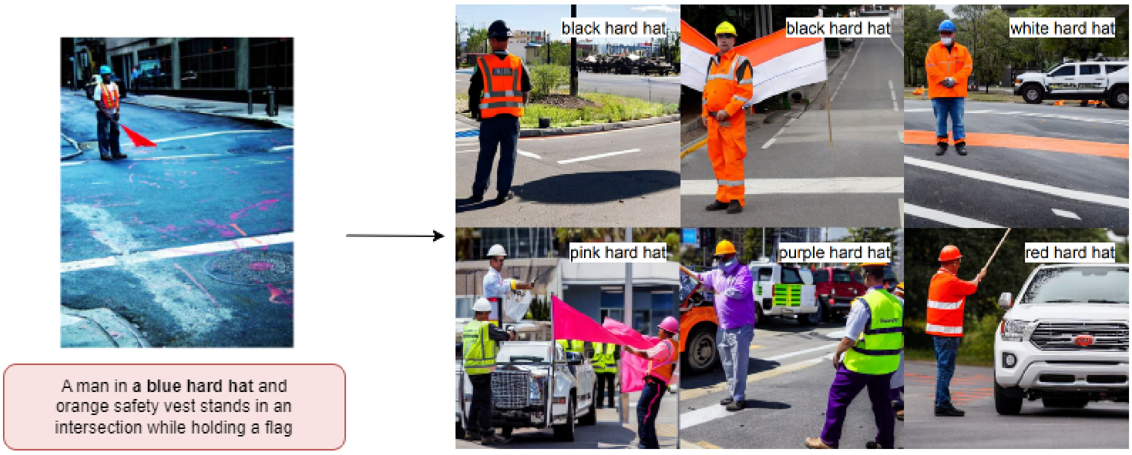

A man in a blue hard hat and orange safety vest stands in an intersection while holding a flag
black hard hat
black hard hat
white hard hat
pink hard hat
purple hard hat
red hard hat

Figure 5.4: GLIGEN generation results with the change of a color attribute

5.3 Future work

5.3.1 Expansion of the attribute dataset

The first suggestion for future work is to expand the attributes dataset. The expansion of the attributes dataset involves incorporating additional attribute types [21]. This expansion should align with the distinct characteristics of objects within each category as specified by the original dataset. For instance, objects in the ”clothing” category might exhibit attributes such as ”material” and ”pattern,” while these descriptors may be less relevant for entities in the ”animals” category. To achieve a comprehensive set of attributes for each category, a systematic division approach can be implemented. For example, the ”clothing” category can be further subdivided into ”top”, ”bottom”, ”headpiece”, ”shoe”, ”bag”, and ”jewelry”. This tailored approach ensures the inclusion of more relevant and diverse attribute samples in the dataset.

5.3.2 Introducing other semantic tasks to dataset generation

One of the suggested directions for future work involves extending the dataset generation process to incorporate other semantic tasks, aiming to explore the broader capabilities of visual grounding models. In particular, emphasis should be placed on tasks related to size and spatial relations, as they offer distinct challenges and delve into the relationships between objects depicted in images and described in sentences.

The integration of size and spatial relations tasks into the dataset generation process necessi-

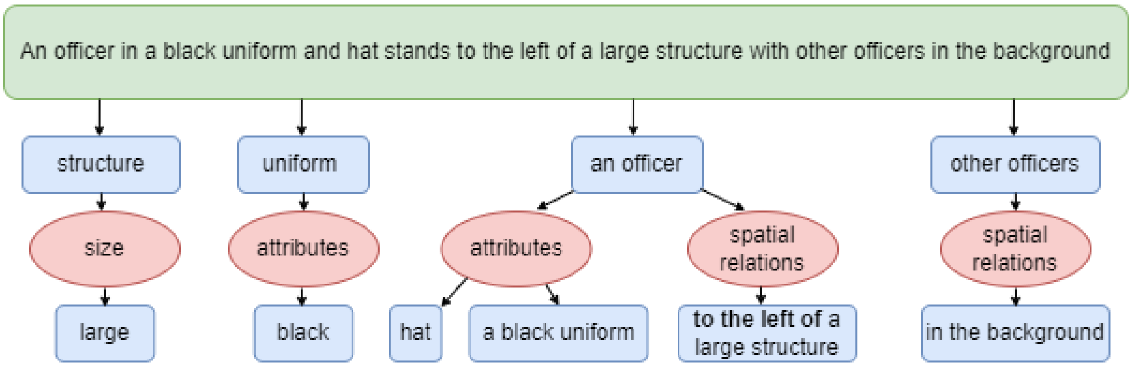

An officer in a black uniform and hat stands to the left of a large structure with other officers in the background
structure
uniform
an officer
other officers
size
attributes
attributes
spatial relations
spatial relations
large
black
hat
a black uniform
to the left of a large structure
in the background

Figure 5.5: Representation of the semantics of the sentence

tates preliminary sentence preprocessing.

The extraction of information from sentences is facilitated by the use of a linguistic tool called a syntax tree. Syntax refers to the rules, principles, and processes governing the structure of sentences in a natural language. At its core, syntax entails how various elements such as subjects, verbs, nouns, noun phrases, etc., are arranged within a sentence. A syntax tree serves as a visual representation of language structure, graphically depicting the grammatical hierarchy.

Noteworthy tools in the field of NLP for constructing syntax trees include The Natural Language Toolkit (NLTK) [22] and Stanza.

In the constructed tree, the clear relationship emerges between ”an officer” and ”a black uniform and hat”, forming a single noun phrase. Within this structure, ”an officer” serves as the main noun phrase, while ”a black uniform and hat” functions as a prepositional phrase, thereby acting as an attribute of ”an officer”. Within the verb phrase, the central token is ”stands” signifying the action attributed to ”an officer”. Furthermore, the verb phrase indicates that the preposition ”to the left of” establishes a connection between ”an officer” and ”a large structure”, with ”an officer” being the primary entity in this relationship.

Unlike attribute-based tasks, which primarily focus on attributes like color, size and spatial relations tasks involve altering bounding boxes to reflect changes in object size and position as described in the sentences.

For the size semantic task, the dataset includes 7500 training samples and 500 validation samples. These samples undergo bounding box alterations to adjust the size of the objects as described in the sentences. Similarly, for the spatial relations task, the dataset involves changing the position of the bounding boxes to align with their positioning in the sentences.

The training pipeline for the size semantic task follows a similar approach to that of the color attribute dataset. The TransVG model is pre-trained on the size dataset and subsequently

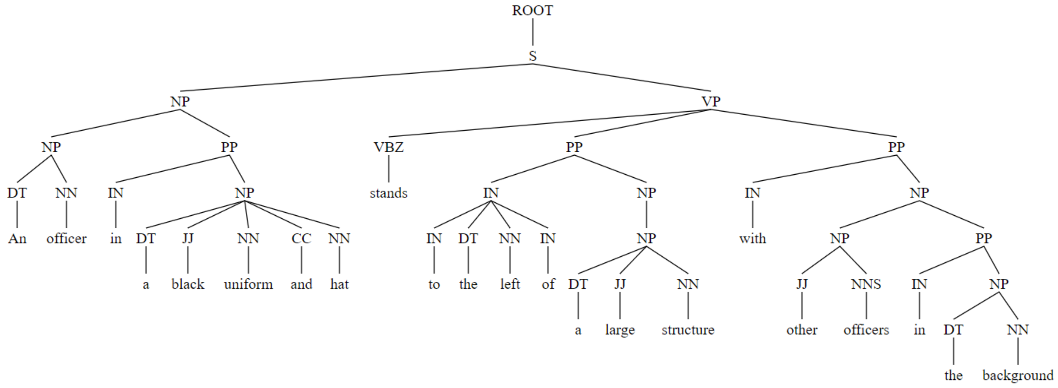


Figure 5.6: Syntax tree of the original sentence generated by Stanza and NLTK tools

fine-tuned on the RefCOCO+ split of the RefCOCO dataset. Despite achieving a validation accuracy of 66.12% on the test split of RefCOCO+, which is 2.9% lower than training on RefCOCO+ from scratch, the results are promising considering the small size of the generated dataset. They demonstrate the feasibility and effectiveness of using artificially generated samples to train VG models.

5.3.3 Usage of other original datasets

Another avenue for future exploration is the incorporation of data from alternative original datasets, such as Visual Genome [12]. Leveraging datasets that already provide information about relationships among entities in both text and images, as well as details like positioning and scene graphs, can mitigate the need for specific sentence preprocessing. This approach offers a potential solution to avoid possible errors during this preprocessing step and can contribute to a more robust and comprehensive dataset.

5.3.4 Generating samples from scratch

The last direction for future research involves developing a methodology for generating sentences from scratch based on specific requirements [23]. This could be achieved through the use of logic templates that define the desired entities at each position and specify the expected relationships among objects. By generating sentences according to the actual needs of the task, this approach may offer a more tailored and flexible solution for attribute-based tasks, providing increased control and customization in the sentence creation process. This avenue opens up possibilities for exploring novel techniques in natural language generation for image-related tasks.

5.4 Concluding words

Visual grounding is the task of associating elements mentioned in natural language queries with their corresponding visual entities in images, establishing a meaningful connection between textual descriptions and visual content. In the context of fully supervised visual grounding, the model undergoes training using pairs of annotated data. Each input in the training set is associated with a corresponding ground truth annotation, meticulously indicating the specific regions or objects in the image referred to by the accompanying language description. Nowadays, the models trained to solve fully supervised visual grounding task are limited by the size of the annotated datasets which are hard and expensive to collect. This thesis introduces a more efficient and cost-effective method for data generation, demonstrating its viability for training fully supervised visual grounding models. The research validates the importance of relevant samples, highlighting the impact of dataset size on model performance.

References

[1] L. Yang, Y. Xu, C. Yuan, W. Liu, B. Li, and W. Hu, “Improving visual grounding with visual-linguistic verification and iterative reasoning,” 06 2022, pp. 9489–9498.

[2] H. Ding, C. Liu, S. Wang, and X. Jiang, “Vision-language transformer and query generation for referring segmentation,” 08 2021.

[3] Z. Yang, B. Gong, L. Wang, W. Huang, D. Yu, and J. Luo, “A fast and accurate one-stage approach to visual grounding,” 08 2019.

[4] J. Deng, Z. Yang, T. Chen, W. Zhou, and H. Li, “Transvg: End-to-end visual grounding with transformers,” 04 2021.

[5] ——, “Transvg: End-to-end visual grounding with transformers,” 04 2021.

[6] Y. Wang, K. Wang, Y. Wang, D. Guo, H. Liu, and F. Sun, “Audio-visual grounding referring expression for robotic manipulation,” 05 2022, pp. 9258–9264.

[7] C. Chen, S. Anjum, and D. Gurari, “Grounding answers for visual questions asked by visually impaired people,” 02 2022.

[8] J. Roh, K. Desingh, A. Farhadi, and D. Fox, “Languagerefer: Spatial-language model for 3d visual grounding,” 07 2021.

[9] B. Plummer, L. Wang, C. Cervantes, J. Caicedo, J. Hockenmaier, and S. Lazebnik, “Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models,” 05 2015.

[10] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. Zitnick, “Microsoft coco: Common objects in context,” 05 2014.

[11] L. Yu, P. Poirson, S. Yang, A. Berg, and T. Berg, “Modeling context in referring expressions,” vol. 9906, 10 2016, pp. 69–85.

[12] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. Shamma, M. Bernstein, and F.-F. Li, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, 05 2017.

[13] Y. Lu, H. Wang, and W. Wei, "Machine learning for synthetic data generation: a review," 02 2023.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[15] D. Kingma and M. Welling, *An Introduction to Variational Autoencoders*, 01 2019.

[16] T. Schopf, D. Braun, and F. Matthes, "Lbl2vec: An embedding-based approach for unsupervised document retrieval on predefined topics," 10 2022.

[17] W. Yin, J. Hay, and D. Roth, "Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach," 01 2019, pp. 3905–3914.

[18] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[19] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. J. Lee, "Gligen: Open-set grounded text-to-image generation," 01 2023.

[20] J. Xie, Y. Li, Y. Huang, H. Liu, W. Zhang, Y. Zheng, and M. Shou, "Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion," 07 2023.

[21] J. Liu, L. Wang, and M.-H. Yang, "Referring expression generation and comprehension via attributes," 10 2017, pp. 4866–4874.

[22] S. Bird, "Nltk: The natural language toolkit," 01 2006.

[23] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," 08 2019.

[24] A. Taylor, M. Marcus, and B. Santorini, "The penn treebank: An overview," 01 2003.

[25] Z. Chen, R. Zhang, Y. Song, X. Wan, and G. Li, "Advancing visual grounding with scene knowledge: Benchmark and method," 07 2023.

[26] C.-H. Ho, S. Appalaraju, B. Jasani, R. Manmatha, and N. Vasconcelos, *YORO - Lightweight End to End Visual Grounding*, 02 2023, pp. 3–23.

[27] J. Deng, Z. Yang, D. Liu, T. Chen, W. Zhou, Y. Zhang, H. Li, and W. Ouyang, "Transvg++: End-to-end visual grounding with language conditioned vision transformer," *IEEE transactions on pattern analysis and machine intelligence*, vol. PP, 07 2023.

[28] Q. Wang, H. Tan, S. Shen, M. Mahoney, and Z. Yao, "Maf: Multimodal alignment framework for weakly-supervised phrase grounding," 10 2020.

[29] L. Yang, Y. Xu, C. Yuan, W. Liu, B. Li, and W. Hu, "Improving visual grounding with visual-linguistic verification and iterative reasoning," 04 2022.

[30] C. Xinpeng, L. Ma, J. Chen, Z. Jie, W. Liu, and J. Luo, "Real-time referring expression comprehension by single-stage grounding network," 12 2018.

[31] Z. Chen, R. Zhang, Y. Song, X. Wan, and G. Li, "Advancing visual grounding with scene knowledge: Benchmark and method," 07 2023.

[32] S. Gupta and J. Malik, "Visual semantic role labeling," 05 2015.

[33] F. Xia, "The part-of-speech tagging guidelines for the penn chinese treebank (3.0)," 11 2000.

[34] A. Agarwal, S. Karanam, and B. Srinivasan, "Learning with difference attention for visually grounded self-supervised representations," 06 2023.

[35] M. Shridhar, D. Mittal, and D. Hsu, "Ingress: Interactive visual grounding of referring expressions," *The International Journal of Robotics Research*, vol. 39, p. 027836491989713, 01 2020.

[36] C. Huang, O. Mees, A. Zeng, and W. Burgard, "Visual language maps for robot navigation," 10 2022.

[37] Y. Zhang, J. C. Niebles, and A. Soto, "Interpretable visual question answering by visual grounding from attention supervision mining," 01 2019, pp. 349–357.

[38] O. Unal, C. Sakaridis, S. Saha, F. Yu, and L. Gool, "Three ways to improve verbo-visual fusion for dense 3d visual grounding," 09 2023.

[39] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *TACL*, vol. 2, pp. 67–78, 12 2014.

[40] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 05 2013.

[41] J. Mao, J. Huang, A. Toshev, O. Camburu, A. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," 06 2016, pp. 11–20.

[42] R. Liu, B. Fang, Y. Tang, and P. Chan, "Synthetic data generator for classification rules learning," 11 2016, pp. 357–361.

[43] S. Mendonca, Y. Brito, C. Santos, R. Lima, T. Araujo, and B. Meiguins, "Synthetic datasets generator for testing information visualization and machine learning techniques and tools," *IEEE Access*, vol. PP, pp. 1–1, 01 2020.

[44] A. Kamath, M. Singh, Y. LeCun, I. Misra, G. Synnaeve, and N. Carion, "Mdetr – modulated detection for end-to-end multi-modal understanding," 04 2021.

Acknowledgments

I would like to thank people that made this project possible and supported me during the whole procedure. I am grateful to my supervisors, Lamberto Ballan, who helped me with a project selection that meets my interests, and Luca Parolari, who stayed in contact with me all the time.