GAFFDNet: GTE-Based Adaptive Fusion and Full-Scale Decoding for Referring Image Segmentation

by

Bushra Rafia Chowdhury
23366010

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
M.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
BRAC University
November, 2025

Declaration

It is hereby declared that

1. The thesis submitted is my own original work while completing M.Sc. in CSE degree at BRAC University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. I have acknowledged all main sources of help.

Student's Full Name & Signature:

Bushra Rafia Chowdhury

Student Id: 23366010

Approval

The thesis titled "GAFFDNet: GTE-Based Adaptive Fusion and Full-Scale Decoding for Referring Image Segmentation" submitted by

- Bushra Rafia Chowdhury (Student ID: 23366010)

of Fall 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of M.Sc. in Computer Science on November 12, 2025.

Examining Committee:

Supervisor:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
BRAC University

Internal Examiner:
(Member)

Dr. Muhammad Iqbal Hossain

Associate Professor
Department of Computer Science and Engineering
BRAC University

External Examiner:
(Member)

Dr. Maheen Islam

Associate Professor and Chairperson
Department of Computer Science and Engineering
East West University, Bangladesh

Program Coordinator:
(Member)

Dr. Md. Sadek Ferdous

Professor
Department of Computer Science and Engineering
BRAC University

Chairperson:
(Chair)

Sadia Hamid Kazi, Ph.D.

Associate Professor
Department of Computer Science and Engineering
Brac University

Ethics Statement

All datasets used in this research are publicly available. These datasets were utilized strictly for academic and research purposes. The work complies with ethical standards concerning data usage, licensing, and privacy, ensuring that no personally identifiable information (PII) is disclosed or misused.
This work aims to advance academic understanding of Referring Image Segmentation while upholding the highest standards of integrity, transparency, and respect for data privacy.

Abstract

Referring Image Segmentation (RIS) requires a precise understanding of both complex visual scenes and natural language expressions to segment target objects accurately. Despite recent advancements, traditional transformer-based methods often face challenges with ambiguous expressions, weak vision-language alignment, and limited contextual reasoning. These limitations hinder the ability to capture nuanced interactions between language and visual context, particularly in cluttered or complex scenes. This research proposed GAFFDNet, an innovative paradigm for referring image segmentation that addresses these challenges. A multi-stage contrastively trained model (GTE) is employed to produce semantically rich and discriminative textual embeddings. A dynamic multimodal fusion module is designed to adaptively integrate visual and linguistic information, which allows the network to focus on the most relevant cues. A full-scale or multi-scale feature aggregation based decoder is incorporated to facilitate dense information exchange across different scales, thereby enhancing both fine-grained spatial detail and global context understanding. Extensive evaluations conducted on three standard benchmark datasets, RefCOCO, RefCOCO+, and G-Ref demonstrate that GAFFDNet consistently outperforms existing competitive methods. These results validate the effectiveness of the proposed approach in achieving precise, context-aware segmentation aligned with complex referring expressions.

Keywords: Referring Image Segmentation, Image Segmentation, Pixel-Level Segmentation, Vision–Language Understanding, Multimodal Fusion, Contrastive Learning

Dedication

This thesis is dedicated to Almighty Allah, who has granted me strength, wisdom, guidance, and good health throughout this journey.
I also dedicate this work to my family and mentors, whose constant support, encouragement, and guidance have been my source of inspiration and perseverance while completing this research.

Acknowledgement

Firstly, all praise is due to Almighty Allah, by whose blessings this thesis has been completed successfully.
I would like to sincerely thank my supervisor, **Md. Golam Rabiul Alam** Sir, for his kind guidance, valuable advice, and continuous support throughout this research. His help whenever I faced difficulties was instrumental in completing this work.
I would also like to extend my sincere thanks to BRAC University Research Lab for providing with access to High-Performance Computing (HPC) resources, which greatly facilitated the implementation and experimentation phases of this research.
Finally, I am deeply grateful to my parents for their unwavering support, encouragement, and prayers. Without their help, reaching this stage of my academic journey would not have been possible.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviations that will be later used within the body of the document

;	Feature Concatenation
α	Projection Function
β	Projection Function
$\downarrow$	Downsampling
γ	Projection Function
$\uparrow$	Upsampling
A_g	Adaptive Gate
C_i	Feature Dimension
C_t	Number of Channel
EV_i	Enhanced Visual Feature
F_i	Decoder Output
H_i	Height of Feature Map
L_a	Language Attention
L_f	Language Feature
L_g	Language Gate
L_ik	Language Key Matrix
L_iv	Language Value Matrix
MM_i	Multimodal Feature
T	Number of Tokens
V_i	Vision Feature
V_iq	Language Query Matrix
W_i	Width of Feature Map

P_d Double-layer Projection Function

P_s Single-layer Projection Function

Chapter 1

Introduction

Referring Image Segmentation (RIS) is a vision–language task that aims to identify and segment a specific object or region in an image based on a natural language query or description. Unlike conventional segmentation approaches that rely on predefined category labels (e.g., cat, dog, car), RIS interprets open-ended textual queries, allowing for flexible and interactive understanding of visual content. This enables fine-grained control over image content and supports a variety of applications. For example, given a natural language expression, the system should correctly identify and segment only the referred object within the image, even when multiple similar objects are present.

RIS is inherently challenging because it requires the joint modeling of both visual and linguistic modalities. The system must understand the semantics of the input query, reason about relationships between objects in the scene, and accurately map linguistic cues to visual regions. This includes handling spatial relationships ("next to the tree"), attributes ("the smaller cup"), and comparative descriptions ("the taller person in the background"). Such requirements demand sophisticated reasoning beyond standard visual recognition or segmentation tasks. The model must precisely interpret the referring expression, localize the correct object, and generate an accurate pixel-level mask. Building upon tasks such as semantic segmentation, instance segmentation, and referring expression comprehension, RIS integrates these capabilities into a unified framework for language-guided, fine-grained scene understanding. It finds applications in areas such as human–robot interaction, visual grounding, image editing, and assistive vision systems.

1.1 Motivation

Referring Image Segmentation (RIS) has strong potential for advancing vision–language technologies by combining the flexibility of natural language with visual understanding, enabling users to interact with and manipulate digital environments more intuitively through plain language. It has a wide range of real-world applications due to its ability to combine natural language understanding with precise visual segmentation, which can be particularly useful in creative industries, e-commerce, and social media content creation.

RIS goes beyond fixed categories, allowing it to handle diverse and evolving real-world descriptions. This scalability makes it highly adaptable for deployment in dynamic environments where descriptions are unpredictable and constantly chang-

ing. By enabling more natural human–computer interaction, RIS allows users to specify objects of interest using everyday language, facilitating tasks such as text-guided image editing, object replacement, or background removal without requiring technical expertise.
In assistive technologies, RIS improves accessibility for visually impaired or elderly users by helping them locate or identify objects based on spoken descriptions. Moreover, RIS supports advanced visual reasoning by incorporating attributes, spatial relations, and comparisons. For example, it can segment "the smaller dog behind the sofa" or "the person holding the blue umbrella," which is crucial for applications in video surveillance, autonomous driving, and other scenarios requiring fine-grained understanding of object relationships.
RIS can also benefit healthcare and medical imaging by assisting professionals in identifying and segmenting specific organs, tissues, or abnormalities from natural language queries, such as segmenting a tumor in MRI or CT scans to improve diagnostic accuracy. Additionally, RIS enables advanced visual search in applications like e-commerce, allowing users to find items using descriptive phrases rather than fixed category labels, enhancing the efficiency and intuitiveness of search systems.
Overall, the versatility of Referring Image Segmentation (RIS) makes it highly valuable across diverse domains, bridging the gap between vision and language to create smarter, more interactive, and accessible technologies.

1.2 Problem Statement

Over the past few years, significant progress has been made in Referring Image Segmentation (RIS), yet it continues to be a highly challenging task. The primary difficulty arises from the need to accurately bridge nuanced natural language understanding with precise pixel-level visual localization. Referring expressions can often be ambiguous, containing comparative phrases (e.g., "the smaller cup") or relational cues (e.g., "next to the red chair"), which require the model to perform careful contextual reasoning to correctly identify the target object. The complexity further increases in scenes with multiple instances of similar objects, where fine-grained reasoning is essential to distinguish the intended target. This challenge is compounded when objects are small, partially occluded, or surrounded by clutter, demanding highly accurate localization and segmentation capabilities.
Another critical challenge is effective visual–language alignment. Models must capture long-range dependencies between linguistic cues and corresponding visual features to correctly interpret the referring expression. In addition, decoder feature fusion remains a significant problem: the multimodal features extracted by the encoder must be effectively combined in the decoder to generate precise and detailed segmentation masks that respect both semantic and spatial information.
These interrelated challenges make RIS substantially more difficult than traditional segmentation tasks. Success in RIS requires sophisticated reasoning across both visual and linguistic modalities, careful feature integration, and robust handling of ambiguous or complex inputs, in order to produce accurate, contextually consistent, and high-quality segmentation results.

1.3 Research Gap

Despite notable progress in Referring Image Segmentation, existing methods exhibit several limitations. Most models use BERT as the language encoder without incorporating contrastive learning, which reduces the discriminative power of textual representations. In addition, existing multimodal fusion modules, such as PWAM [30], rely on fixed operations, limiting their adaptability to diverse visual–linguistic contexts. Furthermore, decoder feature aggregation is often restricted because plain scale skip connections link only corresponding encoder and decoder layers at the same resolution, hindering cross-scale interaction and preventing effective integration of semantic and spatial features. The aim of this research is to develop an advanced Referring Image Segmentation model that improve the language embeddings, multimodal feature fusion, and decoder feature aggregation to address the challenges in RIS and achieve state-of-the-art performance on standard benchmark datasets.

1.4 Research Objectives

The primary objective of this research is to develop an advanced framework for Referring Image Segmentation (RIS) that overcomes the limitations of existing methods in multimodal understanding, cross-modal alignment, and fine-grained visual segmentation. To achieve this goal, the study focuses on the following specific research objectives:

- To develop a contrastively trained language encoder that produces semantically rich and highly discriminative textual embeddings. This objective aims to enhance the model's ability to differentiate between similar referring expressions and to strengthen cross-modal alignment between linguistic and visual features.
- To design an adaptive gating mechanism for dynamic multimodal feature fusion. The goal is to create a fusion strategy that selectively emphasizes the most relevant linguistic and visual cues, enabling the model to effectively capture complex visual–language interactions and improve segmentation performance in challenging scenarios.
- To construct a decoder architecture that aggregates multi-scale features from all encoder and decoder stages This objective focuses on achieving precise segmentation by integrating cross-scale semantic context while maintaining fine-grained spatial information throughout the decoding process.
- To evaluate the proposed method on multiple benchmark datasets and establish its performance superiority over existing approaches. This includes conducting comprehensive experiments to demonstrate the model's effectiveness, generalization ability, and capability to achieve state-of-the-art results in Referring Image Segmentation.

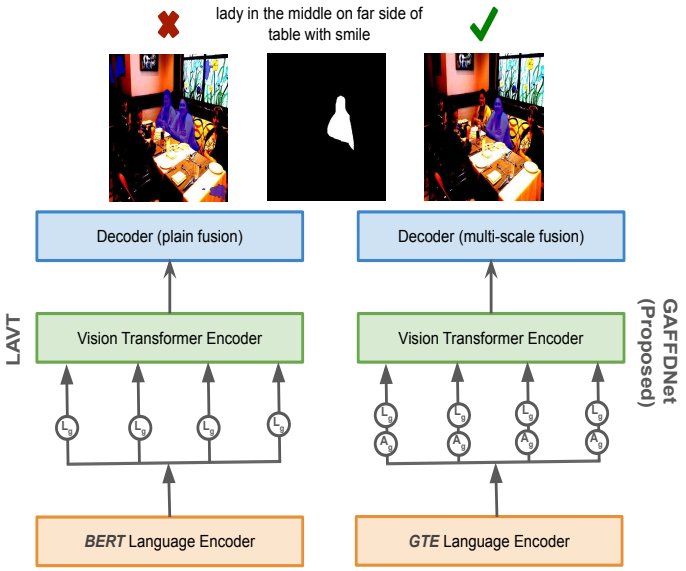

Figure 1.1: Overview of the Proposed GAFFDNet Architecture Compared to Previous Method LAVT.

1.5 Methodology in Brief

In this research, to address limitations in previous language encoders, the proposed method employs a GTE-based multi-stage contrastively trained language model. This produces discriminative embeddings, enhances semantic separation between similar expressions, and improves alignment with visual features. To overcome the fixed operations in the PWAM module [30], which lack adaptability, a learnable adaptive gating mechanism is introduced. This module dynamically weights and fuses visual and linguistic features, allowing the model to emphasize the most relevant modality depending on context. Furthermore, as illustrated in Figure 1.1, the proposed decoder block incorporates full-scale skip connections, enabling dense cross-scale information exchange across all encoder and decoder stages. Unlike plain skip connections, this design allows features to flow in both fine-to-coarse and coarse-to-fine directions, improving semantic reasoning while preserving spatial detail—critical for generating accurate masks in Referring Image Segmentation.

1.6 Structure of the Thesis

The thesis is structured as follows: Chapter 2 presents a detailed review of the relevant literature on Referring Image Segmentation (RIS). This chapter critically analyzes the limitations of existing methods, highlighting challenges which motivate the need for improved approaches. Chapter 3 provides the necessary background and theoretical foundations for the models and techniques used in this research. Chapter 4 describes the methodology in detail, including top-level architecture, datasets, overall architecture of the proposed GAFFDNet model. This chapter compares GAFFDNet with existing state-of-the-art methods on multiple benchmark datasets. Chapter 5 presents implementation process and a detailed evaluation of the proposed method. Chapter 6 concludes the thesis, summarizing the key findings and contributions, and suggesting directions for future research.

Chapter 2

Related Work

Recent advancements in Referring Image Segmentation (RIS) have evolved through diverse multimodal learning approaches, focusing on vision–language alignment, context modeling, and segmentation precision. Below, this section summarizes the progression of RIS research, starting from early foundational works to modern Transformer-based methods, and then highlights the gaps addressed by the proposed approach.

2.1 Early Foundational Models

Early foundational works, such as Hu et al. [3] with Segmentation from Natural Language Expressions (SNLE), combined visual features extracted by Fully Convolutional Networks (FCNs) and sentence embeddings from LSTMs, augmented with spatial cues for segmentation. In this model, the text is encoded using a Recurrent LSTM, which converts the referring expression into a fixed-length vector. The image is processed by a Fully Convolutional Network, or FCN, which produces a spatial feature map. These features are then combined to create a spatial map that highlights the target object. However the limitations of language model that is LSTM which is slow and suffers from an information bottleneck because it compresses the entire sentence into a single vector. The FCN works for basic segmentation, but it struggles to produce precise, context-aware masks, especially for complex sentences or cluttered scenes.
Subsequent models like Li et al. [9] proposed Recurrent Refinement Networks (RRN) that incorporated multi-scale visual semantics and iterative refinement via ConvLSTM, while Liu et al.[6] introduced Recurrent Multimodal Interaction (RMI), which fused language and visual features recurrently on a word-by-word basis. These early approaches primarily relied on separate modality processing followed by simple fusion, limiting their capacity to capture long-range dependencies and complex vision-language interactions. However, LSTM- and ConvLSTM-based models process sequences sequentially, which can struggle with long-range dependencies; they often fail to align subtle linguistic cues with visual features, and their recurrent computations are computationally expensive for high-resolution feature maps. Moreover, simple fusion strategies in these models cannot fully capture complex cross-modal interactions, leading to coarse segmentation or ambiguity.

2.2 Modular Networks

As RIS evolved, researchers began leveraging structured reasoning over language and visual modalities. Modular methods emerged to explicitly leverage the linguistic structure of referring expressions. Yu et al. [10] proposed MAttNet, which decomposed expressions into appearance, location, and relationship modules, using Mask R-CNN and attention-based Bi-LSTM to align structured textual components with visual regions, enhancing interpretability and accuracy. However, its reliance on Bi-LSTM limits the ability to capture long-range dependencies and subtle linguistic nuances, and dependence on Mask R-CNN may reduce flexibility for novel or small objects.

Huang et al. [18] proposed CMPC, which advanced this by detecting candidate regions through entity and attribute words and refining localization via relational words with multi-modal graph reasoning and bilinear fusion. However, the approach can be computationally intensive and may still struggle with fine-grained alignment between complex referring expressions and visual regions.

BRINet [15] incorporated joint visual and linguistic cues to model cross-modal dependencies for precise grouping and segmentation, but its fixed fusion strategies limit adaptability to diverse contexts or novel language descriptions. Despite these advancements, recurrent models remained computationally expensive and struggled with long-range dependencies and fine-grained cross-modal alignment.

2.3 Attention-Based Models

Attention mechanisms have played a pivotal role in improving multimodal fusion and global context modeling. Ye et al. [11] proposed STEP which applied cross-modal self-attention to compute dependencies between each visual region and word, guiding segmentation through recurrent refinement. Ye et al. [14] introduced CMSA (cross-modal self-attention) to capture long-range dependencies by fusing every pixel with linguistic tokens, augmented by spatial coordinates. Hu et al.[16] proposed BCAM, a bi-directional cross-modal attention module combining Vision-Guided Linguistic Attention (VLA) and Language-Guided Visual Attention (LVA) to iteratively refine language and visual features for improved segmentation. LSCM by Hui et al. [19] modeled referring expressions as nodes in a graph linked by syntactic relations via dependency parsing, integrating this with cross-modal attention for richer context modeling and enhanced accuracy. These approaches mark a clear progression from simple fusion and recurrent processing toward modular decomposition, graph reasoning, and advanced attention mechanisms that better capture the complex interplay of visual content and natural language in RIS.

2.4 Transformer-Based Models

More recently, Transformer-based methods have further advanced RIS. VLT [22] reformulates RIS as a direct attention problem using multi-head attention. ReSTR [28] is the first convolution-free model, employing separate Transformer encoders for text and image, followed by cross-modal interactions. EFNet [23] integrates linguistic features via cross-attention within vision encoder layers. LAVT [30] lever-

ages BERT-extracted language features fused pixel-wise with visual features through attention, incorporating a language gate to merge text into multi-scale pyramid features and applying language-aware attention throughout the Swin Transformer backbone for strong cross-modal alignment and efficient decoding.
Building on this, the proposed method enhances multimodal fusion strategy and incorporates multi-scale feature aggregation in the decoder for more precise segmentation. Wang et al. [29] proposed CRIS, a method built upon CLIP, which faces challenges with ambiguous language and complex scenes. This approach addresses these limitations by incorporating a text encoder with multi-stage contrastive learning, thereby enhancing vision-language alignment and improving performance in such challenging scenarios.
In summary, recent works have shown that most RIS methods use BERT [30] as the language encoder. While effective for general NLP tasks, BERT lacks a contrastive learning objective. Without contrastive learning, performance often degrades on short or ambiguous phrases, resulting in suboptimal feature alignment. To address this, proposed method adopt GTE, a multi-stage contrastively trained language model, which produces discriminative embeddings, enhances semantic separation between similar expressions, and improves alignment with visual features. Existing multimodal fusion in [30] PWAM module relies on fixed operations that lack adaptability. This study proposes a learnable adaptive gating mechanism that dynamically weights and fuses visual and linguistic features. Furthermore, existing encoder–decoder method [30] use plain scale connections in the decoder. These connections preserve some spatial information but only link corresponding encoder and decoder layers at the same resolution, which limits cross-scale interaction. Consequently, the model struggles to combine coarse semantic features from deeper layers with fine spatial features from shallow layers, often producing masks that lack contextual consistency and boundary precision. Proposed method designs this decoder like full-scale connections, enabling dense cross-scale information exchange across all encoder and decoder stages. This allows features to flow in both fine-to-coarse and coarse-to-fine directions, improving semantic reasoning while preserving spatial detail, which is crucial for accurate segmentation mask generation.

Chapter 3

Background Study

3.1 Swin Transformer

The choice of the vision backbone is crucial in Referring Image Segmentation (RIS), as it must extract multi-scale visual features that capture both fine-grained object details and global contextual relationships. RIS is the task of localizing and segmenting an object in an image based on a natural language description. This requires the model to understand object-level visual details (such as color, shape, and boundaries) and contextual cues (like spatial relations and relative positions).
Early approaches often relied on convolutional neural network (CNN) backbones, which were effective at capturing local features but struggled to capture long-range dependencies and global context, both of which are essential for accurately identifying the referred object in complex scenes. In this work, the Swin Transformer [26], introduced by Liu et al. (2021) in "Hierarchical Vision Transformer using Shifted Windows", is employed as the vision encoder. Swin is a hierarchical vision transformer designed to reduce the high computational cost of global self-attention while maintaining strong performance on large-scale dense prediction tasks.
A key innovation of Swin Transformer is its hierarchical feature representation, conceptually similar to convolutional neural networks. This design allows it to capture information at multiple scales, making it particularly effective for dense prediction tasks such as semantic segmentation. The hierarchical, multi-scale features produced by Swin make it highly suitable for RIS. Local window attention focuses on object details, while the shifted window mechanism enables global context propagation. Additionally, Swin's computational efficiency allows it to handle high-resolution images effectively, which is essential for producing accurate segmentation masks. Among its variants, the Base model (Swin-B) is widely used due to its balanced trade-off between accuracy and efficiency. The main characteristics are described below.
In patch splitting the input image is divided into fixed-size patches, and each patch is embedded into a feature vector of dimension. Then self-attention is applied locally within non-overlapping windows to capture fine-grained details. In subsequent layers, the windows are shifted to create overlapping regions, enabling cross-window communication and global context modeling. The model produces multi-scale feature maps across four stages, similar to CNNs: stage 1 has 96 channels, 2 Transformer blocks; stage 2 has 192 channels, increased depth; stage 3 has 384 channels, 18 Transformer blocks; stage 4 has 768 channels, 2 Transformer blocks. At first it pre-

trains on ImageNet-22K (22,000 categories) to learn general visual representations. Then it fine-tunes on ImageNet-1K (1,000 categories) for task-specific adaptation. Overall, Swin-B contains approximately 88 million parameters with a computational cost of around 15.4 GFLOPs for 224×224 input images.
It achieves Top-1 accuracy of 86.4% on ImageNet, surpassing traditional CNN baselines such as ResNet-50/101. It also serves as the backbone for state-of-the-art models in object detection and segmentation. In summary, the Swin Transformer combines computational efficiency, multi-scale feature learning, and strong segmentation performance, making it a robust choice as the vision backbone for RIS models.

3.2 General Text Embedding

Like visual encoder, the choice of language encoder is also crucial for Referring Image Segmentation, as it must produce textual representations that align accurately with visual content, capturing both broad semantics and fine-grained details.
Early approaches often employed BERT due to its strong natural language understanding. While it captures syntax and general semantics effectively, BERT struggles to connect linguistic cues such as attributes, spatial relationships, or comparisons to specific objects in an image, often leading to ambiguity in expressions like 'the smaller dog behind the sofa' or 'the person wearing a red hat'. To address these limitations, this work adopts the General Text Embeddings (GTE) model, introduced by Li et al. (2023) in "Towards General Text Embeddings with Multi-stage Contrastive Learning." GTE is a language representation model specifically designed to produce embeddings that are both semantically rich and visually grounded, making it particularly suitable for multimodal tasks such as RIS. Its multi-stage training paradigm and contrastive learning mechanism enable it to encode nuanced relational and descriptive details that are critical for accurate linguistic grounding.
In this network, paired sentences are processed independently, with semantic similarity enforced via a contrastive loss, ensuring robust and efficient sentence embeddings. The model leverages hard negatives semantically similar but incorrect sentences to improve discrimination between closely related expressions. As a result, semantically similar sentences are positioned closer in the embedding space, enhancing cross-modal alignment with visual features. GTE achieves high-quality embeddings with relatively low computational cost, comparable to BERT. Additionally, it employs a two-stage contrastive learning approach. During unsupervised pre-training, the model is trained on 800M text pairs from diverse domains, learning general semantic relationships between words, phrases, and sentences, and distinguishing between semantically similar and dissimilar pairs. Subsequently, task-specific fine-tuning refines the embeddings to align with downstream tasks, such as mapping linguistic cues to visual features, strengthening the model's ability to capture subtle attributes, spatial relationships, and comparative expressions essential for RIS.
In summary, GTE provides semantically rich, visually grounded embeddings that effectively link language to visual content. Compared to traditional models like BERT, it better captures subtle attributes and relational information, enabling more accurate object localization. By producing discriminative and structured embeddings, GTE reduces the need for complex post-processing and supports precise and efficient segmentation in RIS.

Chapter 4

Methodology

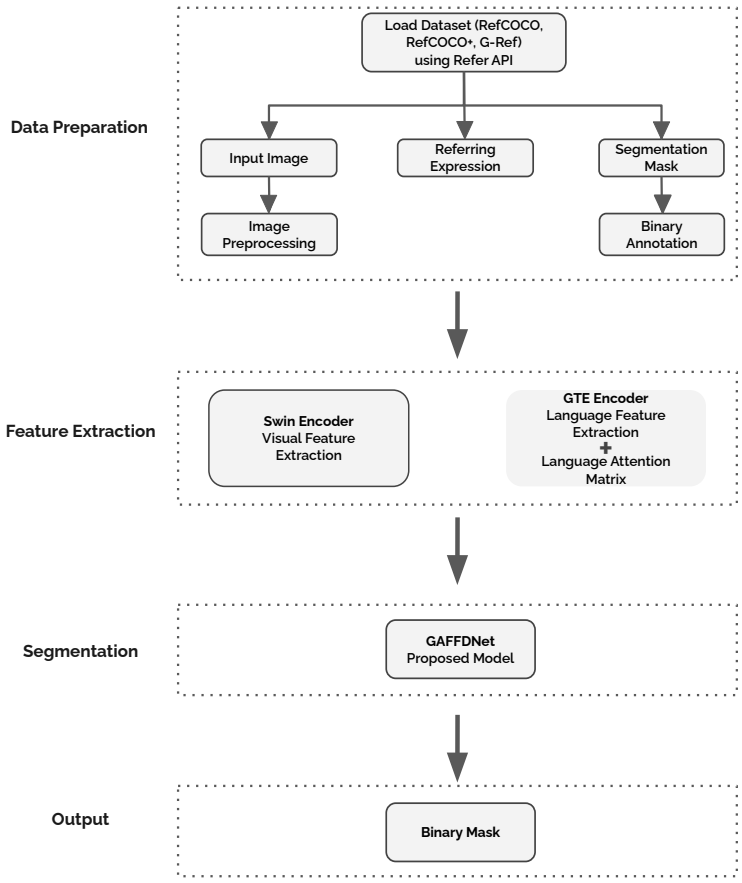

Figure 4.1: Top-level Overview of the Proposed GAFFDNet Model.

Figure 4.1 depicts the top-leve overview or the workflow of the proposed GAFFD-Net model. The primary steps of this workflow include dataset preparation, fature extraction, segmentation and binary mask generation. The dataset preparation step involves selecting datasets from Refer API for training and evaluation, followed by preprocessing to ensure consistency in input formats. The feature extraction step utilizes a vision encoder and a language encoder to obtain visual and linguistic features from images and referring expressions, respectively. These features are then fused using cross-modal attention mechanisms to create multi-modal representations

that capture both visual and linguistic information. Finally, the segmentation step employs a decoder architecture to generate precise segmentation masks for the referred objects based on the fused features. Each of these steps is crucial for the effective functioning of the Referring Image Segmentation model. output

4.1 Dataset Selection

Referring Image Segmentation requires datasets that provide images, pixel-level instance masks, and corresponding referring expressions. Constructing such datasets is particularly challenging and resource-intensive, as it demands both precise manual segmentation and the generation of diverse expressions to describe target objects. Like other researchers in this work, three widely used benchmark datasets is adopted that is built on the MS COCO dataset [2]. These are RefCOCO [5], RefCOCO+ [5], and G-Ref [4] (also referred to as RefCOCOg). All three dataset offers a rich diversity of object categories and real-world scenes.

Table 4.1 summarizes the key characteristics of these three datasets. RefCOCO and RefCOCO+ were collected via an interactive framework known as the ReferItGame, where one annotator described specific objects in MS COCO images using natural language and another annotator identified it. These expressions tend to be short and task-oriented, averaging approximately 3.5 words per expression. RefCOCO+, however, introduces an additional constraint: it explicitly prohibits the use of spatial terms such as "left," "right," or "behind." This restriction forces models to rely more on visual appearance and attribute-based reasoning, rather than positional cues. Both datasets frequently contain multiple same-category objects per image (on average 3.9), making them ideal for evaluating a models instance-level discrimination abilities means distinguishing between individual objects of the same category within an image, like telling one dog apart from another dog, so the model can identify exactly which specific object a referring expression is talking about. The commonly used UNC partition divides these datasets into training, validation, and two distinct test sets: TestA (images containing people) and TestB (images with non-person objects).

G-Ref (RefCOCOg) differs in both annotation method and linguistic structure. It was constructed using a non-interactive annotation process, where annotators independently wrote referring expressions for target mask without system feedback, resulting in expressions that are longer (8.4 words on average) and more descriptive, often containing complex syntax and spatial cues. While G-Ref was designed to focus on images with 2–4 same-category objects, it contains fewer such instances on average (1.6 per image), reducing the frequency of intra-class distractors compared to RefCOCO and RefCOCO+. The G-Ref dataset is available in two splits: the UMD split, which includes a dedicated test set, and the Google split, which lacks a canonical test set. Due to the descriptive and ambiguous nature of its referring expressions, G-Ref presents greater linguistic complexity and challenges related to expression grounding, making it well-suited for evaluating models' capacity for natural language understanding in visually complex scenes. Together, these datasets offer a comprehensive evaluation of a model's ability to understand both visual and linguistic cues in referring expression tasks.

Category	RefCOCO	RefCOCO+	G-Ref
Annotation Style	Interactive	Interactive	Non-interactive
Partition Type	UNC	UNC	UMD
Images	19,994	19,992	26,711
Annotated Objects	50,000	49,856	54,822
Annotated Expressions	142,209	141,564	104,560
Train Samples	120,624	120,191	80,512
Validation Samples	10,834	10,758	4,896
TestA Samples	5,657	5,726	-
TestB Samples	5,095	4,889	-
Expression Length	3.5	3.5	8.4
Same Category Objects	3.9	3.9	1.6

Table 4.1: Overview of Datasets

4.2 Data Preparation

To train and evaluate the proposed Referring Image Segmentation model, a dataset pipeline is established that integrates referring expression datasets through the Refer API and applies systematic preprocessing steps. This preparation ensures that images, segmentation masks, and natural language referring expressions are transformed into consistent input formats suitable for deep learning models. Figure 4.2 illustrates the flow of the data preparation process.

The Refer API provides unified access to widely used referring expression datasets, including RefCOCO, RefCOCO+, and RefCOCOg. Each dataset includes multiple train, validation, and test splits defined by different research groups (UNC, Google, UMD). The Refer API acts as a bridge between raw annotation files and usable data samples. For each split, the API retrieves unique identifiers for each referring phrase and the corresponding image where the referred object is located.

Each natural language referring expression is encoded using the General Text Embedding (GTE) tokenizer from the HuggingFace library. To ensure consistent-length textual inputs, sentences are truncated or padded to a fixed length of 20 tokens. The tokenization process produces two components: tensor embeddings representing the tokenized sentence as a sequence of token IDs, and an attention matrix, which is a binary vector indicating valid tokens and padding tokens.

The corresponding image for each referring expression is loaded in RGB format and preprocessed to produce an image tensor. The preprocessing steps include resizing the image to 480×480 pixels, converting it to a PyTorch tensor with pixel values scaled, and normalizing the tensor using ImageNet mean and standard deviation values for each channel.

The Refer API also provides a segmentation mask for the target object. Each mask is converted into a binary annotation, with the foreground (target object) labeled as 1 and the background as 0. The mask is then resized to match the image resolution. After preprocessing, each dataset sample consists of an image tensor, a segmentation mask, a sentence embedding, and an attention matrix.

4.3 Model Architecture

An overview of the proposed architecture is presented in Figure 4.3. In the proposed method, the model employs a general text embedding model and a hierarchical vision

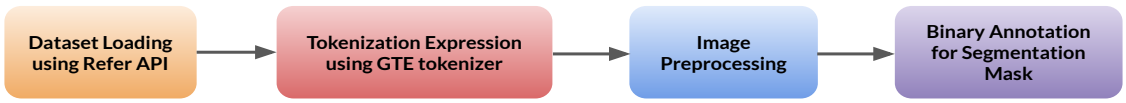

Figure 4.2: Data Preparation

transformer model to perform vision-language encoding. In each stage after feature extraction, using visual features V_i as queries and language features L_f as keys and values, scaled dot-product attention is applied to produce representations for language attention L_a. The resulting language attention L_a and vision features V_i are then fused through an adaptive gating mechanism A_g to produce multi-modal or cross-modal features. Subsequently, a learnable language gate L_g refines the feature and modulates the flow of linguistic information into the Swin Transformer, enabling controlled and effective multi-modal fusion within the Swin Encoder block. The decoder utilizes enhanced vision encoder features EV_i from each transformer block and decoder output F_i through full-scale or multi-scale connections to generate the final segmentation mask for the target object.

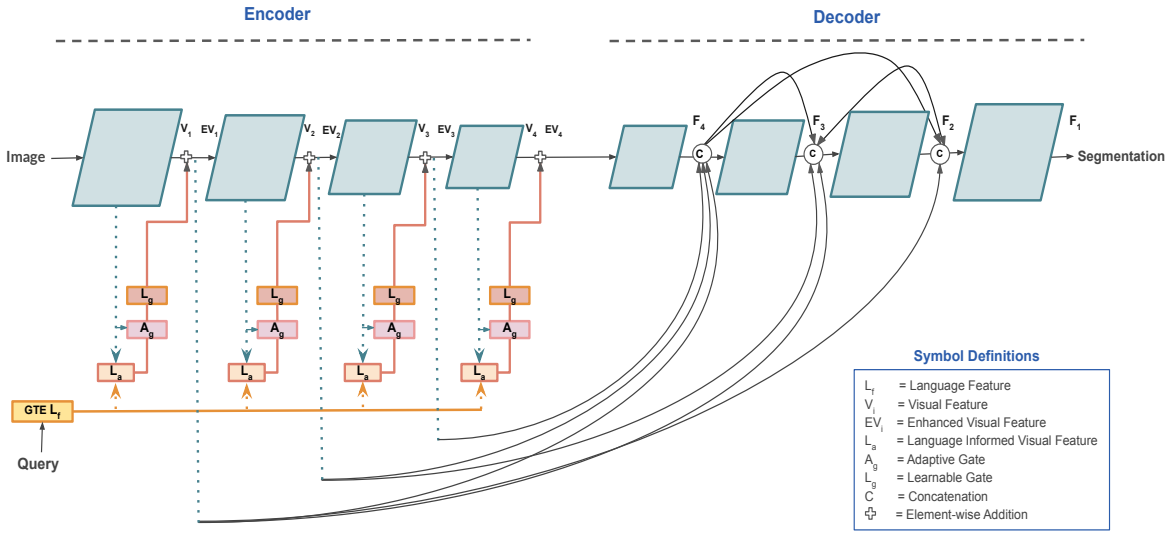

Figure 4.3: Architecture of the Proposed GAFFDNet Model.

The proposed model follows an encoder–decoder architecture. The encoder extracts multi-level visual features by progressively reducing spatial dimensions and capturing high-level semantics, while also integrating visual representations with linguistic features from the referring expression for effective Referring Image Segmentation.

4.3.1 Feature Extraction

To represent linguistic expressions in a high-dimensional semantic space, the proposed model employ the GTE base model [31] from the Hugging Face library. This multi-stage contrastive learning-based model generates highly discriminative and se-

mantically rich embeddings, improving the quality and generalizability of language features for downstream vision–language tasks such as Referring Image Segmentation (RIS). The resulting language feature matrix is represented as:$L_f \in \mathbb{R}^{C_t \times T}$, where $C_t = 768$ is the feature dimension and $T = 20$ is the fixed number of tokens in the input sequence. Each column in L_f corresponds to a token embedding from the referring expression, and each row represents one feature dimension. The token length is set to 20, as the maximum length of referring expressions in the RefCOCO, RefCOCO+, and G-Ref datasets is not more than 8.4 words.
For visual feature extraction, Swin Transformer encoder [26], model is adopted which employs a hierarchical architecture with shifted window attention. This design enables the model to effectively capture both local fine-grained details and global contextual information, making it particularly well-suited for dense prediction tasks such as segmentation. The encoder produces hierarchical visual feature maps across four stages. At each stage $i \in \{1, 2, 3, 4\}$, the output is represented as $V_i \in \mathbb{R}^{C_i \times H_i \times W_i}$, where C_i represents the number of channels (depth per spatial location), and H_i and W_i correspond to the spatial height and width of the feature map, capturing image details at different resolutions.As the network goes deeper, the spatial resolution decreases, while the feature representation becomes more semantic, encoding higher-level visual concepts. This hierarchical representation allows RIS models to leverage both fine object details and broader context for precise object localization and segmentation.

4.3.2 Cross-Modal Attention

To enable semantic alignment between visual and linguistic modalities, this step employs a cross-attention mechanism, which allows one modality to selectively attend to relevant information from another, thereby facilitating effective cross-modal fusion. For each spatial location in the visual input, scaled dot-product attention [8] is computed over the language tokens, forming a language attention matrix L_a as shown in equation (4.1). It transforms the input into three spaces: Queries (V_{iq}), Keys (L_{ik}), and Values (L_{iv}) using weight matrices. Then, it computes the dot product of V_{iq} and L_{ik} to measure similarity between tokens. The result is scaled by $\sqrt{C_i}$ to maintain a stable variance. Next, `softmax` is applied to convert similarity scores into probabilities, yielding the attention weights. Finally, these weights are used to take a weighted sum of the Values (L_{iv}), producing the context vector for each token.

$$L_a = \text{softmax}\left(\frac{V_{iq}^\top L_{ik}}{\sqrt{C_i}}\right) L_{iv}^\top \tag{4.1}$$

Here, the visual feature utilized as queries V_{iq}, while the language features used as keys L_{ik} and values L_{iv}. Visual features are projected through a convolution followed by normalization to stabilize spatial distributions and reduce image-induced variability. In contrast, language features are projected without normalization to preserve semantic diversity, as it could reduce essential distinctions for accurate cross-modal attention.
The resulting attention matrix L_a determines the influence of each word at each image region. The attended outputs are passed through an output projection consisting of a convolution and normalization to produce a refined feature map, which

is the same size as V_i.

4.3.3 Multi-Modal Adaptive Gate

To adaptively integrate visual features V_i with attended language features L_a a learnable gating mechanism is proposed as shown in equations (4.2), (4.3), and (4.4) that dynamically fuses two modalities at each spatial location. This enables the model to determine the relative importance of visual and linguistic cues based on input context.

The obtained visual and language features are combined via element-wise addition to produce a joint representation of the two modalities. This representation is passed through a convolution projection layer denoted as α_i, followed by a sigmoid activation for nonlinearity to produce the adaptive gate A_g. The gating map A_g controls the contribution of each modality per channel and spatial position. The multi-modal feature MM_i is computed by element-wise weighting of the modalities, where the visual features V_i are scaled by the adaptive gate A_g and the language features L_a are scaled by $(1-A_g)$. The two weighted components are then combined through element-wise addition to yield the fused multi-modal representation.

$$\alpha_i = \sigma\left(\text{Conv}(x_i)\right) \tag{4.2}$$

$$A_g = \alpha_i(V_i + L_a) \tag{4.3}$$

$$MM_i = A_g \odot V_i + (1 - A_g) \odot L_a \tag{4.4}$$

Here, $\odot$ denotes element-wise multiplication. A gate value near 1 emphasizes visual features, while a value near 0 favors language. When the gate value is ~ 0.5, both modalities contribute equally, allowing the model to balance visual and linguistic information. This mechanism enables the network to adaptively emphasize modality-specific or jointly informative features based on spatial context.

4.3.4 Learnable Gate

The learnable gating unit assigns weights to each element of the multi-modal feature MM_i and subsequently integrates it with Swin Transformer encoder output V_i through element-wise operations, resulting in enhanced visual representations EV_i enriched with linguistic context. This mechanism is referred to as the *language pathway* in [30], as it enables the Swin Transformer encoder to incorporate contextual language cues into its visual encoding process.

The learnable gating formulation, presented in Equation (4.7), consists of two convolutional blocks, β_i and γ_i, as defined in Equations (4.5) and (4.6). Together, they form a two-layer perceptron with a ReLU activation followed by a tanh nonlinearity, which refines the multi-modal feature MM_i. Subsequently, a residual-style fusion, as shown in Equation (4.8), is employed to preserve the visual representations while effectively integrating multi-modal information.

$$\beta_i = \text{ReLU}\left(\text{Conv}(x_i)\right) \tag{4.5}$$

$$\gamma_i = \tanh\left(\text{Conv}(\beta_i)\right) \tag{4.6}$$

$$L_g = \gamma_i(MM_i) \tag{4.7}$$

$$EV_i = L_g \odot MM_i + V_i \tag{4.8}$$

4.3.5 Decoder Mask Generation

The decoder is designed to hierarchically integrate enhanced visual features EV_i from the Swin Transformer encoder to generate segmentation masks guided by referring expressions. Inspired by the full-scale connection from [17] architecture, it combines both spatially rich encoder features and semantically strong decoder features at each stage, enabling fine-grained object localization conditioned on language. The formulation of the decoder block is shown in equation (4.9).

$$\begin{cases} F_4 = \mathrm{EV}_4 \\ F_i = \mathcal{P}_d\Big(\Big[\mathcal{P}_s(\mathrm{EV}_i);\ \mathcal{P}_s\big(\downarrow(\mathrm{EV}_1, \ldots, \mathrm{EV}_{i-1})\big); \\ \qquad\qquad \mathcal{P}_s\big(\uparrow(F_{i+1}, \ldots, F_4)\big)\Big]\Big), \quad i \in \{3, 2, 1\} \end{cases} \tag{4.9}$$

Here EV_i refers to the enhanced visual feature enriched with linguistic context from the encoder and F_i denotes the fused feature for each decoder stage i, respectively. $\mathcal{P}_s(\cdot)$ and $\mathcal{P}_d(\cdot)$ denote single-layer and double-layer convolutional projection functions. Each $\mathcal{P}_s$ or $\mathcal{P}_d$ projection function is implemented as a convolution block followed by a normalization and LeakyReLU as the activation function. The operators *downarrow* and *uparrow* represent downsampling and upsampling of feature maps, and ; indicates the feature concatenation operation.
At stage $i = 4$, the decoder block is initialized by simply using the encoder output EV_4. For stages $i \in \{3, 2, 1\}$, three categories of features are aggregated: the encoder feature EV_i at the current stage, all shallower encoder features EV_1 to EV_{i-1} after downsampling and projection $\mathcal{P}_s$, and all deeper decoder outputs F_{i+1} to F_4 after upsampling and projection $\mathcal{P}_s$. These are concatenated and passed through $\mathcal{P}_d$ to produce the fused output F_i. The final output from the last decoder stage, F_1, is passed through a 1×1 convolution layer to generate the final binary segmentation mask corresponding to the referring expression.
Downsampling is the process of reducing the spatial resolution of a feature map or image. In neural networks, this is often done to extract more abstract, high-level features while reducing computational cost. For downsampling average pooling is applied to reduce its size while retaining the most important features. Upsampling is the process of increasing the spatial resolution of a feature map or image to recover spatial details lost during downsampling. For upsampling bilinear interpolation is applied. It is used in the decoder to restore the feature map to the original image size so that a precise pixel-level segmentation mask can be generated.
The details are described below:

Decoder Stage 4

At stage $i = 4$, the decoder output is defined as $y_4 = \mathrm{EV}_4$, since it is the deepest encoder feature containing the strongest semantic context, there are no deeper decoder outputs to aggregate, and this provides a clean semantic initialization that will be progressively refined using shallower encoder features in later stages.

Decoder Stage 3

In the existing architecture (LAVT), decoder stage 3 upsamples the previous decoder output y_4 to match the spatial resolution of x_3, concatenates it with x_3, and refines the result with a double-layer projection $\mathcal{P}_d$ to obtain the output for stage 3.

In contrast, as shown in 4.4 proposed architecture aggregates three types of encoder features: x_1, x_2, and x_3. The shallower encoder features (x_1, x_2) are downsampled via max pooling, while the previous decoder output (y_4) is upsampled using bilinear interpolation. Before concatenation, each feature is refined with a single-layer projection $\mathcal{P}_s$. The concatenated features are then processed with a double-layer projection $\mathcal{P}_d$, producing the output for decoder stage 3.

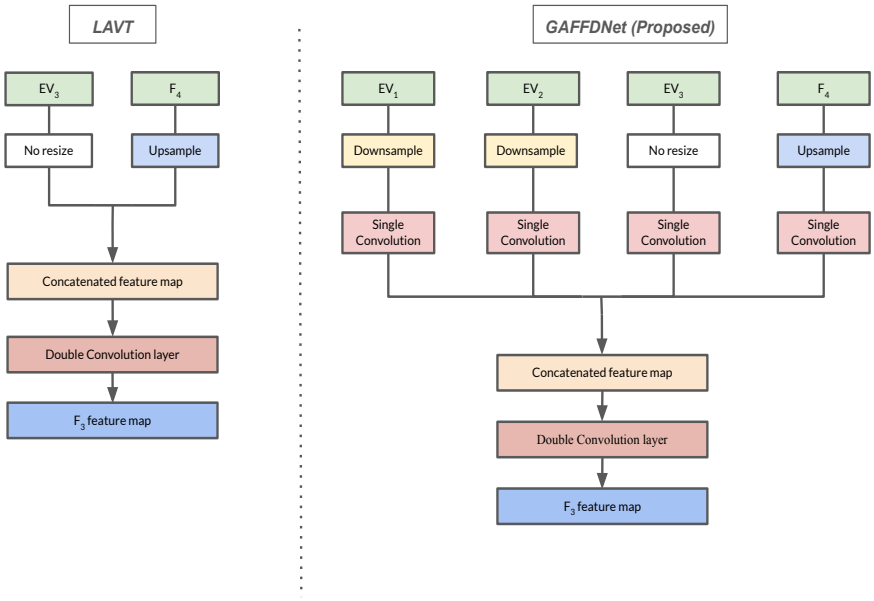

Figure 4.4: Architectural differences at decoder Y_3 stage.

Decoder Stage 2

In the existing architecture (LAVT), decoder stage 2 upsamples the previous decoder output y_3 to match the spatial resolution of x_2, concatenates it with x_2, and refines the result with a double-layer projection $\mathcal{P}_d$ to obtain the output for stage 2.
In contrast, as shown in 4.5 proposed architecture aggregates two types of encoder features: x_1, and x_2. The shallower encoder features (x_1 is downsampled via max pooling, while the previous decoder output (y_3) and (y_4) are upsampled using bilinear interpolation. Before concatenation, each feature is refined with a single-layer projection $\mathcal{P}_s$. The concatenated features are then processed with a double-layer projection $\mathcal{P}_d$, producing the output for decoder stage 2.

Decoder Stage 1

In the existing architecture (LAVT), decoder stage 1 upsamples the previous decoder output y_2 to match the spatial resolution of x_1, concatenates it with x_1, and refines the result with a double-layer projection $\mathcal{P}_d$ to obtain the output for stage 1.
In contrast, as shown in 4.6 proposed architecture uses x_1 feature from endoer and in this step there is no shallower encoder features, and the previous decoder output (y_2), (y_3) and (y_4) are upsampled using bilinear interpolation. Before concatenation, each feature is refined with a single-layer projection $\mathcal{P}_s$. The concatenated features

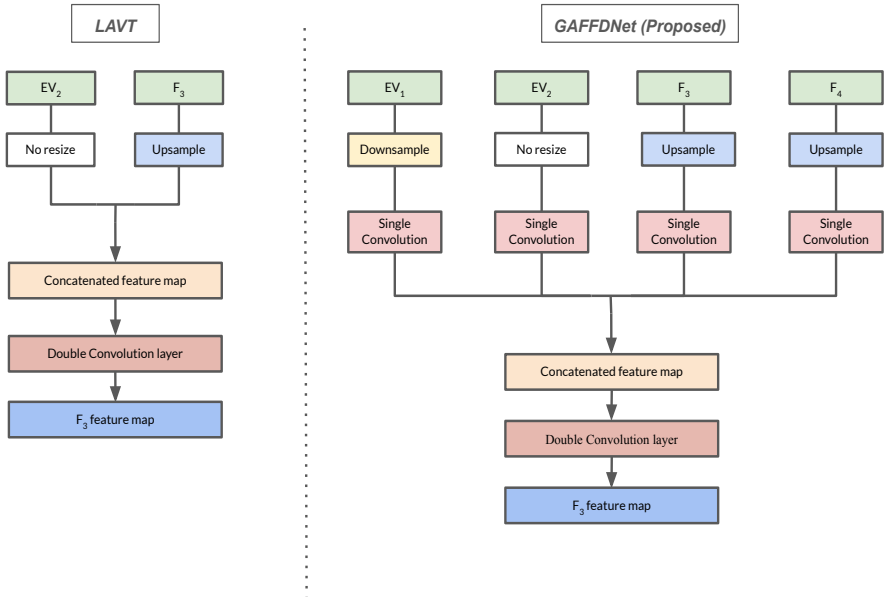

Figure 4.5: Architectural Differences at Decoder Y_2 Stage.

are then processed with a double-layer projection $\mathcal{P}_d$, producing the output for decoder stage 1.
After obtaining the final decoder output F_1, it is passed through a 1×1 convolution layer to generate the final binary segmentation mask corresponding to the referring expression.

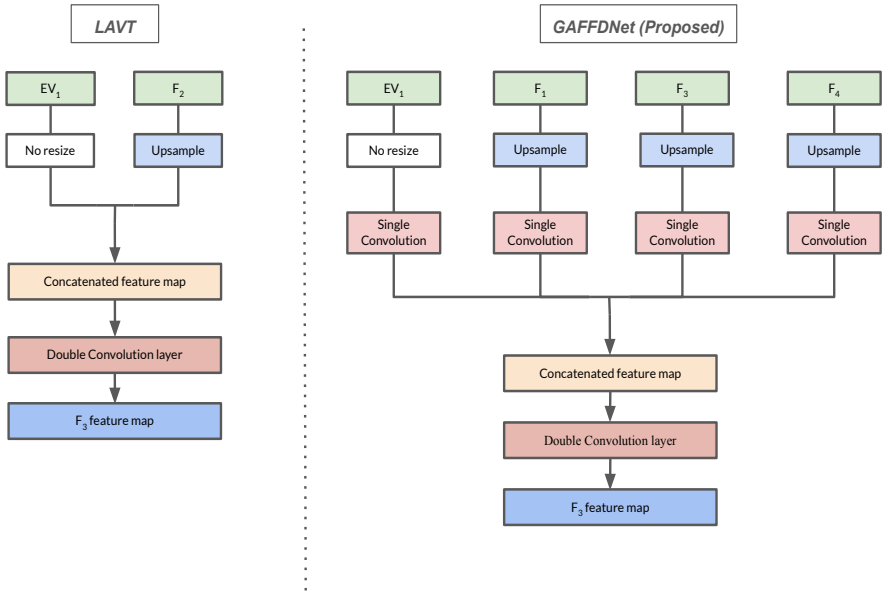

Figure 4.6: Architectural Differences at Decoder Y_1 Stage.

Chapter 5

Performance Evaluation

This section will detail the implementation specifics of the proposed Referring Image Segmentation (RIS) model, including training hyperparameters, inference procedures, and evaluation metrics. Additionally, the results compared the proposed method against state-of-the-art approaches on benchmark datasets will be presented.

5.1 Implementation

The proposed method is implemented in PyTorch [13] and uses the GTE implementation from HuggingFace's Transformer library. Swin Transformer [26] layers are initialized with classification weights pre-trained on ImageNet-22K [1]. All the images are resized to 480×480, and no data augmentation techniques are applied because RIS heavily depends on accurate spatial-text alignment. Augmentations risk introducing mismatches between visual features, segmentation masks, and referring expressions. Additionally, keeping the data unaltered ensures a fair benchmark comparison and leverages the robustness of the pretrained Swin Transformer. The model is trained on a single NVIDIA RTX 4090 GPU with 24GB memory. The training hyperparameters used for the implementation are summarized in Table 5.1.

Hyperparameter	**Value**
Epochs	40
Effective Batch Size	32
Optimizer	AdamW
Weight decay	0.01
Learning Rate	5×10^{-5}
Loss Function	Weighted Binary Cross-Entropy
Precision	Mixed precision (fp16)

Table 5.1: Hyperparameters

These hyperparameters are carefully chosen to balance stability, memory efficiency, and precise mask prediction, which are critical for the proposed model.
The model is trained for 40 epochs as it provided the best balance between performance and computational cost, with no significant improvements beyond this point. RIS models are computationally expensive as they jointly process visual and linguistic features. Training for too few epochs can lead to underfitting, where the model fails to capture complex relationships between text and images. Conversely, training for too many epochs increases the risk of overfitting, where the model memorizes

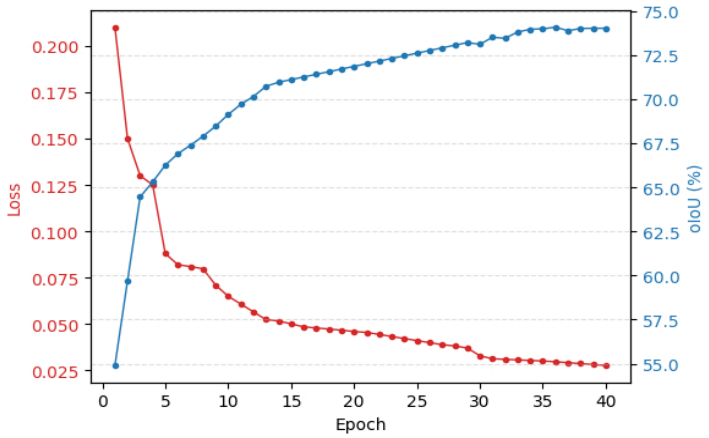

Figure 5.1: Loss vs oIoU Curve on RefCOCO Validation Set.

the training data and performs poorly on unseen samples.The effective batch size for training is set to 32, achieved by using a base batch size of 8 with gradient accumulation over 4 steps. Directly using a batch size of 32 would require a GPU with very large memory, to overcome this limitation, gradient accumulation is applied, where the model processes 8 samples at a time and stores the computed gradients without immediately updating the model weights. This process is repeated four times, and after processing a total of 32 samples, the accumulated gradients are averaged, and a single optimization step is performed. This way it effectively simulates training with a batch size of 32 while staying within the memory constraints of the hardware. With 40 epochs and the complete training process takes approximately 4 days to finish.
Figure 5.1 illustrates the loss and validation curve of the proposed model over 40 epochs on RefCOCO validation set. As training progresses, the loss curve steadily decreases, indicating that the model has learned effectively and minimized prediction errors. Simultaneously, the oIoU accuracy increases, demonstrating improved segmentation accuracy and better alignment between the predicted and ground-truth masks. The inverse relationship between the two curves validates stable convergence as the loss gets smaller, oIoU consistently rises until it reaches around 74.07%, showing that the model achieves strong generalization performance on the validation set.

The model optimization process relies on the AdamW optimizer, which combines the benefits of adaptive learning rates with decoupled weight decay for improved regularization. This choice of optimizer allows the model to converge efficiently while preventing overfitting. To further stabilize training, a learning rate scheduler is applied, which gradually decreases the learning rate following a decay function. This scheduling strategy enables the model to make larger updates during the initial stages of training and progressively smaller refinements as it approaches convergence, ensuring both stability and accuracy in the final learned parameters.
The model uses a weight decay [7] value of 0.01 as a regularization technique to prevent overfitting. Weight decay works by adding a small penalty to the model's loss function, proportional to the magnitude of the weights. This discourages the model from relying too heavily on specific features by keeping the weights small and

well-balanced. As a result, the model learns more generalizable patterns, improving its performance on unseen data. This is particularly important for Referring Image Segmentation (RIS), where the model must process diverse visual and linguistic inputs. A weight decay of 0.01 provides a good balance between effective regularization and maintaining model capacity.
The learning rate is set to 5e-5, which determines the step size for updating model parameters during optimization. Choosing an appropriate learning rate is crucial: too high a value may cause the model to overshoot the optimum, leading to unstable or divergent training, while too low a value results in slow convergence and longer training times. To balance these concerns, a polynomial decay scheduler is applied, which gradually reduces the learning rate from its initial value to near zero over the training process. This approach allows fast learning in the early stages and stable fine-tuning in later stages, preventing overshooting and ensuring smooth convergence. Such a strategy is particularly important for Referring Image Segmentation (RIS) models, where the model must capture complex relationships between natural language descriptions and image regions. Overall, the polynomial decay with an exponent of 0.9 ensures efficient optimization, improved convergence, and better final performance.
For the model optimization a weighted binary cross-entropy (WBCE) loss function is employed, which balances the relative importance of foreground and background classes. This is particularly beneficial for referring image segmentation tasks, where the target object often occupies only a small fraction of the image compared to the background. Without weighting, the model tends to bias toward predicting background regions, leading to poor localization of small or fine-grained objects. To mitigate this imbalance, a slightly higher weight is assigned to the foreground (object) class, ensuring that the model pays greater attention to precise boundary details. By balancing the contributions of background and foreground pixels, the weighted loss encourages the network to produce more accurate segmentation masks, especially in challenging cases where the object is small or complex.
The model is trained using mixed precision (fp16), where most computations are done in 16-bit format to reduce memory usage and speed up training. This allows efficient training of complex RIS models without compromising accuracy.

5.2 Inference Process

During inference, the trained model receives as input an image and a corresponding referring expression. The pipeline consists of the following steps as illustrated in Figure 5.2 and details are described below:

Figure 5.2: Model Inference Process.

Each input image is first loaded and resized to 480 × 480 pixels. The image is then converted into a normalized tensor representation using ImageNet mean and

standard deviation. The referring expression is tokenized using GTE tokenizer from HuggingFace. The token sequence is truncated to a maximum length of 20 tokens, padded with zeros if necessary, and converted into a tensor. An attention mask is generated to distinguish between valid tokens and padding tokens. The proposed segmentation model is instantiated, and its parameters are initialized with weights from the best-performing checkpoint. The model uses CPU device for inference. The image tensor, tokenized sentence, and attention mask are passed through the model. The model outputs a teson. A class-wise arg max operation is applied to produce a binary segmentation mask. This mask is then resized back to the original image resolution using bilinear interpolation. The predicted binary mask is overlaid on the original RGB image to visualize the segmented object. This is achieved by applying a semi-transparent color overlay and contour highlighting on the segmented region. The final visualization is stored in the output directory for inspection. In summary, the inference pipeline begins with multimodal preprocessing, followed by model prediction, and concludes with visualization of the segmentation mask over the original image. This procedure is repeated for each test image and its corresponding referring expression.

5.3 Evaluation Metrics

To assess the performance of the proposed model, three standard metrics are utilized: Overall IoU (OIoU), Mean IoU (mIoU), and Precision@X.
Intersection over Union (IoU) is a common evaluation metric for segmentation tasks, measuring the overlap between the predicted segmentation mask and the ground truth mask. It is calculated as the ratio of the area of intersection to the area of union between the predicted and ground truth masks as defined in Equation (5.1) as follows:

$$IoU_i = \frac{|P_i \cap G_i|}{|P_i \cup G_i|} = \frac{TP}{TP + FP + FN} \tag{5.1}$$

where P_i and G_i denote the predicted masks and ground truth masks, respectively. TP represents the number of pixels correctly predicted, FP is the number of pixels incorrectly predicted, and FN is the number of actual object pixels that were missed by the model.
IoU focuses solely on the positive prediction class, which is crucial in segmentation tasks. This is because the background class often occupies a much larger area. Including True Negatives (TN) in the calculation could lead to artificially inflated accuracy, even if the model performs poorly at detecting the target object. For this reason, recall is not a reliable metric for evaluating the accuracy of Referring Image Segmentation models.
Intersection over Union (IoU) measures the overlap between a predicted segmentation mask and the corresponding ground truth. Overall Intersection over Union (oIoU) measures the total overlap between all predicted masks and ground truth masks across the dataset. It is calculated by dividing the sum of the intersection areas of all predictions and ground truths by the sum of their union areas. This metric provides a global view of the model's segmentation performance, giving higher

weight to larger objects or more frequent classes. Equation (5.2) defines oIoU:

$$\text{oIoU} = \frac{\sum_{i=1}^{N} |P_i \cap G_i|}{\sum_{i=1}^{N} |P_i \cup G_i|} \tag{5.2}$$

Mean Intersection over Union (mIoU) calculates the average IoU for each individual sample in the dataset. For each predicted mask, IoU is computed with its corresponding ground truth, and then these values are averaged. Unlike oIoU, mIoU treats all samples equally, regardless of their size, providing a more balanced assessment of the model's performance across diverse objects and images. It is defined in Equation (5.3), where N is the total number of samples.

$$\text{mIoU} = \frac{1}{N} \sum_{i=1}^{N} \frac{|P_i \cap G_i|}{|P_i \cup G_i|} \tag{5.3}$$

Precision@X measures the percentage of samples whose IoU exceeds a certain threshold X, where $X \in \{0.5, 0.7, 0.9\}$. In other way, it counts how many samples achieve IoU above a threshold like 0.5, 0.7, or 0.9. For example, Precision@0.5 counts the fraction of predicted masks that overlap more than 50% with the ground truth. This metric is computed per sample, similar to mIoU, but instead of averaging IoU values, it checks whether each sample's IoU passes the threshold as defined in equation (5.4), providing insight into how often the model produces highly accurate segmentations.

$$\text{Precision@X} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(IoU_i > X), \quad \mathbf{1} = \begin{cases} 1, & \text{if true} \\ 0, & \text{otherwise} \end{cases} \tag{5.4}$$

5.4 Result Analysis

Table 5.2 presents the performance of the proposed GAFFDNet on three widely used referring image segmentation datasets: RefCOCO, RefCOCO+, and G-Ref. The evaluation metrics include Overall IoU, Mean IoU, and Precision at thresholds 0.5, 0.7, and 0.9.

GAFFDNet demonstrates strong performance across all datasets, achieving the highest Overall and Mean IoU values, which indicate accurate pixel-level segmentation of the referred objects. Precision@0.5 and Precision@0.7 metrics show that the model reliably segments most objects with good overlap to the ground truth, while Precision@0.9 highlights the model's performance for highly stringent segmentation thresholds, which is naturally lower due to the difficulty of perfectly matching object masks.

The results indicate that GAFFDNet effectively handles diverse referring expressions and challenging visual scenes, outperforming previous methods in terms of both accuracy and robustness across multiple splits (validation and test sets). The consistent improvement across all datasets validates the effectiveness of the proposed enhancements in language encoding, adaptive multimodal feature fusion, and cross-scale decoder aggregation.

In this section, visualization masks of the proposed model's performance is presented. Specifically, examples from the RefCOCO, RefCOCO+, and G-Ref datasets are illustrated to visually compare the segmentation results of the proposed GAFFDNet model with the existing LAVT model.

Metrics	RefCOCO			RefCOCO+			G-Ref	
	val	test A	test B	val	test A	test B	val (U)	test (U)
Overall IoU	74.07	76.93	70.44	64.69	69.31	56.27	64.05	64.58
Mean IoU	75.63	78.03	72.90	67.65	71.93	60.36	66.30	66.85
Precision@0.5	85.94	89.20	82.06	76.20	82.05	67.00	75.00	75.43
Precision@0.7	77.60	81.92	71.80	67.92	74.17	57.37	63.66	64.79
Precision@0.9	36.49	36.93	36.96	30.76	32.73	27.69	27.86	28.74

Table 5.2: GAFFDNet Achieved Results on RefCOCO, RefCOCO+, and G-Ref Datasets on All Evaluation Metrics.

In Figure 5.3, visualization examples from the RefCOCO, RefCOCO+, and G-Ref test sets are presented, showcasing the ground-truth masks alongside the predicted masks. The tests of the proposed GAFFDNet model is compared with the existing LAVT model [30]. Given an image and a natural language query, the segmentation masks predicted by GAFFDNet demonstrate more accurate localization of the target object. In contrast, the LAVT model often fails to segment the target mask precisely, highlighting the effectiveness of the proposed approach.

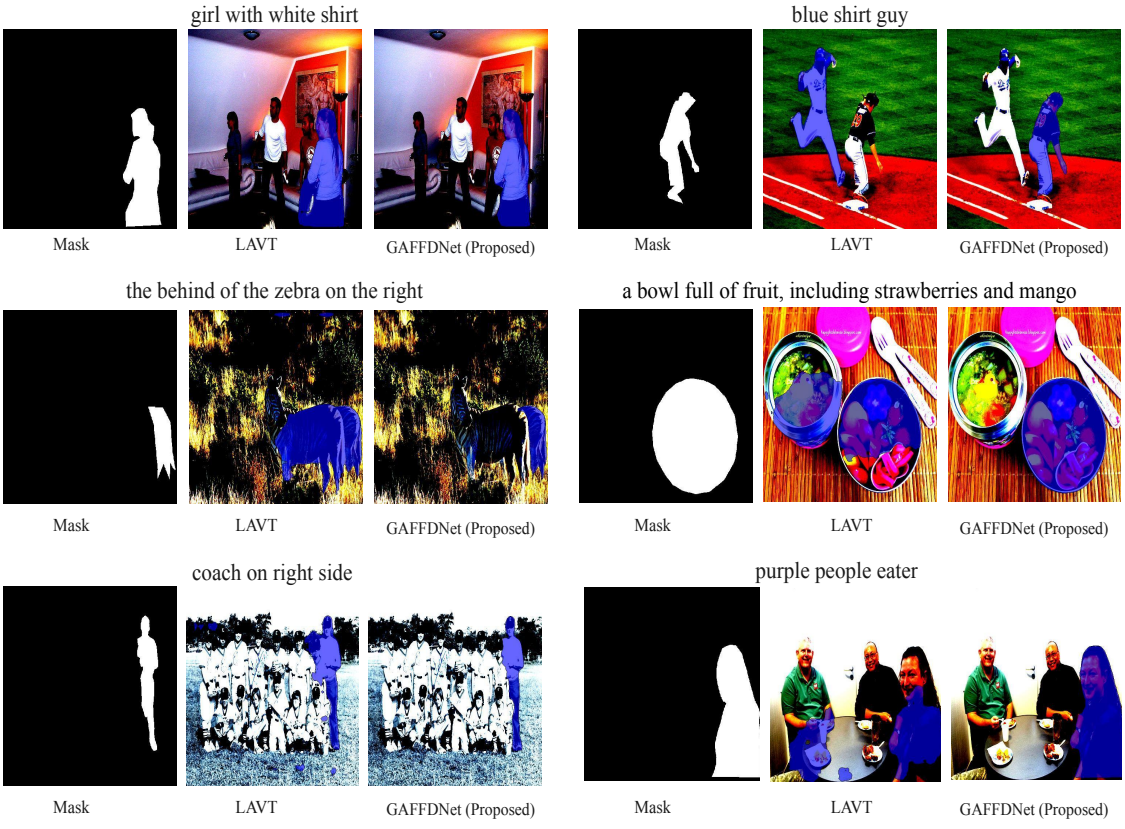

Figure 5.3: Visualization of GAFFDNet and LAVT predicted masks on the RefCOCO, RefCOCO+, and G-Ref test sets.

Furthermore the analysis is extended in Figure 5.4, which illustrates cases where GAFFDNet performs almost as accurately as the ground-truth mask, while the LAVT model segments an entirely incorrect region. Although GAFFDNet occasion-

ally misses very small portions of the target object, the majority of its predictions are correct. Importantly, unlike LAVT, it does not produce completely erroneous segmentations. These observations not only validate the robustness of GAFFDNet but also provide insights into its limitations, which can guide directions for future research.

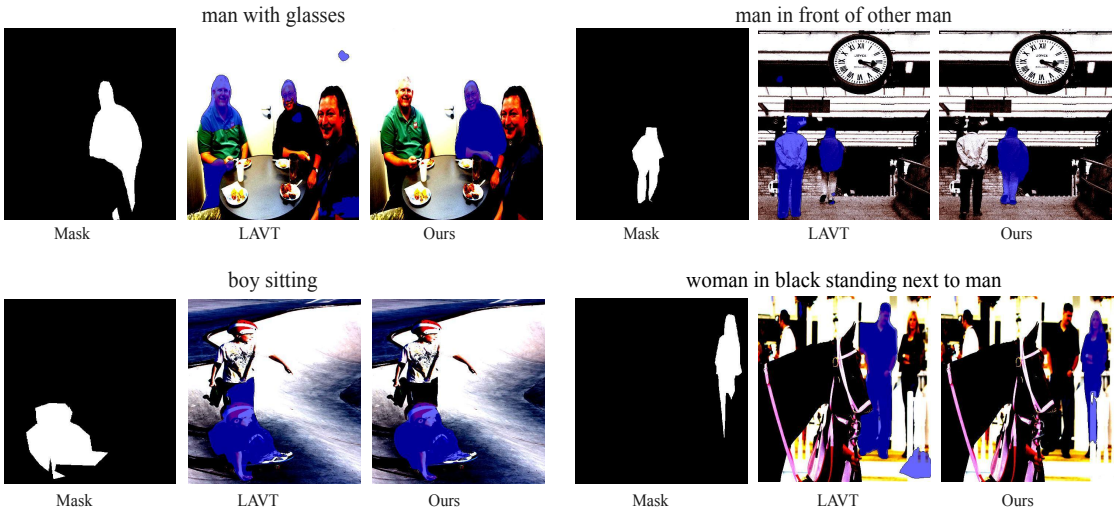

Figure 5.4: Visualization Comparison of GAFFDNet Missed Small Regions

5.5 Ablation Studies

Several ablation experiments were conducted to evaluate the contributions of the key components in the proposed GAFFDNet network. It was observed that each component—the GTE language model, the adaptive gating mechanism, and the full-scale decoder connections—significantly contributes to the overall performance. The results highlight that rich language embeddings, dynamic multimodal feature fusion, and cross-scale feature aggregation are essential for achieving state-of-the-art referring image segmentation accuracy.

Multi-modal adpative gate: The proposed gate was experienced with reverse scaling order $MM_i = A_g \odot L_a + (1 - A_g) \odot V_i$, which results in degraded performance. This observation confirms that the proposed gating formulation more effectively balances the contributions of visual and language-informed visual features during fusion. The sigmoid activation in the proposed adaptive gate is responsible for learning a set of spatial weight maps that dynamically control the contribution of multi-modal information. Alternative nonlinear activation functions, such as *tanh*, were also experimented with, which resulted in degraded performance.

Component-based experiment: To evaluate the contribution of each component in the proposed architecture, a series of ablation studies was performed. In each experiment, a specific module was removed while keeping the rest of the network unchanged, enabling quantification of its impact on segmentation performance.

As highlighted in Table 5.3 (a), the choice of language model plays a vital role. Keeping the adaptive gate A_g and full-scale decoder connections intact, the GTE

Components	P@0.5	P@0.7	P@0.9	oIoU	mIoU
(a) Language Model L_f					
SigLIP	84.48	75.53	34.65	72.85	74.55
CLIP	84.51	75.88	35.52	72.98	75.12
BERT	85.63	76.50	35.97	73.10	75.34
GTE	**85.94**	**77.60**	**36.49**	**74.07**	**75.63**
(b) Adaptive Gate A_g					
Without Gate	85.49	76.15	35.39	73.17	75.05
With Gate	**85.94**	**77.60**	**36.49**	**74.07**	**75.63**
(c) Decoder Connections F_i					
Plain Connections	85.00	76.08	35.20	73.12	74.77
Full Scale Connections	**85.94**	**77.60**	**36.49**	**74.07**	**75.63**

Table 5.3: Ablation Studies on RefCOCO Validation Set.

language model is replaced by SigLIP, CLIP, and BERT. Using SigLIP led to performance drops of 1.22 in overall IoU and 1.08 in mean IoU. CLIP led to performance drops of 1.09 in overall IoU and 0.51 in mean IoU. BERT also resulted in reductions of 0.97 and 0.29 in overall and mean IoU, respectively, along with declines in all Precision@X metrics.

To assess the impact of the adaptive gate, the GTE language model and full-scale decoder structure were fixed and compared performance with and without A_g as shown in Table 5.3 (b). Removing the gate resulted in drops of 0.45 in overall IoU and 0.58 in mean IoU, along with declines in all Precision@X metrics.

The effectiveness of the proposed full-scale skip connections by comparing them against plain decoder connections, while keeping both GTE and A_g fixed as shown in Table 5.3 (c). Replacing full-scale connections with plain ones resulted in drops of 0.94 in overall IoU and 1.52 in mean IoU, along with declines in all Precision@X metrics.

5.6 Discussion

In this section, the results are discussed in relation to the research objectives, including the enhancement of language embeddings, adaptive multimodal feature fusion, cross-scale feature aggregation, and comprehensive evaluation on benchmark datasets.

Enhanced language embeddings: The proposed GAFFDNet model leverages the GTE language encoder to generate rich textual embeddings that effectively capture the semantics of referring expressions. Compared to prior methods using language encoders such as LSTM or CLIP, GTE provides a more nuanced understanding of complex phrases. This is reflected in the improved segmentation accuracy across all benchmark datasets, demonstrating that richer language representations lead to more precise localization of referred objects. Furthermore, by applying contrastive learning objectives, the model can effectively distinguish between semantically similar and dissimilar expressions, enhancing cross-modal alignment with visual features.

Adaptive multimodal feature fusion: The proposed adaptive gate mechanism dynamically balances the contributions of visual features and language informed visual features during multi-modal fusion. This allows the model to selectively emphasize relevant information based on the context of the referring expression.

Cross-scale feature aggregation: The full-scale skip connections in the decoder enable effective integration of multi-level features from different stages of the backbone network. This hierarchical feature aggregation captures both fine-grained details and high-level semantics, which is crucial for accurately segmenting objects of varying sizes and complexities.

Evaluation on benchmark datasets: The performance of the GAFFDNet model was evaluated in terms of overall IoU, mean IoU, and Precision@X metrics across multiple benchmark datasets. In Table 5.4, the proposed model GAFFDNet is compared with state-of-the-art referring image segmentation methods across three benchmark datasets using the oIoU metric. GAFFDNet consistently outperforms all prior approaches across all evaluation subsets. The proposed approach builds upon the work of [30] and is directly compared against it on the RefCOCO dataset [5]. GAFFDNet achieves absolute improvements of 1.34%, 1.11%, and 1.65% over LAVT on the validation, testA, and testB subsets, respectively. For RefCOCO+ [5], GAFFDNet achieves further improvements with margins of 2.55%, 0.93%, and 1.17% across the same splits. Notably, on the more challenging G-Ref dataset (Google split) [4], GAFFDNet achieves significant gains of 2.51% and 2.49% on the validation and test subsets respectively.

Method	**Language Encoder**	**RefCOCO**			**RefCOCO+**			**G-Ref**	
		val	test A	test B	val	test A	test B	val (U)	test (U)
RRN [9]	LSTM	55.33	57.26	53.93	39.75	42.15	36.11	-	-
BRINet [15]	LSTM	60.98	62.99	59.21	48.17	52.32	42.11	-	-
CMPC [18]	LSTM	61.36	64.53	59.64	49.56	53.44	43.23	-	-
LSCM [19]	LSTM	61.47	64.99	59.55	49.34	53.12	43.50	-	-
CMPC+ [25]	LSTM	62.47	65.08	60.82	50.25	54.04	43.47	-	-
MAttNet [10]	Bi-LSTM	56.51	62.37	51.70	46.67	52.39	40.08	47.64	48.61
CAC [12]	Bi-LSTM	58.90	61.77	53.81	-	-	-	46.37	46.95
STEP [11]	Bi-LSTM	60.04	63.46	57.97	48.19	52.33	40.41	-	-
MCN [21]	Bi-GRU	62.44	64.20	59.71	50.62	54.99	44.69	49.22	49.40
EFN [23]	Bi-GRU	62.76	65.69	59.67	51.50	55.24	43.01	-	-
CGAN [20]	Bi-GRU	64.86	68.04	62.07	51.03	55.51	44.06	51.01	51.69
LTS [24]	Bi-GRU	65.43	67.76	63.08	54.21	58.32	48.02	54.40	54.25
VLT [22]	Bi-GRU	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65
BUSNet [27]	Self-Att	63.27	66.41	61.39	51.76	56.87	44.13	-	-
ReSTR [28]	GloVe	67.22	69.30	64.45	55.78	60.44	48.27	-	-
CRIS [29]	Transformer	70.47	73.18	66.10	62.27	68.08	53.68	59.87	60.36
ETRIS [33]	CLIP	70.51	73.51	66.63	60.10	66.89	50.17	59.82	59.91
CM-MaskSD [32]	CLIP	72.18	75.21	67.91	64.47	69.29	56.55	62.67	62.69
VPD [34]	CLIP	73.25	-	-	62.69	-	-	61.96	-
LAVT [30]	BERT	72.73	75.82	68.79	62.14	68.38	55.10	61.24	62.09
GAFFDNet (proposed)	**GTE**	**74.07**	**76.93**	**70.44**	**64.69**	**69.31**	**56.27**	**64.05**	**64.58**

Table 5.4: Comparison with Competitive Methods in Terms of Overall IoU on Three Benchmark Datasets.

In Table 5.5, the proposed model GAFFDNet is compared with the previous best method, LAVT [30], on the RefCOCO validation set using the mIoU and Precision@X metrics. GAFFDNet achieves an absolute improvement of 1.17% in mIoU, along with gains of 1.47%, 2.32%, and 2.19% in Precision@0.5, @0.7, and @0.9, respectively. In addition, the proposed model is evaluated on the RefCOCO testA and testB sets, where GAFFDNet consistently outperforms LAVT by margins of 1.14% and 1.96% in mIoU, respectively, along with corresponding improvements in Precision@X metrics.

Dataset (RefCOCO)	P@0.5	P@0.7	P@0.9	mIoU
Validation set				
LAVT [30]	84.46	75.28	34.30	74.46
GAFFDNet	**85.94**	**77.60**	**36.49**	**75.63**
TestA set				
LAVT [30]	88.07	79.90	35.69	76.89
GAFFDNet	**89.20**	**81.92**	**36.93**	**78.03**
TestB set				
LAVT [30]	79.12	69.17	34.45	70.94
GAFFDNet	**82.06**	**71.80**	**36.96**	**72.90**

Table 5.5: Precision@X and mIoU Comparison on RefCOCO Dataset.

In Table 5.6, proposed model GAFFDNet is compared with the previous best method, LAVT [30], on the RefCOCO+ validation set using the mIoU and Precision@X metrics. GAFFDNet achieves an absolute improvement of 1.84% in mIoU, along with gains of 1.76%, 2.34%, and 0.53% in Precision@0.5, @0.7, and @0.9, respectively. Further the proposed model is evaluated on the RefCOCO+ testA and testB sets, where GAFFDNet consistently outperforms LAVT by margins of 0.96% and 1.13% in mIoU, respectively, along with corresponding improvements in Precision@X metrics.

Dataset (RefCOCO+)	P@0.5	P@0.7	P@0.9	mIoU
Validation set				
LAVT [30]	74.44	65.58	30.23	65.81
GAFFDNet	**76.20**	**67.92**	**30.76**	**67.65**
TestA set				
LAVT [30]	80.68	72.90	32.36	70.97
GAFFDNet	**82.05**	**74.17**	**32.73**	**71.93**
TestB set				
LAVT [30]	65.66	55.94	27.24	59.23
GAFFDNet	**67.00**	**57.37**	**27.69**	**60.36**

Table 5.6: Precision@X and mIoU Comparison on RefCOCO+ Dataset.

In Table 5.7, the proposed GAFFDNet model is compared with the previous best method, LAVT [30], on the G-Ref validation set using the mIoU and Precision@X metrics. GAFFDNet achieves an absolute improvement of 2.96% in mIoU, along with gains of 4.19%, 5.06%, and 5.13% in Precision@0.5, @0.7, and @0.9, respectively. Further the proposed model is evaluated on the G-Ref test set, where GAFFDNet consistently outperforms LAVT by a margin of 3.23% in mIoU, along with corresponding improvements in Precision@X metrics.

Dataset (G-Ref)	P@0.5	P@0.7	P@0.9	mIoU
Validation set				
LAVT [30]	70.81	58.60	22.73	63.34
GAFFDNet	**75.00**	**63.66**	**27.86**	**66.30**
Test set				
LAVT [30]	71.54	59.00	23.10	63.62
GAFFDNet	**75.43**	**64.79**	**28.74**	**66.85**

Table 5.7: Precision@X and mIoU Comparison on G-Ref Dataset.

Overall, these results demonstrate that GAFFDNet consistently surpasses the previous state-of-the-art method, LAVT, across multiple benchmark datasets (RefCOCO, RefCOCO+, and G-Ref) and evaluation metrics (oIoU, mIoU and Precision@X). The consistent improvements across both validation and test splits, indicate that GAFFDNet is more effective at accurately localizing and segmenting target objects

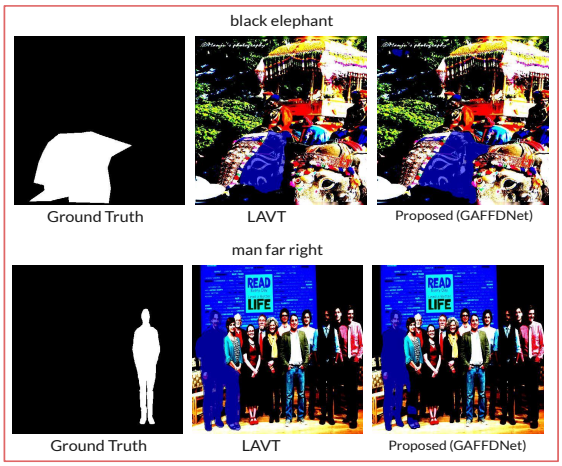

Figure 5.5: GAFFDNet Failure Cases

based on referring expressions, highlighting its robustness and generalization capability in the Referring Image Segmentation task.
The visualization results confirm that GAFFDNet achieves more precise and reliable segmentation than LAVT across diverse datasets and queries, demonstrating both its robustness and practical effectiveness, while also highlighting areas for potential improvement in handling very small or challenging object regions.

5.7 Limitations

Several failure cases highlight situations in which the proposed model struggles to accurately segment target objects. Figure 5.5 illustrates failure cases where the model struggles to accurately segment the target objects. In the first example, the model fails to segment the black elephant due to limited knowledge of this instance, insufficient representation of similar instances, and the low contrast between the object and its background. In the second example, the model incorrectly segments the man on the far left instead of the man on the far right caused by contextual confusion among visually similar human instances.

Chapter 6

Conclusion

6.1 Summary

In this work, we present the proposed GAFFDNet, a novel framework for referring image segmentation that significantly advances language grounding, multimodal fusion, decoder feature aggregation, and overall segmentation accuracy. The model leverages a multi-stage contrastively trained *GTE* language encoder, producing semantically rich and highly discriminative embeddings that improve alignment between linguistic expressions and visual content.

To enhance cross-modal interaction, an adaptive gating mechanism dynamically fuses visual and linguistic features, allowing the model to emphasize the most relevant modality depending on context. The decoder employs full-scale multimodal feature aggregation, integrating information across all encoder and decoder stages, which facilitates rich semantic reasoning while preserving fine spatial details. This design enables precise mask generation even in complex scenes with multiple objects or ambiguous referring expressions.

Extensive experiments on three widely used benchmark datasets—RefCOCO, RefCOCO+, and G-Ref—demonstrate that GAFFDNet achieves state-of-the-art performance across multiple evaluation metrics, validating the effectiveness of each proposed component. While the model demonstrates strong overall performance, accurately segmenting tiny, occluded, or highly cluttered objects remains challenging. Overall, GAFFDNet establishes a robust foundation for more context-aware, semantically precise, and reliable visual grounding systems.

6.2 Future Work

While GAFFDNet achieves competetive performance in referring image segmentation, several research directions can be explored to further enhance its capabilities and applicability.

Future work could explore integrating other sensory modalities, such as depth information from sensors or audio cues from videos. Depth data could help disambiguate objects in cluttered or overlapping scenes, while audio cues could provide contextual information, such as identifying a speaker or source associated with the referred object. Multimodal fusion of these additional signals could improve contextual understanding and robustness in real-world environments Referring image segmentation is computationally intensive. Research could focus on optimizing the model for

real-time performance without significant loss of segmentation quality. Techniques such as model pruning, knowledge distillation, or efficient transformer architectures could reduce inference time and enable deployment on resource-constrained devices, such as mobile platforms or robotics systems.

Despite strong overall performance, segmentation of very small or partially occluded objects remains challenging. Future improvements could involve multi-scale feature refinement, attention-based mechanisms to better capture subtle object cues, or specialized decoder strategies that emphasize fine-grained spatial details. Additionally, leveraging external context or scene priors may help the model infer hidden object boundaries more accurately.GAFFDNet relies on large annotated datasets, which can be expensive and time-consuming to obtain. Future research could investigate semi-supervised or unsupervised learning techniques, allowing the model to learn from unlabeled or partially labeled data. Methods like self-training, pseudo-labeling, or contrastive learning on unannotated images could expand the model's generalization to diverse scenarios and domains. Applying GAFFDNet to datasets or real-world scenarios beyond the benchmarks used in this study may reveal domain adaptation challenges. Future work could explore transfer learning, domain adaptation, or adversarial training strategies to ensure the model remains robust when encountering new object categories, environments, or visual styles.

Since RIS has applications in human–computer interaction, visual search, and assistive technologies, future research could involve user studies or human-in-the-loop experiments. These studies could measure the system's practical usability, assess how effectively it interprets natural language queries in real-world tasks, and guide enhancements for interactive and adaptive systems. Finally, making GAFFDNet's decisions interpretable could increase trust and adoption in sensitive domains, such as healthcare or autonomous systems. Future work could involve visualization of cross-modal attention maps, explanation of language grounding, or quantifying the contribution of each modality to final predictions, enabling users and researchers to better understand and debug the model's reasoning.

Bibliography

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "Imagenet: A large-scale hierarchical image database," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, Jun. 2009.

[2] T.-Y. Lin et al., "Microsoft coco: Common objects in context," in *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, Springer, 2014, pp. 740–755.

[3] R. Hu, M. Rohrbach, and T. Darrell, "Segmentation from natural language expressions," *ArXiv*, vol. abs/1603.06180, 2016.

[4] V. K. Nagaraja, V. I. Morariu, and L. S. Davis, "Modeling context between objects for referring expression understanding," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, Springer, 2016, pp. 792–807.

[5] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, Springer, 2016, pp. 69–85.

[6] C. Liu, Z. L. Lin, X. Shen, J. Yang, X. Lu, and A. L. Yuille, "Recurrent multimodal interaction for referring image segmentation," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1280–1289, 2017.

[7] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[8] A. Vaswani et al., "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, Long Beach, California, USA: Curran Associates Inc., 2017, pp. 6000–6010, ISBN: 9781510860964.

[9] R. Li et al., "Referring image segmentation via recurrent refinement networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2018.

[10] L. Yu et al., "Mattnet: Modular attention network for referring expression comprehension," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1307–1315, 2018.

[11] D.-J. Chen, S. Jia, Y.-C. Lo, H.-T. Chen, and T.-L. Liu, "See-through-text grouping for referring image segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019.

[12] Y.-W. Chen, Y.-H. Tsai, T. Wang, Y.-Y. Lin, and M.-H. Yang, "Referring expression object segmentation with caption-aware consistency," in *British Machine Vision Conference*, 2019.

[13] A. Paszke, "Pytorch: An imperative style, high-performance deep learning library," *arXiv preprint arXiv:1912.01703*, 2019.

[14] L. Ye, M. Rochan, Z. Liu, and Y. Wang, "Cross-modal self-attention network for referring image segmentation," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 494–10 503, 2019.

[15] Z. Hu, G. Feng, J. Sun, L. Zhang, and H. Lu, "Bi-directional relationship inferring network for referring image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020.

[16] Z. Hu, G. Feng, J. Sun, L. Zhang, and H. Lu, "Bi-directional relationship inferring network for referring image segmentation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4423–4432.

[17] H. Huang et al., "Unet 3+: A full-scale connected unet for medical image segmentation," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1055–1059, 2020.

[18] S. Huang et al., "Referring image segmentation via cross-modal progressive comprehension," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 485–10 494, 2020.

[19] T. Hui et al., "Linguistic structure guided context modeling for referring image segmentation," *ArXiv*, vol. abs/2010.00515, 2020.

[20] G. Luo et al., "Cascade grouped attention network for referring expression segmentation," pp. 1274–1282, Oct. 2020.

[21] G. Luo et al., "Multi-task collaborative network for joint referring expression comprehension and segmentation," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 031–10 040, 2020.

[22] H. Ding, C. Liu, S. Wang, and X. Jiang, "Vision-language transformer and query generation for referring segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 16 321–16 330.

[23] G. Feng, Z. Hu, L. Zhang, and H. Lu, "Encoder fusion network with co-attention embedding for referring image segmentation," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15 501–15 510, 2021.

[24] Y. Jing, T. Kong, W. Wang, L. Wang, L. Li, and T. Tan, "Locate then segment: A strong pipeline for referring image segmentation," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9853–9862, 2021.

[25] S. Liu, T. Hui, S. Huang, Y. Wei, B. Li, and G. Li, "Cross-modal progressive comprehension for referring segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, pp. 4761–4775, 2021.

[26] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[27] S. Yang, M. Xia, G. Li, H.-Y. Zhou, and Y. Yu, "Bottom-up shift and reasoning for referring image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 11 266–11 275.

[28] N.-W. Kim, D. Kim, C. Lan, W. Zeng, and S. Kwak, "Restr: Convolution-free referring image segmentation using transformers," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18 124–18 133, 2022.

[29] Z. Wang et al., "Cris: Clip-driven referring image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 686–11 695.

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

[31] Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie, and M. Zhang, "Towards general text embeddings with multi-stage contrastive learning," *arXiv preprint arXiv:2308.03281*, 2023.

[32] W. Wang et al., "Cm-masksd: Cross-modality masked self-distillation for referring image segmentation," *IEEE Transactions on Multimedia*, vol. 26, pp. 6906–6916, 2023.

[33] Z. Xu, Z. Chen, Y. Zhang, Y. Song, X. Wan, and G. Li, "Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17 457–17 466, 2023.

[34] W. Zhao, Y. Rao, Z. Liu, B. Liu, J. Zhou, and J. Lu, "Unleashing text-to-image diffusion models for visual perception," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2023, pp. 5729–5739.