Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FAKULTÄT
FÜR MATHEMATIK, INFORMATIK
UND NATURWISSENSCHAFTEN

BACHELORTHESIS

Analyse und Leistungsbewertung von Transformer-basierten Objekterkennungs- und Segmentierungsmodellen

Florina Sophie Ludwig

Studiengang: Informatik
Matrikelnummer: 7509743
1. Prüfer: Prof. Dr. Chris Biemann, Universität Hamburg
2. Prüfer: Florian Schneider, Universität Hamburg

Language Technology
Fachbereich Informatik
Fakultät für Mathematik, Informatik und Naturwissenschaften

Universität Hamburg
Hamburg, Germany

Bachelorthesis eingereicht für den

Bachelor of Science (B. Sc.)

Abgabedatum 28. Juli 2025

Analyse und Leistungsbewertung von Transformer-basierten Objekterkennungs- und Segmentierungsmodellen

Bachelorthesis eingereicht von: Florina Sophie Ludwig

Abgabedatum: 28. Juli 2025

Betreuer:
Florian Schneider, Universität Hamburg

Komitee:
1. Prüfer: Prof. Dr. Chris Biemann, Universität Hamburg
2. Prüfer: Florian Schneider, Universität Hamburg

Universität Hamburg, Hamburg, Germany
Fakultät für Mathematik, Informatik und Naturwissenschaften
Fachbereich Informatik

Language Technology

Abstract

Im Rahmen der vorliegenden Arbeit wird Grounded SAM 2 (H Zhou et al. 2024; Ren und Shen, 2025), ein transformer-basiertes multimodales System, entwickelt für die Bildsegmentierung, bezüglich seiner Segmentierungsleistung mit unterschiedlichen Modell-Backbone-Kombinationen aus Florence 2 (Xiao et al. 2023), Grounding DINO (Liu et al. 2024), Hiera-large und Hiera-small (Ryali et al. 2023) analysiert. Dazu werden die Kombinationen auf den Datensätzen Microsoft COCO 2017 val. (Lin et al. 2015), RefCOCO und RefCOCO+ (Kazemzadeh et al. 2014) getestet und hinsichtlich Average Recall und Average Precision zu unterschiedlichen Objektgrößen sowie ihrer Laufzeit verglichen. Außerdem erfolgt eine Usability-Studie auf einer Webanwendung, die die modellgestützte Methode der manuellen gegenüberstellt. Die Ergebnisse zeigen, dass Florence 2 insbesondere bei großen Objekten präzise arbeitet, während kleine häufiger von allen Modell-Backbone-Kombinationen übersehen werden. Grounding DINO bietet mit Hiera-small einen guten Kompromiss aus Rechenzeit und Genauigkeit. In der Studie wurde gezeigt, dass Grounded SAM 2 eine deutliche Zeitersparnis und höhere Qualität bei der Segmentierung bietet. Die Arbeit zeigt das Potenzial der verschiedenen Modell-Backbone-Kombinationen von Grounded SAM 2 in praxisnahen Annotationstools auf.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Bilder und Videos sind aus dem Alltag einer modernen Gesellschaft nicht mehr wegzudenken. Jeden Tag werden große Mengen an Bild- und Video-Daten erzeugt und verarbeitet. Da die manuelle Auswertung bei großen Datenmengen in der Regel zu zeitaufwendig ist, bedarf es in vielen Bereichen Methoden, diese Daten automatisiert auswerten zu können. Dafür werden Verfahren benötigt, die in der Lage sind, Inhalte in Bildern und Videos zuverlässig zu lokalisieren (Objektdetektion) und Bereiche im Bild auf Pixelebene abzugrenzen (Segmentierung) (Szeliski, 2022).

Die Computer Vision, ein Teilgebiet der Informatik, beschäftigt sich mit dieser Aufgabe. Typische zu lösende Probleme sind hier die Identifikation und Klassifikation von Objekten, die Bestimmung ihrer Position im Bild sowie die detaillierte Segmentierung von Bildbereichen. In den letzten Jahren haben sich die Methoden dazu stark weiterentwickelt. Moderne Ansätze basieren immer mehr auf Künstlicher Intelligenz (KI), die eigenständig Bilddaten auswerten und dabei komplexe Zusammenhänge erkennen können (Szeliski, 2022). Außerdem sind neue, sogenannte multimodale Ansätze in der Lage, zum Beispiel Text oder Sprache als Information mit einzubeziehen und gleichzeitig mit Bild oder Video zu verarbeiten, um Aufgaben so kontextabhängig zu lösen (Caffagni et al. 2024).

In dieser Arbeit werden solche multimodalen Ansätze genauer untersucht. Im Mittelpunkt steht hier der Vergleich verschiedener Modelle und die Frage, wie gut diese im praktischen Einsatz funktionieren.

1.1 Motivation

Die Bilddetektion und die Segmentierung von Objekten bilden die Grundlage vieler verschiedener relevanter Anwendungsfelder, wie unter anderem medizinische Bildanalyse oder intelligente Überwachungssysteme. Fortschritte in der Computer Vision und insbesondere der Bildsegmentierung tragen dazu bei, Bilddaten automatisiert zu analysieren und auszuwerten, so dass die Ergebnisse für den jeweiligen Anwendungsfall nutzbar gemacht werden (Szeliski, 2022).
Die Entwicklung multimodaler KI-Modelle, die in der Lage sind, visuelle Informationen nicht nur einzeln, sondern zusammen mit anderen Modalitäten wie unter anderem Text zu analysieren, führt

zu erheblichen Fortschritten in dem Bereich der Bildsegmentierung. Die Architekturen können also komplexe Aufgaben wie die Segmentierung eines bestimmten Objekts auf Basis eines natürlichsprachlichen textuellen Prompts (Texteingabe) lösen (Caffagni et al. 2024).
Die Segmentierung bei solchen Modellen erfolgt typischerweise in mehreren Bearbeitungsschritten. Zunächst werden auf Grundlage der Bild- und Textinformationen relevante Bereiche im Bild (häufig in Form von Bounding-Boxes (rechteckigen Begrenzungsrahmen)) lokalisiert. Danach wird eine Segmentierungsmaske für den entsprechenden Bereich erzeugt, die für jeden Pixel aufzeigt, ob er zu dem entsprechenden Objekt gehört oder nicht.

Grounded SAM ist ein Beispiel für ein solches multimodales System, welches von Ren et al. (2024) vorgestellt und speziell für Bildsegmentierungsaufgaben entwickelt wurde und mehrere multimodale Ansätze kombiniert. Grounded SAM 2 (Ren und Shen, 2025) ist eine weiterentwickelte Version von Grounded SAM, die unter anderem zusätzliche Modellkomponenten integriert und so umfassendere Experimente ermöglicht. In Grounded SAM 2 sind Komponenten wie Florence 2 (Xiao et al. 2023), ein multimodales Modell von Microsoft, das Aufgaben wie Objektdetektion, Bildbeschreibung und Bildsegmentierung bearbeiten kann, oder Grounding DINO (Liu et al. 2024), ein Modell speziell für die textbasierte Objektdetektion, integriert. Grounded SAM 2 nutzt Florence 2 und Grounding DINO, um Bounding-Boxes zu erstellen. Außerdem bindet Grounded SAM 2 unterschiedliche Hiera-Modelle ein (Ryali et al. 2023). Die unterschiedlichen Hiera-Modelle, darunter Hiera-small und Hiera-large, werden in Grounded SAM 2 als visuelle Backbones innerhalb des Segmentierungsmoduls SAM 2 (Ravi et al. 2024) eingesetzt, um exakte Extraktionen von Bildmerkmalen zu ermöglichen und die Segmentierungsleistung weiter zu verbessern. SAM 2 kann die von Modulen wie Florence 2 oder Grounding DINO erzeugten Bounding-Boxen übernehmen und weiterverarbeiten.

1.2 Ziel der Arbeit

Ziel dieser Arbeit ist die Analyse der Leistungsfähigkeit des multimodalen Segmentierungssystems Grounded SAM 2 (GSAM2). Dabei wird GSAM2 mit verschiedenen Kombinationen aus den Komponenten Florence 2 und Grounding DINO und den Backbones Hiera-small und Hiera-large getestet (Xiao et al. 2023; Ren und Shen, 2025; Liu et al. 2024). Die Frage, welche Modell-Backbone-Kombination hinsichtlich der Segmentierungsgenauigkeit und Effizienz die besten Ergebnisse liefert, soll dabei im Mittelpunkt stehen. Dabei wird die Segmentierungsgenauigkeit auf den Datensätzen COCO, RefCOCO und RefCOCO+ getestet (Lin et al. 2015; Kazemzadeh et al. 2014).

Zur Unterstützung der Analyse wird eine Webanwendung entwickelt, die eine Visualisierung der Segmentierungsergebnisse der unterschiedlichen Modell- und Backbone-Kombinationen ermöglicht. Außerdem wird mithilfe der Webanwendung eine Usability-Studie durchgeführt, die die Praxistauglichkeit von Grounded SAM 2 testen und diese mit manueller Segmentierung vergleichen soll, um die Effizienz unter realistischen Bedingungen zu testen.

Die Bewertung der Ergebnisse erfolgt anhand der Metriken Average Precision (AP) und Average Recall (AR) die als Standard für die Evaluierung von COCO-Datensätzen gelten (Lin et al. 2015). Für die Auswertung auf dem RefCOCO+ und RefCOCO-Datensatz wird zusätzlich die mean Intersection over Union (mIoU) berücksichtigt, da diese Metrik in der Literatur gängig zur Bewertung der referenzbasierten Datensätze ist (Xiao et al. 2023). Da bisher kein direkter Vergleich der genannten Modellvarianten vorliegt, soll diese Lücke im Rahmen der Bachelorarbeit geschlossen werden. Für diesen Zweck werden im Folgenden konkrete Forschungsfragen spezifiziert.

1.2.1 Forschungsfragen

Auf der Basis der in Abschnitt 1.2 beschriebenen Zielsetzung ergeben sich für die Arbeit folgende Forschungsfragen.

1. *Wie wirken sich die in Grounded SAM 2 eingesetzten Modell-Backbone-Kombinationen, Florence 2 bzw. Grounding DINO, jeweils mit Hiera-small oder Hiera-large, auf die Segmentierungsgenauigkeit und Rechenzeit aus?*

2. *Welche Unterschiede zeigen sich in der Segmentierungsgenauigkeit für kleine, mittlere und große Objekte zwischen den Modell-Backbone-Kombinationen innerhalb von Grounded SAM 2?*

3. *Wie wirkt sich die Nutzung von Grounded SAM 2 im Vergleich zur manuellen Segmentierung auf die Verarbeitungszeit und Effizienz von Segmentierungsaufgaben aus?*

1.3 Struktur der Arbeit

Die Arbeit gliedert sich insgesamt in sieben Kapitel. Nach der einführenden Zielsetzung in Kapitel 1 folgt im zweiten Kapitel der theoretische Hintergrund, einschließlich der Erklärung grundlegender Konzepte wie Transformer-Modelle, Objektdetektion und Bildsegmentierung sowie einer kurzen Beschreibung der benutzten Datensätze und der Evaluationsmetriken. In Kapitel 3 werden einige verwandte Arbeiten dargestellt. Kapitel 4 beschreibt die Ergebnisse der Segmentierung und Bounding Boxes der durchgeführten Experimente auf den Datensätzen COCO, RefCOCO und RefCOCO+. Anschließend werden im fünften Kapitel die Methodik und Ergebnisse einer Usability-Studie vorgestellt, bei der die verschiedenen Modell-Backbone-Kombinationen sowie das manuelle Segmentieren hinsichtlich Effizienz und Nutzerzufriedenheit miteinander verglichen werden. Kapitel 6 diskutiert die Ergebnisse im Kontext der in 1.2.1 spezifizierten Forschungsfragen und vergleicht diese mit bestehenden Arbeiten. Zum Schluss fasst das siebte Kapitel wesentliche Erkenntnisse kurz zusammen und gibt einen Ausblick auf Forschungsrichtungen sowie Verbesserungsmöglichkeiten.

2 Theoretischer Hintergrund und Datensätze

2.1 Computer Vision

Die Computer Vision befasst sich mit der Interpretation und Analyse visueller Daten und ist ein Teilbereich der künstlichen Intelligenz. Es werden Informationen aus Bildern und Videos extrahiert und auf dieser Grundlage zum Beispiel Objekte klassifiziert. Dabei werden unterschiedliche Verfahren aus der Bildverarbeitung und dem maschinellen Lernen eingesetzt (Szeliski, 2022).

2.2 Multimodale Systeme

Multimodale Systeme können Informationen aus verschiedenen Quellen wie Bildern, Videos und Texten gleichzeitig verarbeiten. Sie verknüpfen Daten miteinander, um komplexe Inhalte besser zu erfassen und Zusammenhänge zu erkennen, welche bei Betrachtung einzelner Modalitäten nicht sichtbar wären (Baltrušaitis et al. 2019).
Ein bekanntes Beispiel einer multimodalen Anwendung ist ChatGPT-4 (Achiam et al. 2024). Neben Text kann das Modell ebenfalls Bilder und Dokumente gleichzeitig verarbeiten und so beispielsweise ein Bild im Kontext einer eingegebenen Frage analysieren und daraufhin eine textuelle Antwort generieren.
Ein weiteres Beispiel sind autonome Fahrzeuge. Sie verwenden unterschiedliche Sensoren wie Kameras, Radar oder GPS, um ihre Umgebung möglichst genau zu erfassen. Die Informationen dieser verschiedenen Quellen werden gemeinsam ausgewertet, um zum Beispiel die Spur auch bei Dunkelheit verfolgen zu können (Yao et al. 2024).

2.3 Transformer-Modelle

Viele moderne multimodale Systeme wie auch Grounding DINO, SAM 2 oder Florence 2 basieren auf Transformer-Modellen. Transformer-Modelle sind neuronale Netze, die in der Regel mit einer Encoder-Decoder-Architektur arbeiten, die es ermöglicht, sowohl Informationen einer einzelnen Eingabe, wie Bild oder Text, als auch zwischen unterschiedlichen Eingaben effektiv zu verarbeiten (Vaswani et al. 2023). Dabei werden die zwei zentralen Mechanismen Self- und Cross-Attention genutzt (abgebildet in 2.1).

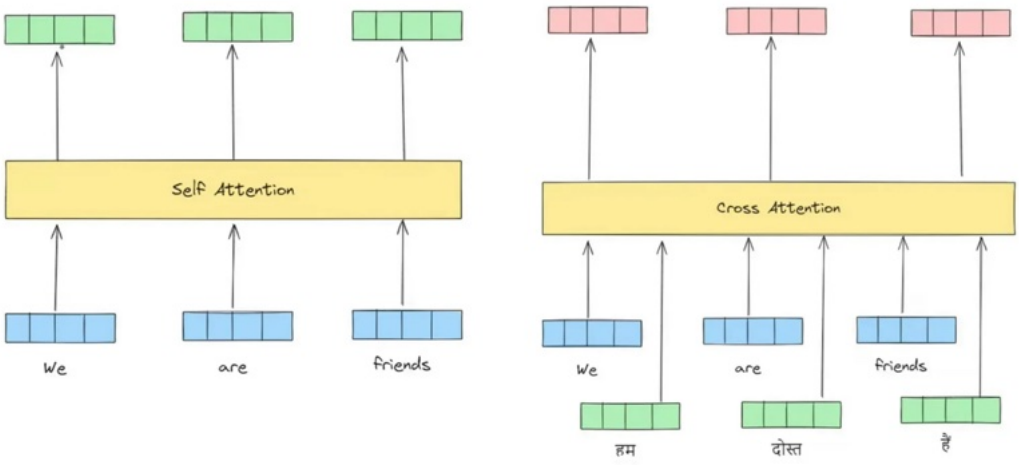

Abbildung 2.1: Self- vs. Cross-Attention. Grafik entnommen aus Sachinsoni (2024)

.

2.3.1 Self-Attention

Self-Attention wird vor allem im Encoder von Transformer-Architekturen verwendet. Wie in Abbildung 2.1 auf der linken Seite dargestellt, erhält der Encoder einzelne Eingabetokens und ermittelt durch Self-Attention Beziehungen zwischen den einzelnen Tokens (Vaswani et al. 2023).

2.3.2 Cross-Attention

Der Decoder verwendet Cross-Attention, um die verschiedenen Modalitäten (z.B. Text und Bild), die vom Encoder verarbeitet wurden, miteinander zu verknüpfen. Dabei wird jedes Bildtoken mit jedem Texttoken in Beziehung gesetzt, um so kontextualisierte Informationen zu erhalten. Dies ist auf der rechten Seite von Abbildung 2.1 dargestellt. Durch die daraus entstehende gemeinsame Repräsentation der Modalitäten können Aufgaben wie textbasierte Segmentierung oder Objektdetektion ausgeführt werden (Luo et al. 2021).

2.4 Objektdetektion

Die Objektdetektion ist ein Teilgebiet der Computer Vision und bezieht sich auf die Lokalisierung von Objekten auf Bildern oder in Videos. Das Ziel kann zum Beispiel sein, die Position eines bestimmten Objektes auf einem Bild zu identifizieren. Dabei werden in der Regel so genannte Bounding Boxes, also rechteckige Rahmen, die die Position eines Objektes angeben, verwendet (Szeliski, 2022).

Abbildung 2.2: Beispiel für Bounding Boxes von einer Person und einem Surfboard. Das Bild stammt aus dem COCO 2017 Datensatz und wurde mit Grounded SAM 2 annotiert (Lin et al. 2015).

Auf der Abbildung 2.2 wurden zwei Objekte detektiert und mit Bounding Boxes gekennzeichnet. Hier wird jeweils einmal die Person (rote BBox)und einmal das Surfboard, auf dem die Person sich befindet (lila BBox) von einer Bounding Box umschlossen.

2.5 Bildsegmentierung

Die Bildsegmentierung erweitert die Objektdetektion durch feinere Lokalisierung von Objekten auf Pixelebene. Das bedeutet, die Segmentierung zielt darauf ab, jedem Pixel eine semantische Bedeutung zuzuordnen und so die exakte Kontur von Objekten zu erfassen (Szeliski, 2022; He et al. 2018). Zusätzlich wird zwischen semantischer, instanz-basierter und referenz-basierter Segmentierung unterschieden. Während es bei der semantischen Segmentierung darum geht, allen Objekten derselben Klasse (z.B. alle Autos im Bild) dieselbe Labelung zu geben, unterscheidet die instanz-basierte Segmentierung zusätzlich zwischen unterschiedlichen Instanzen innerhalb einer Klasse (z.B. „Auto 1“, „Auto 2“). Bei der referenz-basierten Segmentierung wird über den Textprompt nur ein spezifisches Objekt referenziert, anstatt eine gesamte Klasse oder Instanzen zu markieren (He et al. 2018).

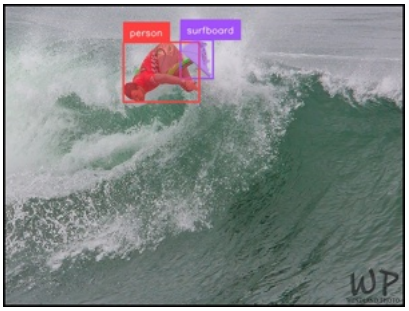

Abbildung 2.3: Beispiel für Segmentierung von einer Person und einem Surfboard. Das Bild stammt aus dem COCO 2107 evaluations Datensatz und wurde mit Grounded SAM 2 annotiert (Lin et al. 2015)

.

Auf Abbildung 2.3 ist ein Beispiel für die Segmentierung von zwei Objekten. Hier wird einmal die Person (in rot) segmentiert und einmal das Surfboard (in lila). Auf dem Bild sind zusätzlich die Bounding Boxes aus dem vorherigen Abschnitt abgebildet.

2.6 Komponenten von Grounded SAM 2

Grounded SAM 2 setzt sich aus mehreren verschiedenen Komponenten, welche teilweise zusammenarbeiten, um präzise Segmentierung zu ermöglichen, zusammen (Ren und Shen, 2025). Während Florence 2 und Grounding DINO Bounding Boxes auf Basis von sprachlichen Prompts erzeugen, übernimmt SAM 2 die eigentliche Segmentierung. In den folgenden Abschnitten werden die einzelnen Komponenten von Grounded SAM 2 näher erläutert.

2.6.1 Florence 2

Florence 2 ist einer der optionalen textbasierten Komponenten von Grounded SAM 2 und ist in der Lage, verschiedene komplexe Bildverarbeitungsaufgaben zu lösen. Neben der Fähigkeit, Objekte zu detektieren und zu segmentieren, kann Florence 2 beispielsweise auch Bildbeschreibungen generieren. Aufgrund seiner Zero-Shot-Fähigkeiten kann das Modell diese Aufgaben auch bewältigen, wenn keine spezifischen Trainingsdaten für die zu lösenden Aufgaben oder Objekte vorliegen. Zero-Shot-fähig bezeichnet die Fähigkeit eines Modells, Aufgaben bearbeiten zu können, die während des Trainings nicht explizit gesehen wurden.

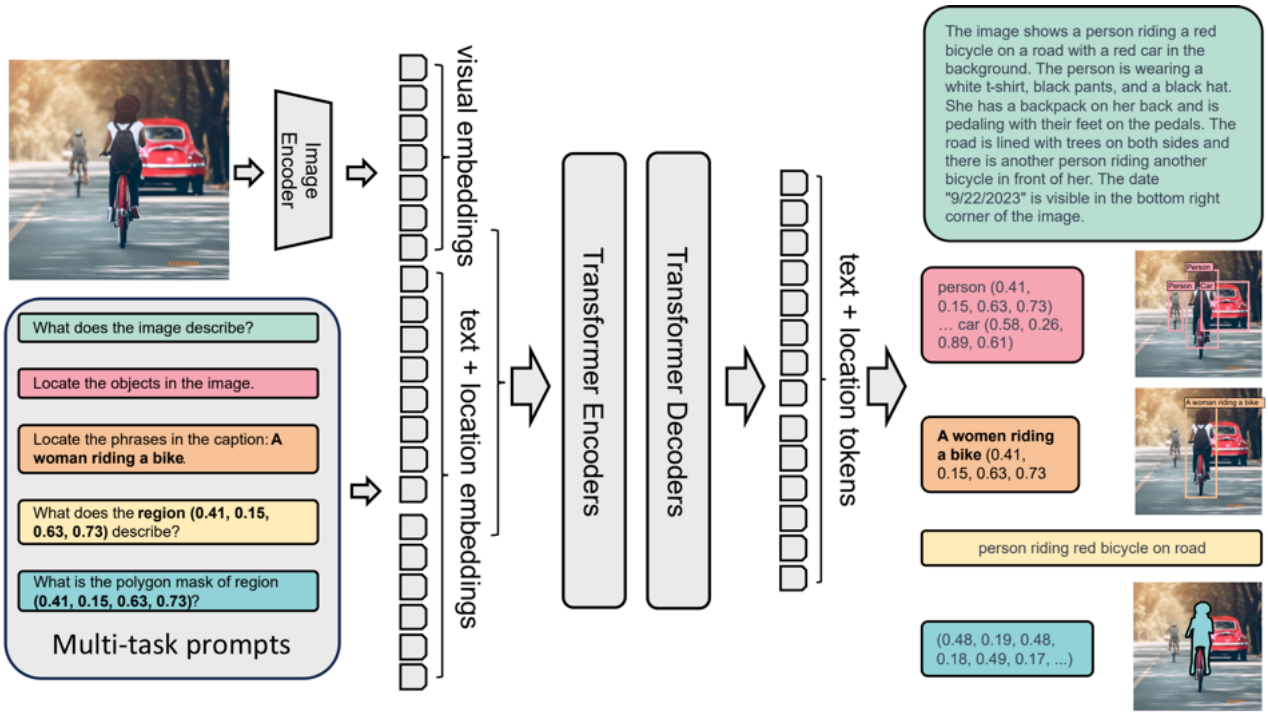

Abbildung 2.4: Architektur von Florence 2. Diese Abbildung wurde aus dem Originalpapier von Xiao et al. (2023) entnommen.

Die Abbildung 2.4 zeigt, wie Florence 2 Aufgaben in der Objekterkennung löst. Dem Modell wird ein Bild sowie ein Textprompt übergeben, und das Bild wird zunächst in einem **Image Encoder** verarbeitet.

Der Image Encoder wandelt das Bild in eine für Florence 2 nutzbare Merkmalsdarstellung (Feature Embedding) um. Anstatt das Bild in Pixelform weiterzugeben, werden dabei visuelle Informationen wie Farben, Formen oder Kanten extrahiert. Die resultierenden Merkmale werden anschließend mit den Textinformationen in einer **Transformer-Encoder-Decoder-Architektur** kombiniert und weiterverarbeitet. Der Transformer Encoder erzeugt eine kontextualisierte Darstellung aller Tokens (z.B. Bildausschnitte oder Teilwörter) und der Decoder generiert auf dieser Basis eine spezifische Ausgabe. Abhängig von der jeweiligen Aufgabe können auf dieser Basis unterschiedliche Ergebnisse wie etwa Bounding Boxes, Segmentierungsmasken oder Bildbeschreibungen erzeugt werden. (Xiao et al. 2023; Ren und Shen, 2025)

2.6.2 Grounding DINO

Grounding DINO (Liu et al. 2024) ist eine weitere textbasierte Komponente von Grounded SAM 2 (Ren und Shen, 2025) und wurde speziell für die Objekterkennung entwickelt. Das Modell ist ebenfalls Zero-Shot-fähig und kann auf Grundlage textueller Beschreibungen/Kategorien relevante Objekte in Bildern lokalisieren, ohne dabei auf vordefinierte Begriffe angewiesen zu sein.

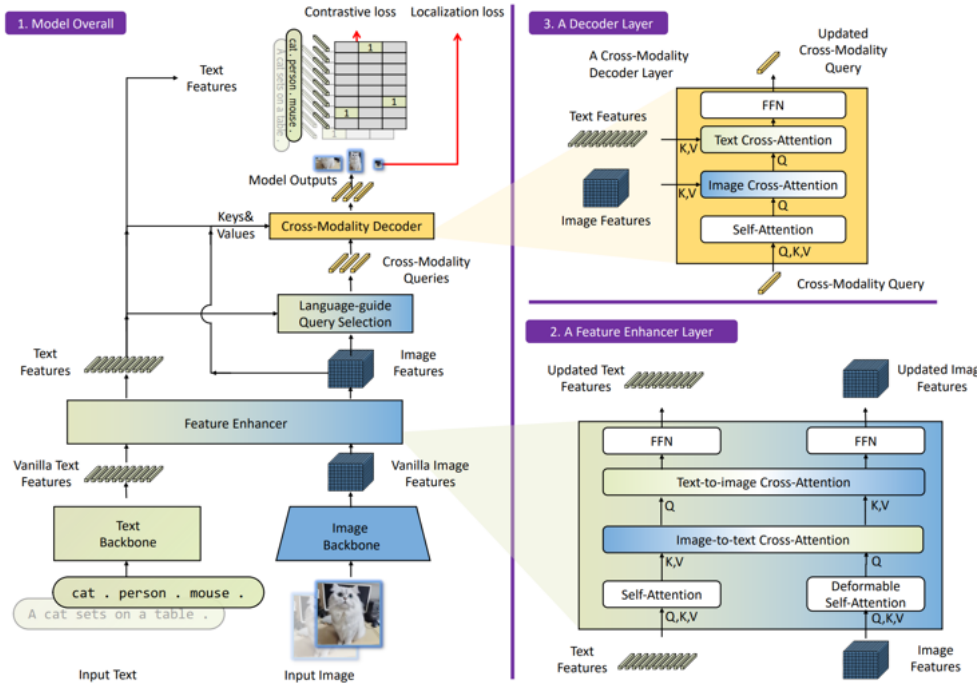

Abbildung 2.5: Architektur von Grounding DINO. Diese Abbildung wurde aus dem Originalpapier von Liu et al. (2024) entnommen.

In Abbildung 2.5 ist die Architektur von Grounding DINO dargestellt, in der gezeigt wird, wie das Modell Informationen miteinander kombiniert, um Objekte zu lokalisieren. Ein Bild sowie ein Textprompt werden im ersten Schritt an das Modell übergeben, die anschließend den Feature Enhancer durchlaufen. Der Feature Enhancer versucht unter anderem durch Text-to-Image Cross Attention sowie Image-to-Text Cross Attention Bild- und Textinformationen zu kombinieren und so beispielsweise bestimmte Bildbereiche hervorzuheben. In der Language Guide Query Selection werden Bildbereiche Textabschnitten zugeordnet und daraufhin im Cross-Modality Decoder kombiniert. Der Cross-Modality Decoder setzt Text/Image Cross Attention ein, um letztendlich Bounding Boxes zu generieren (Liu et al. 2024).

2.6.3 Segment Anything Model 2 (SAM 2)

SAM 2 (Ravi et al. 2024), eine Weiterentwicklung von dem ursprünglich von Meta vorgestellten Modell SAM (Kirillov et al. 2023), wurde für die Segmentierung von Objekten in Bildern sowie Videos entwickelt. Es werden verschiedene Arten von Prompts unterstützt, die angeben, welche Bereiche im Bild oder Video segmentiert werden sollen. Es können zum Beispiel Bounding Boxes (BBoxes), Masken oder Punkte als Eingabe in Bildern oder Videos verarbeitet werden.

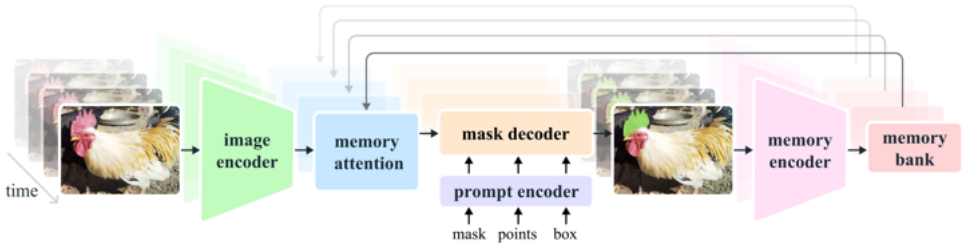

Abbildung 2.6: Architektur von SAM 2. Diese Abbildung wurde aus dem Originalpapier von Ravi et al. (2024) entnommen.

Da SAM 2 einzelne Objekte segmentiert, also Masken pro Instanz ausgibt, handelt es sich bei SAM 2 um Instanz-Segmentierung, bei der verschiedene Instanzen derselben Klasse separat segmentiert werden (Kirillov et al. 2023). Die Abbildung 2.6 beschreibt die Architektur von SAM 2 und zeigt, wie das Modell Bilder oder Videos verarbeitet, wobei bei Videos ein Frame nach dem anderen analysiert wird. Die Hauptkomponenten der Architektur sind:

Image-Encoder (Hiera-Modelle)

In Segmentierungsmodellen wie SAM 2 kommen häufig Backbones, was neuronale Netzwerke sind, die Merkmale aus dem Eingabebild oder Videoframe extrahieren, zum Einsatz (Szeliski, 2022). Konkret verwendet SAM 2 verschiedene Varianten von Hiera-Modellen als Backbone, die als Image Encoder eingesetzt werden. Die unterschiedlichen Hiera-Varianten unterscheiden sich hauptsächlich in ihrer Größe, also zum Beispiel in der Anzahl ihrer Parameter. Größere Modelle wie Hiera-large verfügen über mehr Parameter und können dadurch komplexere Muster erkennen. Kleinere Varianten wie Hiera-small sind dafür effizienter, benötigen weniger Rechenleistung und eignen sich so besonders gut für Aufgaben mit begrenzten Ressourcen. Die unterschiedlichen Modelle beeinflussen die Genauigkeit der Segmentierung (Ryali et al. 2023).

Da eine umfassende Analyse aller verfügbaren Hiera-Modelle den Rahmen dieser Arbeit überschreiten würde, beschränkt sich die Evaluation auf die beiden Varianten **Hiera-small** und **Hiera-large**.

Memory-Encoder, Memory-Bank und Memory-Attention

Wenn Objekte in Videos segmentiert werden sollen, werden vorangegangene Masken und Bounding-Boxes (BBoxes) durch den Memory-Encoder mit den aktuellen zusammengefasst, um Informationen der vorangegangenen Frames nutzen zu können. In der Memory-Bank werden diese Informationen, also Masken und BBoxes, gespeichert und dann durch die Memory-Attention für den aktuellen Frame genutzt.
Die Komponenten kommen in SAM 2 ausschließlich bei der Verarbeitung von Videos zum Einsatz und sind für die Analyse einzelner Bilder nicht relevant.

Prompt-Encoder

Der Prompt-Encoder verarbeitet den Input (also den Prompt) also zum Beispiel übergebene BBoxes, Masken oder Text und gibt die resultierenden Merkmale an den Mask-Decoder weiter.

Mask-Decoder

Der Mask-Decoder arbeitet mit dem Prompt-Encoder zusammen, um eine Segmentierungsvorhersage zu generieren. Wenn der Prompt unklar ist, werden mehrere Masken erstellt und die beste ausgewählt. (Ravi et al. 2024; Mukherjee, 2024).

2.7 Architektur von Grounded SAM 2

Grounded SAM 2 kombiniert visuelle und textuelle Merkmale, um Objekte in Bildern sowie Videos zu segmentieren. Die Verarbeitung erfolgt dabei in zwei Stufen. Als erstes werden mittels textgesteuerter Modelle relevante Bildbereiche lokalisiert, auf denen dann als nächstes durch ein Segmentierungsmodul eine genaue Maske erstellt wird.

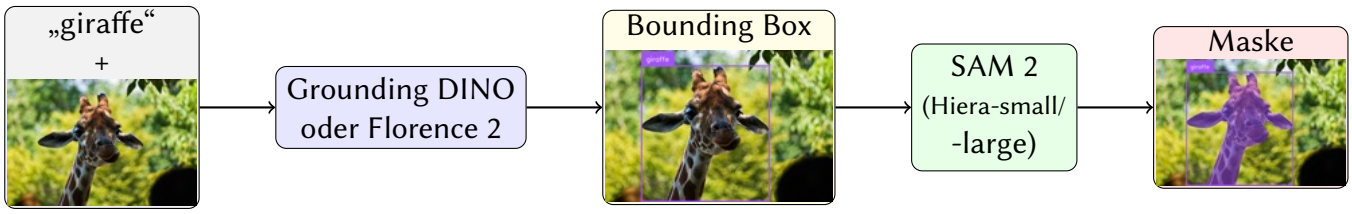

Abbildung 2.7: Abbildung zeigt den Ablauf von Grounded SAM 2 anhand eines Bildes von einer Giraffe und dem Textprompt „giraffe“. Das Bild stammt Ursprünglich aus dem COCO Datensatz (Lin et al. 2015) und wurde mit Grounded SAM 2 annotiert.

Das multimodale System unterstützt für die Lokalisierung der Objekte unterschiedliche Komponenten, wie zum Beispiel Grounding DINO (siehe Abschnitt 2.6.2), DINO 1.5, DINO 1.6, DINO X und Florence 2 (siehe Abschnitt 2.6.1). Diese Modelle verarbeiten jeweils ein Eingabebild gemeinsam mit einem Textprompt und generieren Bounding Boxes, welche die Grundlage für die darauf folgende genaue Segmentierungsmaske bilden (siehe Abbildung 2.7). Zusätzlich wird jeder Vorhersage ein Konfidenzwert (Confidence-Score) zugewiesen, der angibt, wie sicher das Modell sich mit seiner Vorhersage ist. Dieser Score wird später bei der Auswertung der Segmentierungsleistung verwendet.

Im nächsten Schritt wird mit SAM 2 eine Maske auf Grundlage der zuvor übergebenen Bounding Boxes erstellt. SAM 2 unterstützt die Integration unterschiedlicher Hier-Modelle, darunter Hiera-Small, Hiera-Medium und Hiera-Large. In dieser Arbeit werden vier Konfigurationen betrachtet: Florence 2 + Hiera-large, Florence 2 + Hiera-small, Grounding DINO + Hiera-large sowie Grounding DINO + Hiera-small (Ren und Shen, 2025; Ren et al. 2024).

2.8 Datensätze für die Evaluation

Um die Leistung von Grounded SAM 2 in Verbindung mit Florence 2, Grounding DINO und den verschiedenen Hiera-Modellen zu evaluieren, wird Grounded SAM 2 auf dem Microsoft COCO (Common Objects in Context) (Lin et al. 2015) Datensatz von 2017 sowie dem RefCOCO und dem RefCOCO+ getestet (Kazemzadeh et al. 2014).

2.8.1 COCO Datensatz

Für die Experimente auf dem COCO Datensatz wird der Validation-Split verwendet. Der COCO Validation-Split enthält 5000 Bilder und insgesamt 36781 zu segmentierende Objekte. Es gibt 80 Kategorien, wie zum Beispiel „person“oder „car“, wobei jedes zu segmentierende Objekt einer bestimmten Kategorie zugeordnet ist. Der Datensatz enthält 11473 kleine Objekte (Fläche $< 32^2$ Pixel), 12825 mittlere Objekte ($32^2 \leq$ Fläche $< 96^2$ Pixel) und 12483 große Objekte (Fläche $\geq 96^2$ Pixel). Die Ground-Truth-Daten sind im JSON-Format strukturiert. Dabei gibt es für jedes Objekt eine Instanz, welche unter anderem die Kategorie-ID, die Segmentierungsmaske sowie Daten für die Bounding Box, die das Objekt einschließt, enthält (Lin et al. 2015).

Abbildung 2.8: Beispielbild aus dem COCO 2017 Datensatz. Die Objektkategorien für dieses Bild sind: „hair drier“, „chair“, „person“

Abbildung 2.9: Beispielbild aus dem COCO 2017 Datensatz. Die Objektkategorien für dieses Bild sind: „orange“, „apple“, „bottle“, „oven“, „toaster“.

In Abbildung 2.8 und 2.9 sind zwei Beispiele aus dem Datensatz, mit den zu den Bildern zugehörigen Kategorien abgebildet. Die Kategorien wurden als Textprompts übernommen.

2.8.2 RefCOCO und RefCOCO+ Datensatz

RefCOCO, RefCOCOg und RefCOCO+ (Kazemzadeh et al. 2014) basieren auf dem COCO Datensatz, ergänzen diesen aber durch unterschiedliche von Menschen erstellte „Referring Expressions “, also genauere Beschreibungen davon, welche Objekte segmentiert werden sollen. Die drei Varianten des RefCOCO Datensatzes; RefCOCO, RefCOCO+ und RefCOCOg unterscheiden sich in der Art und Weise, wie die Referring Expressions formuliert sind sowie der Anzahl der Bilder und Objekte leicht. Die Datensätze enthalten alle drei ausschließlich große Objekte (Fläche $\geq 96^2$ Pixel) ((Yu et al. 2016). Für jedes Objekt sind zwischen einer und drei Referring Expressions vorhanden. Für die Objekte, für die weniger als drei vorhanden waren, wurde die erste Referring Expression

repliziert. Die Experimente konnten so jeweils für die erste, die zweite und die dritte Prompt-Anweisung auf allen Bildern durchgeführt werden. In dieser Arbeit werden der RefCOCO sowie der RefCOCO+ Datensatz verwendet.

RefCOCO

Für die Experimente auf dem RefCOCO Datensatz wird der Test-split, welcher 5000 Referenzen auf insgesamt 4527 Bildern enthält, verwendet. Es gibt 1-3 Beschreibungen für jedes Objekt, welche durchschnittlich aus 3,5 Wörtern bestehen. Neben Beschreibungen, die auf dem Aussehen der Objekte basieren, enthält der Datensatz auch Beschreibungen, die die Position des Objekts beschreiben (Kazemzadeh et al. 2014).

Abbildung 2.10: Beispielbild aus dem RefCOCO-Test-Datensatz. Auf diesem Bild wird ein Objekt Refferenziert, für das es drei Prompts gibt. „cat“, „Cat on right“, „cat, but not in reflection“

Abbildung 2.11: Beispielbild aus dem RefCOCO-Test-Datensatz. Auf diesem Bild wird ein Objekte Refferenziert, für das es drei Prompts gibt. „man“, „man on right“, „man in background“

Auf den Abbildungen 2.12 und 2.11 sind Beispiele aus dem RefCOCO-Datensatz mit den jeweils dazugehörigen Prompts in der Bildunterschrift.

REFCOCO+

Der REFCOCO+ Datensatz wurde auf dem validation-Split getestet. Dieser enthält 3805 Annotationen zu 1500 unterschiedlichen Bildern. Die Objektbeschreibungen bestehen wie auch bei REFCOCO im Schnitt aus 3,5 Wörtern und es gibt ebenfalls 1-3 Beschreibungen für jedes Objekt. Im Gegensatz zu REFCOCO enthalten die Beschreibungen im REFCOCO+ Datensatz keine Informationen darüber, wo im Bild sich das Objekt befindet (Kazemzadeh et al. 2014).

Abbildung 2.12: Beispielbild aus dem RefCOCO+-Validation-Datensatz. Auf diesem Bild werden zwei Objekte Refferenziert, für die es jeweils drei Prompts gibt. „White bowl with vertical stripes“, „white bowl on corner“, „WHITE BOWL NEXT TO RICE“ und „bowl of carrots“, „carrots“, „bowl of carrots“

Abbildung 2.13: Beispielbild aus dem RefCOCO+-Validation-Datensatz. Auf diesem Bild werden zwei Objekte Refferenziert, für die es jeweils drei Prompts gibt. „woman sitting in chair“, „so sorry, tired..lady in white pants“, „adult in chair“ und „The kid“, „baby“, „kid“

Auf den Abbildungen 2.12 und 2.13 sind zwei Beispielbilder aus dem RefCOCO+ Datensatz mit den zugehörigen Prompts in der jeweiligen Bildbeschriftung zu sehen.

2.9 Evaluationsmetriken

Um die Segmentierungsqualität von Grounded SAM bewerten zu können, werden die in der Literatur als Standard geltenden COCO-Evaluationsmetriken verwendet (Bochkovskiy et al. 2020; He et al. 2018; Lee und Park, 2020; L Zhang et al. 2022; Wang et al. 2019). Dabei werden Average Recall (AR) und Average Precision (AP) mit jeweils unterschiedlichen Intersection over Union (IoU) - Werten erhoben. Ergänzend wird auf den Datensätzen RefCOCO und RefCOCO+ zusätzlich die mean Intersection over Union (mIoU) für die Segmentierungsmaske berechnet, da diese in der Referring-Image-Segmentation-Literatur als gängige Metrik gilt (Xiao et al. 2023; Y Zhang et al. 2025), weil in den Datensätzen für jeden Prompt nur ein Objekt segmentiert werden soll. Bei COCO steht eher die Bewertung der gleichzeitigen Detektion und Segmentierung im Vordergrund, weswegen die mIoU-Werte hier nicht separat erhoben werden.

2.9.1 Intersection over Union (IoU)

Durch IoU wird gemessen, wie genau die tatsächliche Segmentierung/BBox mit der vorhergesagten übereinstimmt. IoU berechnet sich durch:

$$\text{IoU} = \frac{\text{Fläche der Überschneidung}}{\text{Fläche der Vereinigung}}$$

Eine IoU von 1 ist mit perfekter Übereinstimmung gleichzusetzen und eine IoU von 0 mit gar keiner. Verschiedene IoU-Schwellenwerte (zum Beispiel 0.50, 0.75) legen fest, wie genau eine BBox oder Maske sein muss, um als „korrekt“ zu gelten (Lucas, 2023). IoU wird zur Berechnung der anderen Metriken erhoben und mIoU ist die mittlere IoU über alle Vorhersagen hinweg.

2.9.2 Average Recall (AR)

Average Recall ist eine Metrik, die angibt, wie viele der tatsächlich in einem Bild vorhandenen Objekte richtig von dem Modell erkannt werden. Recall berechnet sich durch:

$$\text{Recall} = \frac{\text{Anzahl korrekt erkannter Objekte}}{\text{Anzahl aller tatsächlichen Objekte}}$$

Ein Modell mit hohem AR erkennt also viele richtige Objekte. Dabei wird aber nicht beachtet, ob das Modell zusätzlich falsche Objekte detektiert. Zur Berechnung des AR wird über 10 IoU-Schwellenwerte von 0,50 bis 0,95 in Abständen von 0,05 geprüft, wie viele Objekte mit maximal k Vorhersagen pro Bild erkannt werden. Die einzelnen Recall-Werte werden dann über die IoU-Schwellen gemittelt. AR berechnet sich also durch:

$$\text{AR}_{\text{max}=k} = \frac{1}{10} \sum_{t=1}^{10} \text{Recall}_{IoU=t}(k) \tag{2.1}$$

Genauer werden Metriken mit folgenden maximalen Vorhersagen pro Bild erhoben.

- $\mathbf{AR}_{\mathbf{max}=1}$: Average Recall bei maximal 1 Vorhersage pro Bild.
- $\mathbf{AR}_{\mathbf{max}=10}$: Average Recall bei maximal 10 Vorhersagen pro Bild.
- $\mathbf{AR}_{\mathbf{max}=100}$: Average Recall bei maximal 100 Vorhersagen pro Bild.
- $\mathbf{AR}_s$, $\mathbf{AR}_m$, $\mathbf{AR}_l$: Average Recall für für kleine ($area < 32^2$), mittlere ($32^2 < area < 96^2$) und große ($area > 96^2$) Objekte bei maximal 100 Vorhersagen pro Bild.

AR legt Fokus auf die Vollständigkeit der Segmentierung. (Lucas, 2023).

2.9.3 Average Precision (AP)

Um die Average Precision (außerhalb von COCO-Kontexten häufig als *mean Average Precision (mAP)* bezeichnet) zu berechnen, werden zunächst alle Vorhersagen nach ihrem Confidence Score sortiert und jede Vorhersage basierend auf dem vorher bestimmten IoU-Wert mit dem Ground-Truth-Objekt abgeglichen. Mit jedem Vorhersageschritt wird ein kumulierter Recall und ein kumulierter Precision-Wert berechnet. Precision berechnet sich durch:

$$\text{Precision} = \frac{\text{Anzahl korrekt erkannter Objekte}}{\text{Anzahl aller vorhergesagten Objekte}}$$

Eine hohe Precision-Wer bedeutet, dass unter allen erkannten Objekten nur wenige falsch erkannte enthalten sind. Damit ergänzt die Precision den Recall. Aus den resultierenden Precision- und Recall-Paaren wird eine Precision-Recall-Kurve erstellt. Dann erfolgt eine Interpolation über 101 gleichmäßig verteilte Recall-Werte zwischen 0.0 und 1 in 0.01 Schritten. Für jeden dieser Punkte wird die höchste gemessene Precision bei gleichem oder höherem Recall-Wert verwendet. Die AP für eine Objektklasse ergibt sich aus der Fläche unter der Kurve. Um die AP über alle Objektklassen zu berechnen, wird der Durchschnitt aller APs der einzelnen Objektklassen genommen (Padilla et al. 2021). Genauer werden folgende AP-Werte über alle Klassen erhoben.

- **AP**$_{50:95}$: Mittlere Average Precision über 10 IoU-Schwellen (0.50 bis 0.95 in 0.05-Schritten), gemittelt über alle Objekte.
- **AP**$_{50}$: Average Precision bei IoU-Schwelle 0.50.
- **AP**$_{75}$: Average Precision bei IoU-Schwelle 0.75.
- **AP**$_s$, **AP**$_m$, **AP**$_l$: Average Precision für kleine ($Fl\cup che < 32^2 Pixel$), mittlere ($32^2 < Fl\cup che < 96^2 Pixel$) und große ($Fl\cup che > 96^2 Pixel$) Objekte. Hier wird auch über 10 IoU Schwellen (0.50 bis 0.95 in 0.05-Schritten) gemittelt.

$AP_{0,50:0,95}$ ist eine der wichtigsten Metriken, da Recall sowie Precision mit einbezogen werden und sie einen Überblick über die Modellleistung geben kann (Lin et al. 2015; Lucas, 2023).

Hinweis zu Anwendbarkeit der Metriken: Bei bestimmten Datensätzen wie RefCOCO und RefCOCO+, die ausschließlich Objekte enthalten, die nach COCO als groß klassifiziert werden (Fläche $\geq$ 962 Pixel), können Metriken für kleine (AP_s, AR_s) und mittlere (AP_m, AR_m) Objekte nicht bestimmt werden. Dadurch ist außerdem $AP_l = AP_{0,50:0,95}$, wodurch AP_l redundant ist. Da in den beiden Datensätzen maximal 10 Objekte auf einem Bild annotiert sind, entfällt $AR_{max=100}$ ebenfalls. In der Auswertung (siehe Kapitel 4.3 und 4.2) werden diese Metriken daher nicht angegeben.

3 Verwandte Arbeiten

3.1 Unimodale Bildsegmentierungsmodelle

Die Bildsegmentierung umfasst unterschiedliche Aufgaben und Anforderungen, die verschiedene Techniken zur Abgrenzung der Bildbereiche erfordern. Es gibt bereits verschiedene Modelle, die Aufgaben in diesem und ähnlichen Bereichen zu lösen versuchen. Ein Beispiel für so ein Modell ist Masked-attention Mask Transformer. Masked-attention Mask Transformer (Mask2Former) (Cheng et al. 2022), welcher 2022 vorgestellt wurde, wurde entwickelt, um eine einheitliche Architektur für verschiedene Segmentierungsaufgaben bereitzustellen. Durch die Nutzung von "masked attention", einer Methode, die gezielt einzelne Bildregionen verarbeitet, erreicht Mask2Former auf bekannten Datensets wie COCO vergleichsweise gute Werte. Mask2Former ist besonders gut für automatisch lösbare Aufgaben geeignet, da keine zusätzlichen Benutzereingaben wie textuelle Anweisungen erforderlich sind.

3.2 Multimodale Modelle

Neben rein visuellen Modellen gibt es verschiedene multimodale Ansätze, wie Grounded SAM 2, die Bild- und Textinformationen miteinander verknüpfen. Dadurch wird ermöglicht, die verschiedenen Aufgaben interaktiv mit dem Benutzer zu verbinden und so bestimmte Bereiche im Bild über Textbefehle zu segmentieren. Ein weiterer multimodaler Ansatz ist MDETR (Modulated Detection for End-to-End Multi-Modal Understanding) (Kamath et al. 2021), der Text und Bilder als Eingabe entgegennimmt und daraus Bounding Boxes für referenzierte Objekte generiert. Anders als Grounded SAM 2 liefert MDETR keine Masken, sondern dient nur als Objektdetektor.

BLIP (Li et al. 2022) ist neben Florence 2 (Yuan et al. 2021) und Grounding DINO (Caron et al. 2021), die von Grounded SAM 2 genutzt werden, ein weiteres fortschrittliches multimodales Modell. BLIP wurde dafür entwickelt, Bild-Text-Aufgaben, wie zum Beispiel die Erstellung von Bildunterschriften, zu lösen.

Modelle wie diese sind besonders in Bereichen der Bildbearbeitung (Xiao et al. 2023), virtueller Realität (Szeliski, 2022) oder visueller Suche (Santini et al. 2023) von Nutzen.
Kosmos-1 (Huang et al. 2023) von Microsoft ist ein weiteres multimodales Modell, das darauf abzielt, ein umfassendes Verständnis von Text- und Bildinformationen zu schaffen. Das Modell wurde mit einzelnen Texten, Bildern, aber auch zusammengehörenden Bild-Text-Paaren trainiert, wodurch es Bilder und auch Texte versteht, was die Möglichkeit eröffnet, bestimmte Fragen zu Bildern zu beantworten.

3.3 Anwendung von SAM und Grounded SAM

SAM wurde in einigen Studien bereits in unterschiedlichen Anwendungsfällen untersucht (Ji et al. 2024; C Zhang et al. 2023; H Zhou et al. 2024). Ein aktuelles Survey von T Zhou et al. (2024) betont, dass Prompt-basierte Modelle wie SAM es schaffen, eine Vielzahl von Aufgaben mit minimalem Training zu bewältigen und Modelle wie CLIP oder SAM die Bildsegmentierung verbessert haben. Allerdings wurden die in Grounded SAM 2 genutzten Modelle, also SAM (mit den unterschiedlichen Hiera-Modellen) jeweils zusammen mit Florence 2 oder Grounding DINO, bisher nicht systematisch untersucht. In einer Studie von He et al. (2023) wird die Leistungsfähigkeit von SAM auf verschiedenen medizinischen Bildsegmentierungsdatensätzen getestet. Das Ergebnis zeigt, dass SAM ohne Anpassungen oder Fine-Tuning nicht mit Modellen, die speziell für die medizinische Bildverarbeitung entwickelt wurden, mithalten kann. Mumuni und Mumuni (2024) untersuchen die Integration von Grounding DINO und SAM auf verschiedenen Datensätzen und zeigen, dass die Kombination zu einer erhöhten Genauigkeit im Gegensatz zu anderen Ansätzen führt. EVF-SAM (Early Vision-Language Fusion for Text-Prompted Segment Anything Model) ist ein weiterer aktueller Ansatz zur textgesteuerten Segmentierung, der von Y Zhang et al. (2025) vorgestellt wurde und mit SAM arbeitet. Im Gegensatz zu Grounded SAM setzt EVF-SAM auf eine frühe Fusion von Text- und Bildeingaben. Das bedeutet, visuelle und textuelle Informationen werden bereits in den ersten Verarbeitungsschritten gemeinsam betrachtet.

4 Experimente

4.1 Grounded SAM 2 auf dem COCO Datensatz

Zur Evaluation wurde Grounded SAM 2 zunächst auf dem COCO-validation-Datensatz von 2017 (Lin et al. 2015) evaluiert. Der Datensatz enthält eine Mischung aus alltäglichen Objekten und eignet sich gut, um eine erste Einschätzung zur allgemeinen Segmentierungsleistung verschiedener Modulkonfigurationen zu erhalten. Getestet wurden jeweils Kombinationen mit den Komponenten Grounding DINO und Florence 2, in Verbindung mit jeweils Hiera-small sowie Hiera-large als Backbone. Ziel war es, Unterschiede in der Segmentierungsqualität sichtbar zu machen.

Zur Veranschaulichung wurden einige beispielhafte Segmentierungsergebnisse ausgewählt. Die verwendeten Beispielbilder stammen ursprünglich aus dem COCO-Datensatz (Lin et al. 2015). Um visuelle Unterschiede möglichst direkt vergleichbar zu machen, wird jeweils dasselbe Bild mit allen vier Modell-Backbone-Kombinationen dargestellt. In Abbildung 4.1 und 4.2 sind die Ergebnisse von Grounding DINO in Kombination mit Hiera-large und Hiera-small zu sehen und die Ergebnisse von Florence 2 mit Hiera-large bzw. Hiera-small sind in Abbildung 4.3 und 4.4 abgebildet.

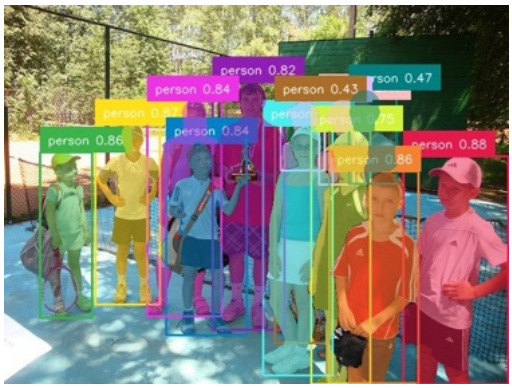

Abbildung 4.1: Grounded SAM 2 mit Grounding DINO + Hiera-large auf einem Bild mit verschiednenen Personen.

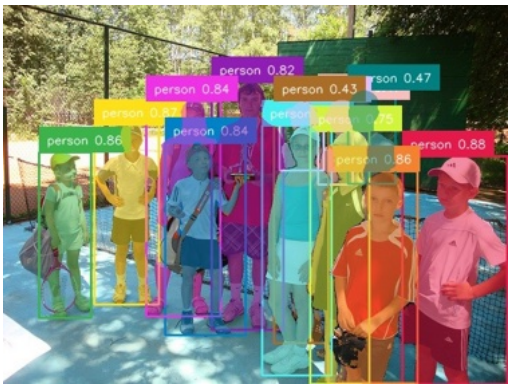

Abbildung 4.2: Grounded SAM 2 mit Grounding DINO + Hiera-small auf einem Bild mit verschiednenen Personen.

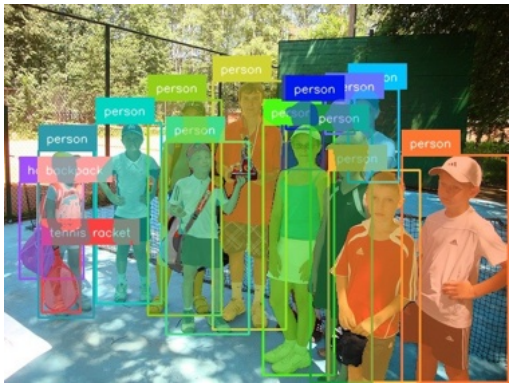

Abbildung 4.3: Grounded SAM 2 mit Florence 2 + Hiera-large auf einem Bild mit verschiedenen Personen.

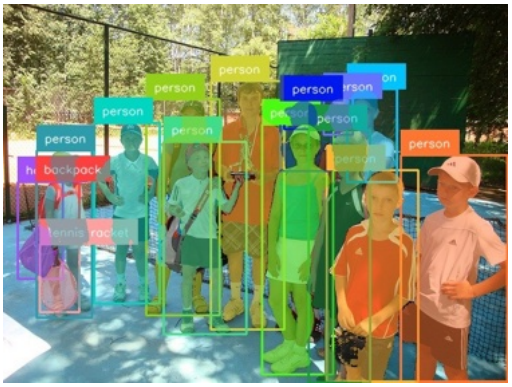

Abbildung 4.4: Grounded SAM 2 mit Florence 2 + Hiera-small auf einem Bild mit verschiedenen Personen.

Beim Vergleich der vier Masken fällt auf, dass sich die Segmentierungsergebnisse sowohl hinsichtlich der erfassten Konturen als auch der insgesamt erkannten Objekte unterscheiden. Besonders deutlich wird dies bei den Variante mit Grounding DINO (Abbildung 4.1 und 4.2). Einige Objekte, wie der Tennisschläger oder der Rucksack des Kindes auf der linken Seite, wurden gar nicht erkannt. Außerdem sind ebenfalls bei dem Kind auf der linken Seite leichte Unterschiede in der Segmentierung zwischen Hiera-large (Abbildung 4.1) und Hiera-small (Abbildung 4.2) zu erkennen.

Florence 2 erkennt dagegen insgesamt mehr Objekte, insbesondere die, die von etwas anderem teilweise verdeckt werden. Auch hier ist beispielsweise bei dem Tennisschläger auf der linken Seite zu erkennen, dass die Kombination aus Florence 2 mit Hiera-large in Abbildung 4.3 die Konturen des Tennisschlägers etwas präziser erfasst als die Variante mit Hiera-small (Abbildung 4.4). Ähnliche Beobachtungen konnten auch bei weiteren Bildern gemacht werden, wobei natürlich nicht alle Bilder im Detail analysiert werden konnten. Die Beispiele sollen daher einen repräsentativen, aber nicht abschließenden Eindruck vermitteln.

4.1.1 Segmentierungsergebnisse COCO

In diesem Abschnitt werden die Segmentierungsergebnisse der Experimente mit Florence 2, Grounding DINO, Hiera-large und Hiera-small vorgestellt. Tabelle 4.1 zeigt die Average Precision (siehe Abschnitt 2.9.3) und Tabelle 4.2 die Average Recall (siehe Abschnitt 2.9.2) Ergebnisse.

Modell	Backbone	$AP_{50\text{-}95}$	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
Florence 2	Hiera-large	39.8	58.2	43.5	22.7	43.5	59.3
	Hiera-small	38.7	57.3	41.9	22.3	42.8	57.0
Grounding DINO	Hiera-large	38.8	56.6	42.4	22.5	41.8	56.7
	Hiera-small	37.7	55.9	41.2	22.0	40.6	55.1

Tabelle 4.1: Die Tabelle zeigt die Average Precision (AP) Ergebnisse der Segmentierung mit Florence 2 und Grounding DINO in Kombination mit Hiera-large bzw. Hiera-small auf dem COCO-2017-Datensatz. Alle dargestellten Zahlen geben Prozentwerte an.

Die Ergebnisse in Tabelle 4.1 zeigen, dass Florence 2 in Kombination mit Hiera-large die höchste $AP_{50\text{-}95}$ von 39,8% erreicht. Im Vergleich dazu liegt Grounding DINO mit Hiera-large leicht darunter,

bei 38,8%. Auch bei großen Objekten (AP_l) erzielt Florence 2 mit Hiera-large den besten Wert von 59,3%. Grounding DINO zusammen mit Hiera-small weist insgesamt etwas niedrigere Werte auf.

Modell	Backbone	$AR_{max=1}$	$AR_{max=10}$	$AR_{max=100}$	AR_s	AR_m	AR_l
Florence 2	Hiera-large	34.1	46.8	47.9	29.5	53.6	66.5
	Hiera-small	33.2	45.6	46.9	29.1	52.7	64.7
Grounding DINO	Hiera-large	31.5	43.6	44.4	26.8	48.8	61.9
	Hiera-small	30.8	42.5	43.4	26.4	47.4	60.6

Tabelle 4.2: Die Tabelle zeigt die Average Recall (AR) Ergebnisse für die Segmentierung mit Florence 2 und Grounding DINO in Kombination mit Hiera-large bzw. Hiera-small auf dem COCO-2017-Datensatz. Alle dargestellten Zahlen geben Prozentwerte an.

Auch in Bezug auf die Recall-Metrik erreicht Florence 2 mit Hiera large, wie in Tabelle 4.2 zu sehen ist, die höchsten Werte. Mit $AR_{max=100}$ = 47,9% liegt Florence 2 über den $AR_{max=100}$ Werten von Grounding DINO mit Hiera-large (44,4%). Insbesondere bei größeren Objekten (AR_l) weist Florence 2 mit 66,5% einen besseren Wert als Grounding DINO mit Hiera-large (61,9%) auf. Die Kombination aus Grounding DINO mit Hiera-small zeigt auch bei den AR Werten wieder die schwächste Leistung mit einem $AR_{max=100}$ von 43,4% und einem AR_l von 60,6%. Florence 2 ist hier mit beiden Backbones besser als die Kombinationen mit Grounding DINO.

Insgesamt zeigen die Ergebnisse, dass Florence 2 zusammen mit Hiera-large im Vergleich zu den anderen Kombinationen die beste Leistung erbringt. Insbesondere bei der Segmentierung großer und mittelgroßer Objekte, aber auch bei den Gesamtergebnissen liefert diese Konfiguration immer die besten Ergebnisse. Grounding DINO erzielt ebenfalls gute Resultate, ist aber fast immer leicht schlechter als Florence 2. Die Unterschiede zwischen Hiera-small und Hiera-large deuten darauf hin, dass der gewählte Backbone ebenfalls einen kleinen Einfluss auf die Ergebnisse hat. Hiera-large hat fast immer bessere Ergebnisse als Hiera-small. Außerdem weisen die AP-Werte darauf hin, dass Hiera-large besonders für das Erreichen höherer Genauigkeit sowie das Segmentieren kleinerer Objekte geeignet ist.

4.1.2 Bounding Box (BBox) Ergebnisse COCO

In diesem Abschnitt werden die Ergebnisse der Bounding-Boxen (BBoxen) von Grounded SAM 2 mit den Modellen Florence 2 und Grounding DINO in Kombination mit den Backbones Hiera-large und Hiera-small vorgestellt. In Tabelle 4.3 sind die Average Precision (siehe Abschnitt 2.9.3) und in Tabelle 4.3 die Average Recall (siehe Abschnitt 2.9.2) Ergebnisse präsentiert.

Modell	Backbone	$AP_{50\text{-}95}$	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
Florence 2	Hiera-large	45,3	60,2	49,3	28,5	49,5	64,4
	Hiera-small	45,5	60,0	49,2	29,2	49,9	63,1
Grounding DINO	Hiera-large	45,0	58,2	48,8	28,9	48,0	63,2
	Hiera-small	44,7	57,9	48,6	28,6	47,6	62,6

Tabelle 4.3: Die Tabelle zeigt die Average Precision (AP) Ergebnisse für die BBoxen mit den Modellen Grounding DINO und Florence 2 in Kombination mit den Backbones Hiera-large und Hiera-small. Die Experimente wurden im Rahmen der Evaluation von Grounded SAM 2 auf dem COCO 2017 Datensatz durchgeführt. Alle dargestellten Zahlen geben Prozentwerte an.

Die Bounding Box-AP-Ergebnisse in Tabelle 4.3 zeigen, dass die Unterschiede zwischen den Modellvarianten insgesamt gering ausfallen. Florence 2 erreicht sowohl mit Hiera-large als auch mit Hiera-small ähnliche Werte für $AP_{0.50-0.95}$ von 45,3% bzw. 45,5%. Grounding DINO mit Hiera-large schneidet mit 45,0% etwas schlechter ab und Grounding DINO mit Hiera-small erreicht mit 44,7% den schlechtesten $AP_{0.50-0.95}$ Wert. Auffällig ist, dass Florence 2 mit Hiera-small bei kleinen und mittelgroßen Objekten geringfügig bessere Ergebnisse erreicht als die Kombination mit Hiera-large. Bei großen Objekten erzielt Florence 2 mit Hiera-large hingegen (AP_l) mit 64,4% die beste Leistung.

Modell	Backbone	$AR_{max=1}$	$AR_{max=10}$	$AR_{max=100}$	AR_s	AR_m	AR_l
Florence 2	Hiera-large	38,8	53,5	54,9	34,2	59,7	75,9
	Hiera-small	38,9	54,2	55,9	36,6	60,9	75,5
Grounding DINO	Hiera-large	36,4	50,7	51,7	32,8	54,6	71,2
	Hiera-small	36,3	50,6	51,7	32,8	54,6	71,1

Tabelle 4.4: Die Tabelle zeigt die Average Recall (AR) Ergebnisse für die BBoxen mit den Modellen Grounding DINO und Florence 2 in Kombination mit den Backbones Hiera-large und Hiera-small. Die Experimente wurden im Rahmen der Evaluation von Grounded SAM 2 auf dem COCO 2017 Datensatz durchgeführt. Alle dargestellten Zahlen geben Prozentwerte an.

Die AR Ergebnisse für die BBoxen in Tabelle 4.4 zeigen ein ähnliches Bild wie bei der Segmentierung. Florence 2 erreicht insgesamt die höchsten AR Werte. Bei großen Objekten (AR_l) liegt Florence 2 zusammen mit Hiera-large mit 75,9% über den Grounding DINO-Varianten (mit Hiera-large (71,2%) und mit Hiera-small (71,1%)). Bei kleinen und mittleren Objekten (AR_s, AR_m) erreicht Florence 2 mit Hiera-small die besten Werte und übertrifft somit die Variante mit Hiera-large. Insbesondere wenn mehrere oder größere Objekte auf einem Bild sind, schneidet Florence 2 besser ab als Grounding DINO. Auch bei $AR_{max=100}$ übertrifft Florence 2 in Kombination mit Hiera-large mit 54,9% die anderen Varianten leicht.

Auch bei der Bounding Box Erkennung erzielt Florence 2 in Kombination mit beiden Backbones durchgehend bessere Resultate als Grounding DINO. Besonders auffällig ist auch hier die hohe Erkennungsrate großer Objekte, teils mit fast 76% Recall, was für Aufgaben mit dominanten Bildelementen von Vorteil sein kann. Die Unterschiede zwischen Hiera-large und Hiera-small sind bei Florence 2 eher gering, wobei Hiera-small bei kleinen Objekten besser abschneidet. Der Grund dafür könnte sein, Hiera-large Details früher gröber zusammenfasst, wodurch kleinere Details verloren gehen könnten und Hiera-large so besser für größere Objekte geeignet ist. Dieser Effekt wurde ebenfalls in der Arbeit „Rethinking the backbone architecture for tiny object detection“ von Ning et al. (2023) bestätigt. Grounding DINO mit Hiera-large ist konsistent leicht besser als mit Hiera-small.

4.2 Grounded SAM 2 auf dem RefCOCO Datensatz

Im Anschluss wurde Grounded SAM 2 auf dem RefCOCO-Test-Datensatz (Kazemzadeh et al. 2014) getestet, der zusätzlich sprachliche Beschreibungen beinhaltet, mit denen bestimmte Objekte im Bild referenziert werden. Im Gegensatz zum COCO-Datensatz, bei dem alle Objekte einer Klasse im Bild segmentiert werden sollen, geht es hier immer um ein spezifisches Objekt, wodurch sich der Datensatz zum Testen der textgesteuerten Komponenten von Grounded SAM 2 besonders gut eignet. Die Evaluierung erfolgte auf dem Test-Split und es wurden erneut vier Systemvarianten betrachtet: Grounding DINO in Kombination mit Hiera-large bzw. Hiera-small sowie Florence 2 mit denselben Backbones.
Der RefCOCO-Datensatz enthält nach COCO-Einordnung ausschließlich große Objekte und nicht mehr als 10 Objekte auf einem Bild, weshalb einige Metriken, die zuvor bei dem COCO-Datensatz genutzt worden sind, überflüssig werden (Lin et al. 2015).
($AR_{max=100}$, AR_s, AR_m, AR_l, AP_s, AP_m und AP_l sind also überflüssig)

Die Abbildungen 4.5 bis 4.8 zeigen exemplarische Ergebnisse auf einem Bild mit mehreren Kühen. Die Referenz auf das zu segmentierende Objekt lautete „bottom - right bull“. Das Bild stammt ursprünglich aus dem COCO-Datensatz (Lin et al. 2015)

Abbildung 4.5: Grounded SAM 2 mit Grounding DINO + Hiera-large auf einem Bild mit mehreren Kühen aus dem RefCOCO Datensatz. Hier sollte die Kuh unten rechts segmentiert werden.

Abbildung 4.6: Grounded SAM 2 mit Grounding DINO + Hiera-small auf einem Bild mit mehreren Kühen aus dem RefCOCO Datensatz. Hier sollte die Kuh unten rechts segmentiert werden.

Abbildung 4.7: Grounded SAM 2 mit Florence 2 + Hiera-large auf einem Bild mit mehreren Kühen aus dem RefCOCO Datensatz. Hier sollte die Kuh unten rechts segmentiert werden.

Abbildung 4.8: Grounded SAM 2 mit Florence 2 + Hiera-small auf einem Bild mit mehreren Kühen aus dem RefCOCO Datensatz. Hier sollte die Kuh unten rechts segmentiert werden.

Keine der Varianten hat tatsächlich die referenzierte Kuh unten rechts segmentiert. Stattdessen wählten die Modelle jeweils eine andere Kuh auf dem Bild aus, in allen Fällen eine zentral oder generell auffällig platzierte. Die textuelle Information „unten rechts“ („bottom right“) wurde also nicht korrekt von den Modellen umgesetzt. Es wurde zwar immer eine Kuh segmentiert, jedoch stimmen die Position von Segmentierung und Beschreibung nicht überein. Grounding DINO hat mit beiden Backbone-Varianten anstelle der Kuh unten rechts die wahrscheinlich auffälligste Kuh in der Mitte des Bildes segmentiert. Abgesehen davon sind bezüglich der Genauigkeit der Konturen leichte Unterschiede zu sehen. Die Kombination mit Hiera-small hat einen der Hinterhufen nicht richtig erkannt, während die Kombination mit Hiera-large einen kleinen Teil am Kopf der Kuh nicht segmentiert hat. Florence 2 hat anstatt der Kuh unten rechts die Kuh unten links segmentiert, auch hier sieht man zwischen den beiden Hiera-Varianten leichte Unterschiede. In der Konfiguration mit Hiera-small wurde zusätzlich noch ein Bein einer anderen Kuh segmentiert.

Ähnliche Muster ließen sich auch auf anderen Bildern beobachten. Räumliche Begriffe wie „left“, „right“, „top“ oder „bottom“ scheinen den Modellen oft Schwierigkeiten zu bereiten. Es gelingt zwar, die semantisch richtige Objektklasse zu identifizieren (wie hier z.B. „Kuh“), allerdings scheinen sich die Modelle, anstatt sich konsequent an der beschriebenen Position zu orientieren, häufig auf andere Bildmerkmale, wie Größe oder Position, im Zentrum zu reagieren.

4.2.1 Segmentierungsergebnisse RefCOCO

Im Folgenden werden die numerischen Segmentierungsergebnisse auf dem RefCOCO-Datensatz dargestellt. In Tabelle 4.5 sind die Average Precision- und mIoU-Ergebnisse (siehe Abschnitte 2.9.3 und 2.9.1) und in Tabelle 4.6 die Average Recall-Ergebnisse (siehe Abschnitt 2.9.2) aufgeführt.

Modell	Backbone	$AP_{50\text{-}95}$	AP_{50}	AP_{75}	mIoU
Florence 2	Hiera-large	42,4	54,5	46,8	48,9
	Hiera-Small	43,7	54,3	46,4	48,6
Grounding DINO	Hiera-large	33,6	43,1	35,0	47,9
	Hiera-small	33,0	43,0	36,9	47,6

Tabelle 4.5: Die Tabelle zeigt die Average Precision (AP) Ergebnisse für die Segmentierung mit den Backbones Grounding DINO und Florence 2 in Kombination mit den Modellen Hiera-large und Hiera-small. Die Experimente wurden im Rahmen der Evaluation von Grounded SAM 2 auf dem RefCOCO Test Datensatz durchgeführt. Alle Angaben sind in Prozent.

Die Segmentierungsergebnisse (dargestellt in Tabelle 4.5) des RefCOCO Datensatzes zeigen, dass Florence 2 mit beiden Backbone-Varianten Grounding DINO deutlich übertrifft. Florence 2 in Kombination mit Hiera-small erreicht mit 43,7% den höchsten AP_{50-95} Wert unter allen getesteten Kombinationen. Grounding DINO erreicht in der besten Variante (Hiera-large) nur 33,6% AP_{50-95} und bleibt damit unter dem Niveau von Florence 2. Bei Florence 2 sowie Grounding DINO fallen die Unterschiede zwischen den beiden Hiera-Varianten eher gering aus. In den meisten Fällen ist Hiera-large noch etwas besser. Die mIoU zeigt nur kleine Unterschiede zwischen den Varianten auf.

Modell	Backbone	$AR_{max=1}$	$AR_{max=10}$
Florence 2	Hiera-large	41,6	42,8
	Hiera-Small	41,0	42,1
Grounding DINO	Hiera-large	33,0	36,2
	Hiera-small	32,4	33,7

Tabelle 4.6: Die Tabelle zeigt die Average Recall (AR) Ergebnisse für die Segmentierung mit den Backbones Grounding DINO und Florence 2 in Kombination mit den Modellen Hiera-large und Hiera-small. Die Experimente wurden im Rahmen der Evaluation von Grounded SAM 2 auf dem RefCOCO Datensatz durchgeführt. Alle Angaben sind in Prozent.

Auch bei den Recall-Werten in Tabelle 4.6 zeigt sich ein ähnliches Bild. Florence 2 erzielt wieder mit beiden Backbone-Varianten höhere Ergebnisse als Grounding DINO. Mit einem $AR_{max=10}$ von 42,8% kommt Florence 2 mit Hiera-large auf den besten Wert. Grounding DINO liegt in seiner besten Konfiguration (zusammen mit Hiera-large) bei 36,2%. Die Kombination aus Grounding DINO mit Hiera-small erzielt die niedrigsten Ergebnisse.

Die Segmentierungsergebnisse auf dem RefCOCO-Datensatz bestätigen insgesamt die Tendenz, die bereits bei COCO zu beobachten war. Florence 2 in Kombination mit Hiera-large zeigt erneut die besten Segmentierungsergebnisse, während die beiden Varianten mit Grounding DINO deutlich schlechter abschneiden. Die Unterschiede zwischen Hiera-small und Hiera-large bleiben zwar gering, lassen aber erkennen, dass ein leistungsfähigerer Backbone bei großen Objekten zu präziseren Segmentierungen führen kann.

4.2.2 Bounding Box Ergebnisse RefCOCO

Nach der Analyse der Segmentierungsergebnisse wird im nächsten Schritt untersucht, wie zuverlässig die Varianten beim Vorschlagen von Bounding Boxes arbeiten. Die nachfolgenden Tabellen zeigen die Ergebnisse für Average Precision, mIoU (Tabelle 4.7) und Average Recall (Tabelle 4.8) der vier getesteten Modell-Backbone-Konfigurationen.

Modell	Backbone	$AP_{50\text{-}95}$	AP_{50}	AP_{75}
Florence 2	Hiera-large	48,8	56,8	50,7
	Hiera-Small	48,7	56,7	50,6
Grounding DINO	Hiera-large	38,9	45,2	41,0
	Hiera-small	39,0	45,8	41,0

Tabelle 4.7: Die Tabelle zeigt die Average Precision (AP) Ergebnisse für die BBoxen mit den Backbones Grounding DINO und Florence 2 in Kombination mit den Modellen Hiera-large und Hiera-small. Die Experimente wurden im Rahmen der Evaluation von Grounded SAM 2 auf dem RefCOCO Datensatz durchgeführt. Alle Angaben sind in Prozent.

Bei Betrachtung der Average-Precision-Werte in Tabelle 4.7 zeigt sich, dass Florence 2 in Kombination mit Hiera-large die besten Ergebnisse erzielt. Bei AP_{50-95} erreicht diese Konfiguration einen Wert von 48,8% und liegt damit etwas über Florence 2 mit Hiera-small (48,7%). Grounding DINO folgt mit 38,9% (Hiera-large) und 39% (Hiera-small). Auch in den anderen Metriken AP_{50} sowie AP_{75} bleibt Florence 2 leicht vorn.

Modell	Backbone	$AR_{max=1}$	$AR_{max=10}$
Florence 2	Hiera-large	48,1	49,4
	Hiera-Small	48,0	49,4
Grounding DINO	Hiera-large	38,2	39,7
	Hiera-small	38,2	40,0

Tabelle 4.8: Die Tabelle zeigt die Average Recall (AR) Ergebnisse für die BBoxen mit den Backbones Grounding DINO und Florence 2 in Kombination mit den Modellen Hiera-large und Hiera-small. Die Experimente wurden im Rahmen der Evaluation von Grounded SAM 2 auf dem RefCOCO 2017 Datensatz durchgeführt. Alle Angaben sind in Prozent.

Ein ähnliches Bild wie in Tabelle 4.7 zeigt sich bei den Recall-Ergebnissen in Tabelle 4.8. Florence 2 mit Hiera-large und Hiera-small erreichen beide bei $AR_{max=10}$ einen Wert von 49,4%. Grounding DINO bleibt mit 39,7% (Hiera-large) und 40% (Hiera-small) darunter. Die Unterschiede zwischen Hiera-small und Hiera-large sind hier sehr gering.

Die Bounding Box Ergebnisse auf dem RefCOCO Datensatz fügen sich erneut in das bis jetzt beobachtete Bild ein. Florence 2 liefert konsistent präzisere und vollständigere BBox-Vorhersagen als Grounding DINO. Der Einfluss des eingesetzten Hiera-Modells bleibt bei den Bounding Boxen dabei relativ subtil.

4.3 Grounded SAM 2 auf dem RefCOCO+ Datensatz

Im letzten Evaluationsschritt wurde Grounded SAM 2 auf dem RefCOCO+-Datensatz (Kazemzadeh et al. 2014) getestet. Dadurch, dass sich der Datensatz im Gegensatz zum RefCOCO-Datensatz ausschließlich auf visuelle und nicht auf räumliche Hinweise bezieht, ändern sich die Anforderungen an die semantische Erfassung der Bildinhalte erneut. Auch der RefCOCO+-Datensatz enthält ausschließlich Objekte, die von COCO als große Objekte eingestuft werden, wodurch auch hier einige Metriken überflüssig sind.
(Es fallen $AR_{max=100}$, AR_s, AR_m, AR_l, AP_s, AP_m und AP_l weg) Wie zuvor wurden alle vier Modell-Backbone-Varianten getestet.
Auf den folgenden Abbildungen sind zur Veranschaulichung Segmentierungsergebnisse auf demselben Beispielbild dargestellt, welches ursprünglich aus dem COCO-Datensatz stammt (Lin et al. 2015).

Abbildung 4.9: Grounded SAM 2 mit Grounding DINO + Hiera-large auf einem Bild mit mit drei Katzen. Hier sollte die schwarze Katze auf der rechten Seite segmentiert werden.

Abbildung 4.10: Grounded SAM 2 mit Grounding DINO + Hiera-small auf einem Bild mit mit drei Katzen. Hier sollte die schwarze Katze auf der rechten Seite segmentiert werden.

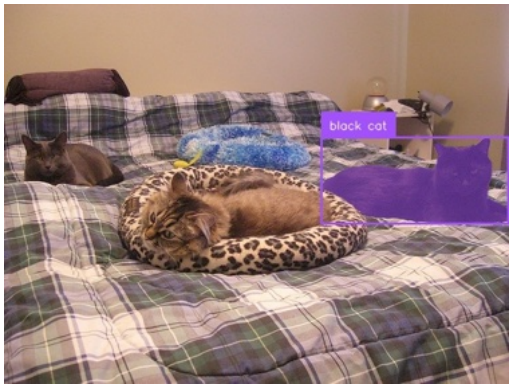

Abbildung 4.11: Grounded SAM 2 mit Florence 2 + Hiera-large auf einem Bild mit mit drei Katzen. Hier sollte die schwarze Katze auf der rechten Seite segmentiert werden.

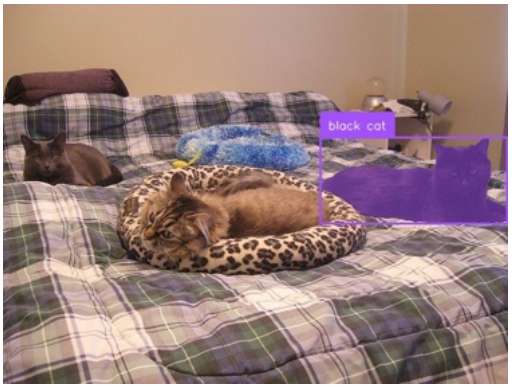

Abbildung 4.12: Grounded SAM 2 mit Florence 2 + Hiera-small auf einem Bild mit drei Katzen. Hier sollte die schwarze Katze auf der rechten Seite segmentiert werden.

Bei der Betrachtung der Ergebnisse der vier Modellvarianten mit dem Prompt „black cat“ fallen deutliche Unterschiede auf. Grounding DINO identifiziert in den Abbildungen 4.9 und 4.10 in beiden Konfigurationen (Hiera-small und Hiera-large) dieselbe falsche Katze im Bild. Die gewählte Bounding

Box liegt jeweils auf einer anderen Katze, die sich mehr im Vordergrund befindet, aber heller als die gesuchte ist und nicht zu der sprachlichen Beschreibung passt. Auch auf einigen anderen Bildern zeigte Grounding DINO Probleme, insbesondere wenn mehrere Adjektive einem Objekt sinnvoll zugeordnet werden sollten. Es scheint so, als würde das Modell eher darauf achten, welches Objekt (hier die Katze) sich im Zentrum oder gut sichtbar im Bild befindet.

Im Gegensatz dazu erkennt Florence 2 in beiden Backbone-Varianten die korrekte schwarze Katze. Florence 2 hatte aber in anderen Beispielen ebenfalls Probleme, das richtige Objekt zu identifizieren. Die Segmentierungen, die anschließend auf Basis der jeweiligen Bounding Box erzeugt wurden, zeigen zwischen den Varianten Hiera-small und Hiera-large keinen nennenswerten sichtbaren Unterschied. Auch wenn häufig nicht das richtige Objekt erkannt wird, wählen die Konfigurationen trotzdem meistens nur ein Objekt aus. Sie verstehen also, dass nur ein Objekt beschrieben wird, haben aber Probleme, diese Beschreibung korrekt zu interpretieren.

4.3.1 Segemntierungsergebnisse RefCOCO+

Die folgenden Tabellen zeigen die Ergebnisse der Segmentierungsauswertung. Tabelle 4.9 enthält die Average-Precision- und die mIoU-Werte (siehe Abschnitt 2.9.3 und 2.9.1) und Tabelle 4.10 die zugehörigen Average-Recall-Werte (siehe Abschnitt 2.9.2).

Modell	Backbone	$AP_{50\text{-}95}$	AP_{50}	AP_{75}	mIoU
Florence 2	Hiera-large	37,1	46,4	40,3	45,2
	Hiera-Small	36,5	48,9	40,6	45,0
Grounding DINO	Hiera-large	31,9	41,7	35,2	46,6
	Hiera-small	31,3	41,7	36,6	44,7

Tabelle 4.9: Die Tabelle zeigt die Average Precision (AP) Ergebnisse für die Segmentierung mit den Backbones Grounding DINO und Florence 2 in Kombination mit den Modellen Hiera-large und Hiera-small. Die Experimente wurden im Rahmen der Evaluation von Grounded SAM 2 auf dem RefCOCO+ Validation Datensatz durchgeführt.

Tabelle 4.9 zeigt, dass Florence 2 die besten Gesamtwerte erreicht. Bei $AP_{50\text{-}95}$ kommen die beiden Kombinationen mit Florence auf 37,1% (Hiera-large) und 36,5% (Hiera-small). Die Kombinationen mit Grounding DINO schneiden etwas schlechter ab, bleiben aber im vergleichbaren Bereich. Die Unterschiede zwischen den Hiera-Varianten sind sehr gering, auffällig ist aber, dass teilweise Hiera-small bei AP_{50} und AP_{75} sogar besser abschneidet als Hiera-large. Die mIoU-Werte liegen bei allen Kombinationen relativ dicht beieinander.

Modell	Backbone	$AR_{max=1}$	$AR_{max=10}$
Florence 2	Hiera-large	36,8	37,4
	Hiera-Small	36,2	36,8
Grounding DINO	Hiera-large	31,4	32,5
	Hiera-small	30,7	30,8

Tabelle 4.10: Die Tabelle zeigt die Average Recall (AR) Ergebnisse für die Segmentierung mit den Backbones DINO und Florence 2 in Kombination mit den Modellen Hiera-large und Hiera-small. Die Experimente wurden im Rahmen der Evaluation von Grounded SAM 2 auf dem RefCOCO+ Datensatz durchgeführt.

Ein ähnliches Bild ergibt sich in Tabelle 4.10 für die Recall-Werte. Florence 2 mit Hiera-large liegt hier in allen Metriken leicht vorn, gefolgt von Florence 2 mit Hiera-small. Die Unterschiede zwischen den Varianten mit Hiera-small und Hiera-large bleiben wie bereits in den vorangegangenen Auswertungen gering. Hier bleibt allerdings Hiera-large konstant etwas besser als Hiera-small.

Die Segmentierungsergebnisse auf dem RefCOCO+ Datensatz bestätigen den Trend der vorherigen Abschnitte. Florence 2 liefert insgesamt auch hier die stabilsten Ergebnisse, kann sich also besser als Grounding DINO auf visuelle Merkmale von Objekten stützen. Die Unterschiede zwischen Hiera-small und Hiera-large bleiben auch hier gering, wobei Hiera-large meistens insgesamt leicht bessere Ergebnisse liefert und.

4.3.2 Bounding Box Ergebnisse RefCOCO+

Die Auswertungen für die Bounding-Boxen (BBoxen) auf dem RefCOCO+ Datensatz sind in Tabelle 4.11 und 4.12 dargestellt.

Modell	Backbone	$AP_{50\text{-}95}$	AP_{50}	AP_{75}
Florence 2	Hiera-large	43,1	51,5	44,1
	Hiera-Small	43,0	51,5	44,0
Grounding DINO	Hiera-large	37,0	44,1	38,2
	Hiera-small	37,0	44,1	38,2

Tabelle 4.11: Die Tabelle zeigt die Average Precision (AP) Ergebnisse für die BBoxen mit Grounding DINO und Florence 2 in Kombination mit den Backbones Hiera-large und Hiera-small. Die Experimente wurden im Rahmen der Evaluation von Grounded SAM 2 auf dem RefCOCO+ Datensatz durchgeführt.

In Tabelle 4.11 ist zu sehen, dass Florence 2 mit Hiera-large bei den Bounding Boxen die besten Ergebnisse liefert. Mit einem AP_{50-95} von 43,1% liegt diese Konfiguration vor den anderen. Die Unterschiede zu Hiera-small sind bei Florence 2 sowie bei Grounding DINO extrem leicht ausgeprägt. Grounding DINO schneidet mit beiden Hiera-Varianten mit einem AP_{50-95} von 37% insgesamt schlechter ab als Florence 2.

Modell	Backbone	$AR_{max=1}$	$AR_{max=10}$
Florence 2	Hiera-large	42,8	43,5
	Hiera-Small	42,7	43,5
Grounding DINO	Hiera-large	36,4	37,7
	Hiera-small	36,4	37,7

Tabelle 4.12: Die Tabelle zeigt die Average Recall (AR) Ergebnisse für die BBoxen mit Groinding DINO und Florence 2 in Kombination mit den Backbones Hiera-large und Hiera-small. Die Experimente wurden im Rahmen der Evaluation von Grounded SAM 2 auf dem RefCOCO+ Datensatz durchgeführt.

Ein ähnliches Muster zeigt sich in Tabelle 4.12 für die Recall-Werte. Florence 2 mit Hiera-large und Hiera-small erzielen mit 43,5%($AR_{max=10}$) den besten Wert, Grounding DINO erreicht einen ($AR_{max=10}$) von 37,7%. Zwischen den Hiera-Varianten sind kaum oder nur kleine Unterschiede.

Auch bei den BBox-Ergebnissen auf dem RefCOCO+-Datensatz setzt sich der Trend zugunsten von Florence 2 fort. Mit stabilen AP- und AR-Werten um die 43% zeigt Florence eine konstante Leistung, die sich durch die Backbone-Wahl kaum beeinflussen lässt. Grounding DINO bleibt dagegen erkennbar zurück und zeigt zwischen den Backbone-Varianten ebenfalls kaum Unterschiede.

4.4 Rechenzeit der Experimente

Zur Bewertung der Einsetzbarkeit der Modell-Backbone-Kombinationen wurde die Zeiten gemessen, die die Kombinationen gebraucht haben, um alle Bilder in den Datensätzen zu bearbeiten. Tabelle 4.13 zeigt die Rechenzeiten in Minuten.

Modell	Backbone	COCO	RefCOCO	RefCOCO+
Florence 2	Hiera-large	292	417	313
	Hiera-Small	261	318	238
Grounding DINO	Hiera-large	198	240	187
	Hiera-small	155	162	108

Tabelle 4.13: Rechenzeit der Modell-Backbone-Kombinationen auf den Datensätzen in Minuten.

Die Ergebnisse zeigen deutliche Unterschiede zwischen den Varianten. Modell sowie Backbone scheinen einen erheblichen Einfluss auf die Rechenzeit zu haben, wobei das Modell entscheidender ist. Florence 2 braucht für den COCO-Datensatz 292 Minuten und die Variante mit Hiera-small etwa 30 Minuten weniger. Grounding DINO braucht mit dem jeweiligen Hiera-Modell deutlich weniger Zeit als die Varianten mit Florence 2. Besonders effizient zeigt sich die Variante aus Grounding DINO mit Hiera-small, die auf allen Datensätzen mit weniger als 170 Minuten auskommt. Das entspricht einer Zeitersparnis von 47 bis 65% im Vergleich zur langsamsten Variante (Florence mit Hiera-large). Die Wahl des Modells sollte also nicht nur hinsichtlich der Segmentierungsleistung, sondern auch im Hinblick auf die Rechenzeit gut überlegt sein.

4.5 Fazit zu den Experimenten

Die Experimente auf den drei Datensätzen COCO, RefCOCO und RefCOCO+ zeigen, dass sowohl die Wahl des Modells als auch die Wahl des Backbones einen Einfluss auf das Segmentierungsergebnis haben können. Florence 2 erreicht hinsichtlich der Average Recall und Average Precision über alle Objektgrößen hinweg meistens die besten Werte. Grounding DINO ist aber insbesondere bei einfachen ein-Wort-langen Prompts, wie sie in den Experimenten auf dem COCO-Datensatz verwendet wurden, leistungsmäßig sehr nah an den Ergebnissen von Florence 2 und kann in Kombination mit Hiera-large teilweise sogar bessere Ergebnisse erzielen als die Florence 2-Variante mit Hiera-small. Bei kleineren Objekten schneidet Florence 2 mit Hiera-small am besten ab. Bei längeren Textprompts schneidet Florence 2 mit beiden Backbone-Varianten deutlich besser ab als Grounding DINO. Der Backbone scheint hier eine weniger relevante Rolle zu spielen. Hinsichtlich der Rechenzeit ist Grounding DINO mit Hiera-small mit deutlichem Abstand am effizientesten. Florence 2 braucht mit Hiera-large auf allen Datensätzen deutlich länger. Bezüglich der Rechenzeit sind also Modell- sowie Backbone-Wahl relevant. Welche Variante bevorzugt wird, hängt letztendlich vom Anwendungskontext sowie der Gewichtung zwischen Genauigkeit und Laufzeit ab.

5 Usability-Studie

5.1 Vorgehensweise zur Durchführung der Studie

Um die Qualität der Segmentierung von Grounded SAM 2 nicht nur anhand festgelegter Datensätze und Prompts, sondern auch unter realen Nutzerbedingungen zu bewerten, wurde eine Usability-Studie durchgeführt. Dabei stand die Beantwortung der dritten Forschungsfrage im Fokus:

„Wie wirkt sich die Nutzung von Grounded SAM 2 im Vergleich zur manuellen Segmentierung auf die Verarbeitungszeit und Effizienz von Segmentierungsaufgaben aus?“

Es sollte ermittelt werden, wie viel Zeit durch die verschiedenen Modell-Backbone-Kombinationen im Vergleich zur manuellen Segmentierung eingespart werden kann und ob Nutzende mit den Ergebnissen der automatischen Segmentierung zufrieden sind. Zur Durchführung wurde eine Webanwendung entwickelt, welche es erlaubt, Prompteingaben zu Bildern mit den unterschiedlichen Modell-Backbone-Kombinationen zu machen oder manuell Bounding Boxes (BBoxes) zu zeichnen und Objekte zu segmentieren.

5.1.1 Webanwendung

Die Webanwendung, die zur Durchführung der Studie entwickelt wurde, basiert auf von Lovable generiertem Quellcode (Lovable, Inc., 2025). Lovable ist eine Plattform, die mittels KI dabei helfen soll, Webseiten zu erstellen und wurde hier größtenteils für das Frontend genutzt.
Die Webanwendung wurde so konzipiert, dass sowohl manuelle als auch modellgestützte Segmentierung möglich ist. Bei der modellgestützten Segmentierung kann zwischen den verschiedenen Modellen und Backbones gewählt werden. Die resultierenden Ergebnisse können in Form von JSON-Dateien für die Bounding Boxes und für die Segmentierungsmaske heruntergeladen werden.
Außerdem wurde ein Pfad implementiert, über den Teilnehmende durch die Studie geleitet werden. Zu Beginn wurden einige einleitende Fragen, unter anderem zur Vorerfahrung mit Bildsegmentierungsmodellen, gestellt. Anschließend wurden die Teilnehmenden in Gruppen aufgeteilt, denen jeweils eine bestimmte Modell-Backbone-Kombination (z.B. Grounding DINO mit Hiera-small) zugeordnet war. Innerhalb dieser Gruppe führten die Teilnehmenden mehrere Segmentierungsaufgaben mit

der ihnen zugewiesenen Systemvariante durch. Jede Aufgabe bestand aus einem vorgegebenen Bild sowie einem Objekt, zu dem ein Textprompt formuliert werden sollte. Außerdem gab es eine Gruppe für die manuelle Segmentierung.
Nach Abschluss aller Bilder war ein weiterer Fragebogen zur Erfassung subjektiver Eindrücke zur wahrgenommenen Qualität der Ergebnisse auszufüllen. Zusätzlich wurde die Zeit der Segmentierungen erfasst.

Zur Veranschaulichung der Benutzeroberfläche wurden im Anhang A.1 exemplarisch Screenshots der Webanwendung ergänzt. Diese zeigen unter anderem die Ansicht zur manuellen Segmentierung, zur modellgestützten sowie zu den Ergebnisseiten.

5.1.2 Aufbau der Studie

Die Usability-Studie wurde mit insgesamt 39 Teilnehmenden durchgeführt, die unterschiedliche Erfahrungsstände im Bereich Bildverarbeitung und anderen KI-Anwendungen mitbrachten. Um ein einheitliches Verständnis sicherzustellen, wurde vor Beginn der Aufgaben allen Teilnehmenden ein Einführungsvideo gezeigt, das die Bedienung der Webanwendung sowie wichtige Begriffe, wie Bounding Box und Segmentierung, erklärte.

Die gesamte Studie, also sowohl Nutzeroberfläche als auch Fragen, wurde in englischer Sprache durchgeführt. Außerdem wurden die Teilnehmenden dazu aufgefordert, ihre Prompteingaben in Englisch zu formulieren.
Bevor mit den Segmentierungsaufgaben begonnen wurde, wurden einleitend allgemeine demografische Fragen gestellt.

Fragen vor der Segmentierung:

- How old are you?
- What is your highest level of education?
  (z.B. Bachelor´s degree, High school diploma, etc.)
- Have you ever used an AI tool for segmentation or object detection? If so, which one?
- In which field do you work or study?
  (STEM, Healthcare, Education, etc.)

Die Fragen dienten dazu, Hintergründe der Teilnehmenden einordnen zu können und so mögliche Einflüsse in den Kontext bringen zu können.
Nach den Einführungsfragen wurden die Teilnehmenden in eine von 5 Gruppen eingeteilt, dabei wurde versucht, darauf zu achten, dass der Anteil an Teilnehmenden, die im MINT-Bereich (Science, Technology, Engineering, and Mathematics (Naturwissenschaften, Technologie, Ingenieurwesen und Mathematik)) arbeiten oder studieren, in jeder der Gruppen ungefähr gleich hoch und der Altersdurchschnitt in jeder Gruppe ähnlich ist. Jede Gruppe führte die Aufgaben mit einer spezifischen Modell-Backbone-Kombination durch.

- Gruppe 1: Florence 2 mit Hiera-large
- Gruppe 2: Florence 2 mit Hiera-small
- Gruppe 3: Grounding DINO mit Hiera-large
- Gruppe 4: Grounding DINO mit Hiera-small
- Gruppe 5: Händisch ohne Grounded SAM 2 segmentieren

In allen Gruppen wurden in der Segmentierungsphase dieselben 20 Bilder (ursprünglich aus dem COCO-Datensatz Lin et al. (2015)) verwendet, auf denen jeweils ein zu segmentierendes Objekt war. Unter den Objekten gab es 6 kleine (small), 6 mittlere (medium) und 8 große (large) Objekte (Größenunterteilung nach Lin et al. (2015). Bei Gruppe 1-4 war die Aufgabe zu jedem Bild:

„*A red circle highlights an object in the image. Please segment it by describing it in the text field using as few words as possible. Please make your input in English. (Ein roter Kreis markiert ein Objekt im Bild. Bitte segmentieren Sie es, indem Sie es im Textfeld mit so wenigen Worten wie möglich beschreiben. Bitte geben Sie Ihren Text auf Englisch ein.)*“

Ein Beispielbild mit einem roten Kreis ist in Abbildung 5.1 zu sehen.

Abbildung 5.1: Beispielbild aus der Studie Gruppe 1-4. Hier sollte der gelbe Frisbee segmentiert werden.

Die Teilnehmenden sollten also zu jedem der zwanzig Bilder einen eigenen Textprompt auf Englisch formulieren. Die durch die Modelle erzeugten Segmentierergebnisse wurden danach angezeigt. Gruppe 5 (die Gruppe, die händisch segmentieren musste) hat die Bilder ohne rote Umrandungen bekommen. Hier lautete die Aufgabe beispielsweise:

„*Segment the frisbee by first drawing a tight Bounding Box around it. Then place points along its edge. Finally, refine the segmentation using the brush and eraser tools.*“

Es sollte also mit den Tools auf der Webanwendung das jeweils beschriebene Objekt segmentiert werden. Alle Gruppen haben identische Bilder bekommen, welche in A.1.1 dokumentiert sind. Während die Teilnehmenden Bilder segmentierten, wurde außerdem in allen Gruppen die Zeit gemessen.

Nach Abschluss aller Bilder folgte ein Frageblock, der die subjektive Erfahrung mit der Segmentierung erfassen sollte. Die meisten Fragen waren als Multiple-Choice-Fragen mit vorgegebenen Antwortskalen gestaltet, überwiegend auf Basis einer fünfstufigen Likert-Skala (Likert, 1932).

Fragebogen zur Bewertung der Segmentierungsergebnisse:

- How satisfied were you with the results of the segmentation tool?
  („Very satisfied“, „Satisfied“, „Neutral“, „Dissatisfied“, „Very dissatisfied“)
- Would you recommend this tool to others?
  („Yes“, „No“, „Maybe“)
- How helpful did you find the segmentation tool?
  („Very helpful“, „Helpful“, „Neutral“, „Not very helpful“, „Not helpful at all“)
- How well were the rectangular boxes (Bounding Boxes) drawn around the objects?
  („Very well“, „Well“, „Average“, „Poorly“, „Very poorly“)
- Did you notice anything specific about the Bounding-Boxes?
- How well were the objects inside the boxes segmented?
  („Very well“, „Well“, „Average“, „Poorly“, „Very poorly“)
- Did you notice anything specific about the segmentation inside the bounding boxes?
- Do you have any additional comments or suggestions for improvement?

Evaluationsmetriken

Die Bewertung der verschiedenen Modell-Backbone-Kombinationen sowie der händischen Segmentierung erfolgte anhand der Metriken Average Precision (AP) und Average Recall (AR) wie sie auch in den anderen Datensätzen verwendet wurden. Die Berechnung der Metriken wurde bereits in Abschnitt 2.9.3 und 2.9.2 erläutert. Da im Rahmen der Studie nur Bilder untersucht wurden, auf denen lediglich ein Objekt zu sehen ist, fallen die Metriken $\mathbf{AR}_{\mathbf{max=10}}$ und $\mathbf{AR}_{\mathbf{max=100}}$ weg. Die Rückmeldungen der einzelnen Gruppen wurden, wenn möglich, auf Grundlage der fünfstufigen Likert-Skala ausgewertet (Likert, 1932).

5.2 Ergebnisse

Im Folgenden werden die Ergebnisse der Studie präsentiert. Zunächst erfolgt eine Auswertung der erhobenen demografischen Daten. Anschließend werden die Segmentierungsergebnisse der Gruppen anhand AP und AR dargestellt und es wird auf einige Beispielbilder eingegangen. Abschließend werden die Resultate des Fragebogens zur subjektiven Einschätzung der Teilnehmenden ausgewertet.

5.2.1 Demografische Daten

Zur Einordnung der erhobenen Ergebnisse wurden zu Beginn der Studie demografische Informationen der jeweiligen Teilnehmenden erfasst. Die erfassten demografischen Daten sind in den folgenden Diagrammen und Tabellen (5.2, 5.3 und 5.1, 5.1) dargestellt.

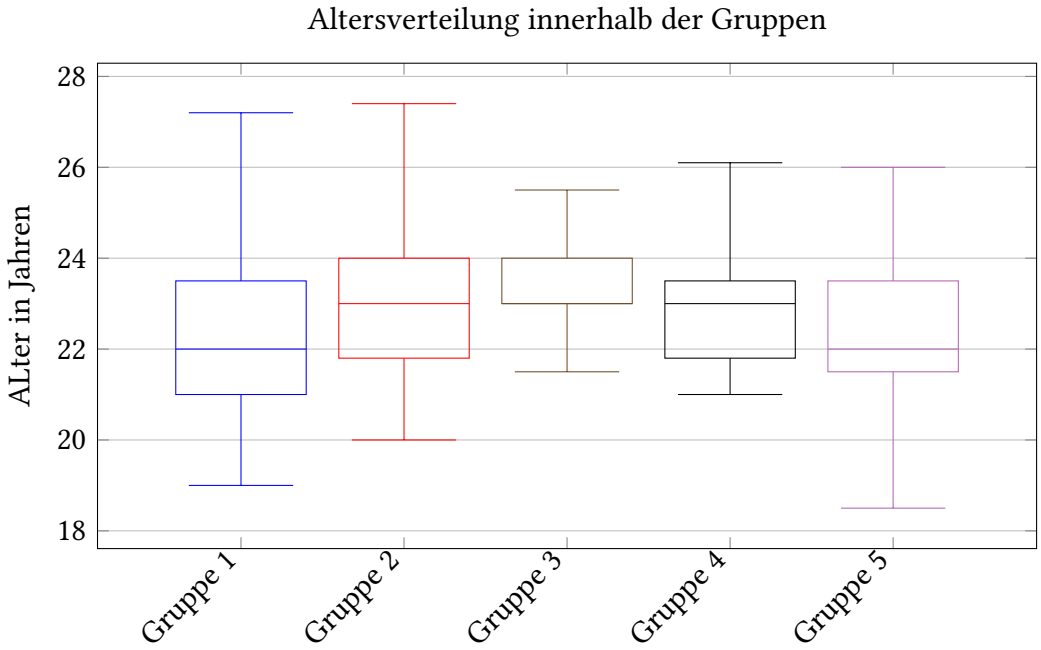

Abbildung 5.2: Boxplot zur Darstellung der Altersverteilung in den fünf Gruppen. Der Boxplot zeigt Median, unteres und oberes Quartil sowie die Spannweite an.

Die Altersverteilung der Teilnehmenden ist in Abbildung 5.2 in Form eines Boxplots dargestellt. Die Boxen zeigen jeweils das untere und obere Quartil sowie den Median innerhalb der fünf Gruppen. Die sogenannten „Whisker“ geben die Spannweite der Daten ohne Ausreißer an. Der Altersdurchschnitt sowie die Altersspannweite in den verschiedenen Gruppen wurden versucht, ähnlich zu halten. Wie in der Abbildung deutlich wird, befand sich der Großteil der Teilnehmenden zwischen 21 und 24 Jahren und es gibt keine signifikanten Ausreißer nach oben oder unten.

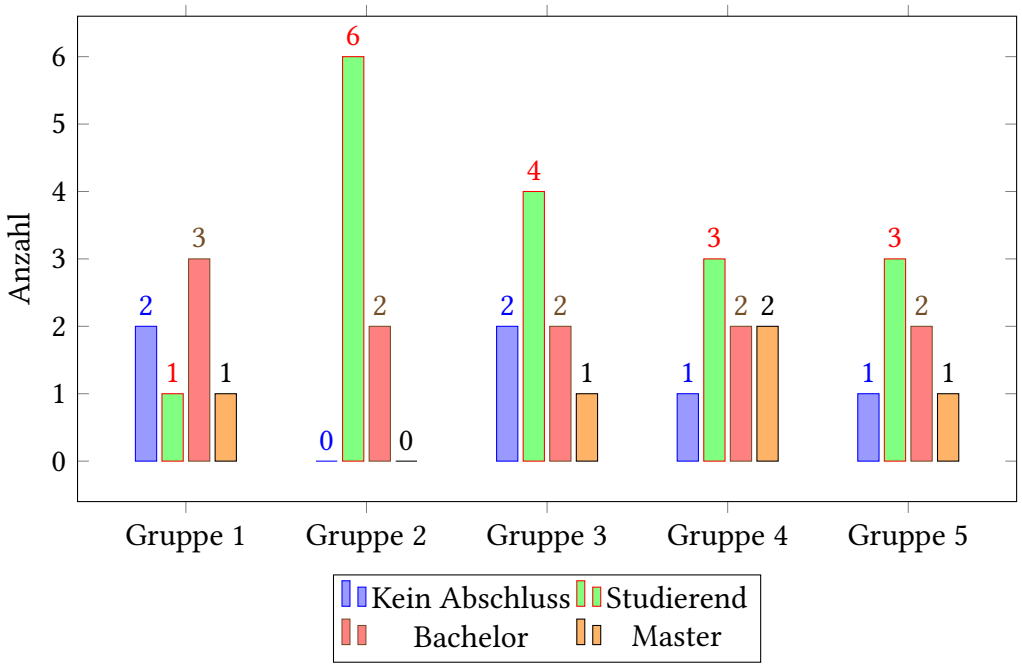

Abbildung 5.3: Gruppenweises Balkendiagramm zur Verteilung des Bildungsstands der Teilnehmenden. Dargestellt sind die vier abgefragten Bildungsniveaus: kein Abschluss (auch Teilnehmende mit Abitur o. ä.), studierend, Bachelor und Master. Die Zahlen über den Balken geben die Anzahl an Teilnehmenden mit dem jeweiligen Abschluss pro Gruppe an.

Die Teilnehmenden hatten unterschiedliche Bildungsstände, wobei studierend, also derzeit im Bachelorstudium, am häufigsten angegeben wurde. In Gruppe zwei sind besonders viele Studierende und in Gruppe eins weniger als in den anderen. Ansonsten haben einige wenige Teilnehmer:innen einen Master, ein paar mehr einen Bachelor und einige keinen Universitätsabschluss oder eine Universitätszulassung.

Gruppe	*n*	Anteil MINT (%)	Segmentiererfahrung (%)
1	7	57	43
2	8	63	25
3	9	66	22
4	8	63	25
5	7	71	29

Tabelle 5.1: Gruppenwesie Teinlnehmeranzahl(n), Anteil der Teilnehmenden die entweder im MINT-Bereich arbeiten oder im MINT-Bereich studieren (Anteil MINT) und Anteil der Teilnehmden die vorher schon einmal mit Segmentierwerkzeugen gearbeitet haben (Segmentiererfahrung).

Neben dem Bildungsstand und dem Alter wurden auch das Studien- bzw. Berufsfeld sowie die Vorerfahrung mit Segmentierwerkzeugen erhoben. In Tabelle 5.1 lässt sich erkennen, dass über alle Gruppen hinweg ein großer Anteil der Teilnehmenden einen MINT-Hintergrund (STEM) hat. Es haben in jeder Gruppe zwischen 57% und 71% einen MINT-Hintergrund, wobei Gruppe 5 mit 71% einen etwas größeren Anteil im Gegensatz zu den anderen Gruppen hat. Damit sind die Teilnehmenden der Studie eher technisch orientiert. Hier wird nur angegeben, ob ein:e Teilnehmende:r im MINT-Bereich tätig ist oder nicht, da genauere Erfahrungen außerhalb dieses Bereichs für die Studie nicht von großer Bedeutung zu sein scheinen. Außerdem haben einige Teilnehmende Segmentiererfahrung,

haben also zuvor schon einmal mit einer Anwendung zum Segmentieren gearbeitet. In Gruppe 1 ist der Anteil an Personen mit Segmentiererfahrung etwas höher als in den anderen Gruppen.

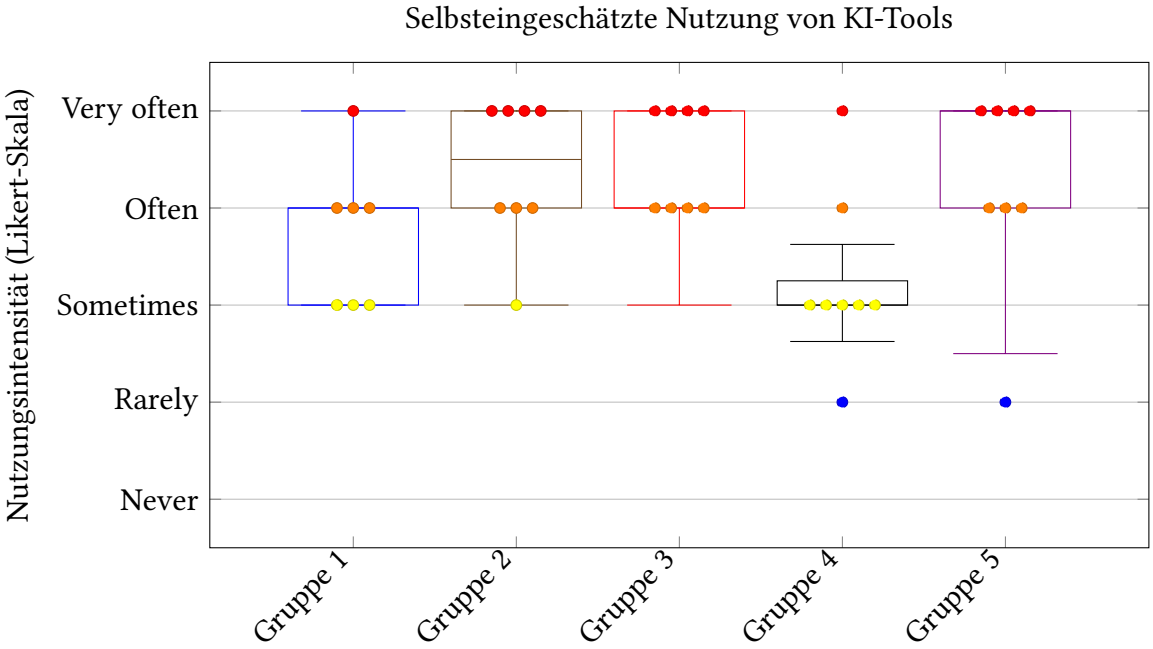

Abbildung 5.4: Boxplot zur Einschätzung der Nutzungsintensität von KI-Tools in den Gruppen (Likert-Skala von „Never“ bis „Very often“). Die Skala wurde zur Visualisierung als intervallskaliert interpretiert.

Ergänzend dazu wurde die generelle Nutzung von KI-basierten Anwendungen erfasst. Die Ergebnisse sind in 5.4 dargestellt. Der Boxplot macht sichtbar, dass besonders die Gruppen 2, 3 und 5 relativ hohe Medianwerte bei der selbsteingeschätzten KI-Nutzung aufweisen. Gruppe 4 zeigt etwas weniger KI-Nutzungshäufigkeit, wobei der Großteil der Teilnehmenden zumindest manchmal (sometimes) mit KI arbeitet. In Gruppe 1 ist die Nutzung noch etwas höher als in Gruppe 4. Insgesamt scheinen aber die meisten der Teilnehmenden manchmal oder sogar oft KI-basierte Anwendungen zu nutzen.

5.2.2 Bounding Boxen (BBox)

In diesem Abschnitt wird die Auswertung der Bounding Boxes bezüglich der Usability-Studie dargestellt. Es werden also AP (Tabelle 5.2) und AR (Tabelle 5.1) für die zwanzig Bilder aus der Studie für die unterschiedlichen Prompts der Teilnehmenden der jeweiligen Gruppe erhoben.

Gruppe	Modell	Backbone	$AP_{50\text{-}95}$	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
1	Florence 2	Hiera-large	61,4	78,9	69,3	15,0	73,3	96,7
2		Hiera-Small	64,9	81,6	69,3	31,9	70,4	94,7
3	Grounding Dino	Hiera-large	60,3	74,3	60,1	38,3	70,3	73,1
4		Hiera-small	61,5	75,4	67,6	41,4	71,0	70,8
5	/	/	44,9	75,4	42,3	13,3	47,9	74,3

Tabelle 5.2: Die Tabelle zeigt die Average Precision (AP) Ergebnisse der automatisch erzeugten Bounding Boxes der Usability-Studie. Die Gruppen 1-4 arbeitete mit einer spezifischen Modell-Backbone-Kombination (Florence 2 bzw. Grounding DINO, kombiniert mit Hiera-large oder Hiera-small). Gruppe 5 hat die Bounding Boxen manuell erstellt.

Die Ergebnisse in Tabelle 5.2 zeigen, dass Gruppe 2, die mit Florence 2 und Hiera-small gearbeitet hat, insgesamt die höchsten AP-Werte erreicht und besonders bei $AP_{50\text{-}95}$ besser abschneidet als die anderen Gruppen. Gruppe 1 erreicht mit Florence 2 und Hiera-large ebenfalls (bis auf AP_s) konsistent hohe Werte. Sehr auffällig ist der vergleichsweise niedrige Wert von AP_s (15%) bei Gruppe 1. Bei Betrachtung der einzelnen Ergebnisse der Teilnehmenden von Gruppe 1 zeigte sich, dass dies nicht an einzelnen Teilnehmenden lag. Keiner der Teilnehmer hat in Gruppe 1 für AP_s einen Wert von über 20% erreicht.
Die Gruppen 3 und 4, welche mit Grounding DINO arbeiten, schneiden im direkten Vergleich mit Gruppe 1 und 2 schlechter ab. Wobei Gruppe 4 (Hiera-small) bei fast allen Metriken bessere Ergebnisse aufweist als Gruppe 3 (Hiera-large). Gruppe 5, die Gruppe, die die Bounding Boxes manuell erstellt hat, liefert insgesamt solide Ergebnisse, bleibt bei höherer Präzision hinter den anderen Gruppen, erreicht aber bei niedrigerer Präzision, bei einer IoU von 50% einen ziemlich guten Wert (75,4% für AP_{50}).

Gruppe	Modell	Backbone	$AR_{max=1}$	AR_s	AR_m	AR_l
1	Florence 2	Hiera-large	61,4	15,0	73,3	96,7
2		Hiera-Small	69,4	31,9	70,4	94,7
3	Grounding Dino	Hiera-large	59,4	38,1	69,3	72,2
4		Hiera-small	61,5	41,4	71,1	70,9
5	/	/	45,1	13,3	47,9	75,5

Tabelle 5.3: Die Tabelle zeigt die Average Recall (AR) Ergebnisse der automatisch erzeugten Bounding Boxes der Usability-Studie. Die Gruppen 1-4 arbeitete mit einer spezifischen Modell-Backbone-Kombination (Florence 2 bzw. Grounding DINO, kombiniert mit Hiera-large oder Hiera-small). Gruppe 5 hat die Bounding Boxes manuell erstellt.

Bei den Average-Recall-Werten in Tabelle 5.3 zeigt sich ein sehr ähnliches Muster wie bei der AP. Gruppe 2 erreicht mit Florence 2 und Hiera-small den besten $\mathbf{AR_{max=1}}$ (69,4%). Gruppe 1 schneidet wieder bei mittleren und großen Objekten besser ab als Gruppe 2. Die Gruppen mit Grounding DINO (3 und 4) erreichen insgesamt schwächere Werte, und Gruppe 5 zeigt wie zuvor eine schlechtere Leistung als die anderen.

Insgesamt ist klar zu erkennen, dass das manuelle Erstellen von Bounding-Boxen solide Ergebnisse liefert und Grounded SAM 2 mit allen getesteten Modell-Backbone-Kombinationen aber eine höhere Segmentierungsgenauigkeit erreicht. Besonders deutlich wird dies bei den hohen Werten für AP_l und AR_l, also bei großen Objekten. Zudem ist auffällig, dass Gruppe 1, die mit der Modell-Backbone-Kombination aus Florence 2 und Hiera-large arbeitet, im Vergleich zu den Gruppen 2, 3 und 4 einen deutlich niedrigeren Wert für AP_s und AR_s aufweist. Das die Average Precision- und Average Recall-Werte sehr nah beieinanderliegen, lässt sich dadurch erklären, dass es auf jedem Bild nur ein Zielobjekt gab und dass es keine Objektklassenunterscheidung gibt (jede Referenz gehört zu einer einzelnen Klasse). Grounded SAM 2 erkennt meistens korrekt, dass pro Prompt nur ein Objekt segmentiert werden soll. In diesem Fall gibt es auf der Precision-Recall-Kurve nur einen Punkt (entweder Precision=Recall=1 oder beides 0). Die Fläche unter der Kurve entspricht dann einfach dem Precision-Wert. Da Precision hier häufig dem Recall-Wert entspricht, sind sich dadurch AP und AR sehr ähnlich. Aus den ähnlichen Werten lässt sich also ableiten, dass Grounded SAM 2 tatsächlich meistens erkennt, dass nur ein Objekt segmentiert werden soll.

Um die Ursachen für den auffällig niedrigen AP_s und AR_s in Gruppe 1 besser einordnen zu können, wurden die AP (AR ist wieder sehr nah an den AP-Werten und wird nicht explizit genannt) Ergebnisse für kleine Objekte von Gruppe 1 detaillierter aufgeschlüsselt. Die kleinen Objekte haben in der Gruppe bei einer IoU von 50% einen Average Precision-Wert von 35,7% und bei einer IoU von 75% nur 11,9%. Das bedeutet, die Modell-Backbone-Kombination scheint kleinere Objekte zwar manchmal korrekt zu erkennen, tut dies aber in vielen Fällen nicht präzise genug. In Abschnitt 5.2.4 wird dies anhand der Ergebnisbilder in Gruppe 1 genauer untersucht.

5.2.3 Segmentierung

In diesem Abschnitt werden die Ergebnisse der Segmentierungsleistung bezüglich der Usability-Studie präsentiert. In Tabelle 5.4 werden die Average Precision und in Tabelle 5.5 die Average Recall Ergebnisse dargestellt.

Gruppe	Modell	Backbone	$AP_{50\text{-}95}$	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
1	Florence 2	Hiera-large	66,7	81,6	76,3	31,7	80,2	85,0
2		Hiera-Small	66,0	83,3	76,3	39,7	73,5	85,5
3	Grounding Dino	Hiera-large	58,1	74,3	66,8	32,7	66,7	75,0
4		Hiera-small	59,7	75,4	71,2	36,4	66,6	75,6
5	/	/	56,6	86,9	64,0	23,3	67,8	76,7

Tabelle 5.4: Die Tabelle zeigt die Average Precision (AP) Ergebnisse der Segmentierung der Usability-Studie. Die Gruppen 1-4 arbeitete mit einer spezifischen Modell-Backbone-Kombination (Florence 2 bzw. Grounding DINO, kombiniert mit Hiera-large oder Hiera-small). Gruppe 5 hat die Segmentierung manuell erstellt.

Die Ergebnisse in Tabelle 5.4 zeigen, dass Gruppe 1 sowie Gruppe 2, die beide mit Florence 2 gearbeitet haben, wie auch bei den Bounding Boxes, deutlich höhere AP-Werte erzielen als die Gruppen mit Grounding DINO. Gruppe 1 (Florence 2 mit Hiera-large) erreicht leicht höhere Werte bei $AP_{50–95}$ und AP_m als Gruppe 2 und Gruppe 2 erreicht bei AP_{50}, AP_s und AP_l minimal höhere Werte. Welche der beiden Gruppen insgesamt besser abgeschnitten hat, lässt sich nicht eindeutig bestimmen. Eindeutig ist allerdings, dass beide Gruppen mit ihren Konfigurationen bessere Ergebnisse als Gruppe 3 und 4 (Grounding DINO) aufweisen. Zwischen den beiden Gruppen mit Grounding DINO fallen die Unterschiede auch eher gering aus. Gruppe 5 (manuelle Segmentierung) liefert hier auch wieder solide Ergebnisse. Die anderen Gruppen sind Gruppe 5 allerdings insbesondere bei höherer IoU ($\mathbf{AP_{75}}$) und kleinen Objekten (AP_s) überlegen.

Gruppe	Modell	Backbone	$AR_{max=1}$	AR_s	AR_m	AR_l
1	Florence 2	Hiera-large	66,6	31,7	80,2	85,0
2		Hiera-Small	66,0	39,7	73,5	85,5
3	Grounding Dino	Hiera-large	58,3	32,7	66,7	75,0
4		Hiera-small	59,7	36,3	66,6	75,5
5	/	/	56,7	23,3	67,9	76,6

Tabelle 5.5: Die Tabelle zeigt die Average Recall (AR) Ergebnisse der Segemntierung der Usability-Studie. Die Gruppen 1-4 arbeitete mit einer spezifischen Modell-Backbone-Kombination (Florence 2 bzw. Grounding DINO, kombiniert mit Hiera-large oder Hiera-small). Gruppe 5 hat die Segmentierung manuell erstellt.

Die Average Recall Ergebnisse in Tabelle 5.5 unterscheiden sich, wie auch bei den Bounding Boxes, nur sehr gering von den Average Precision Werten (Tabelle 5.4). Auch hier sind sich die beiden Gruppen, die Florence 2 (Gruppen 1 und 2) genutzt haben, sehr leistungsnah und schneiden insgesamt besser ab als Gruppe 3 und 4 (Grounding DINO). Gruppe 5 bleibt insgesamt zurück.

Insgesamt zeigen die Ergebnisse der Segmentierung, dass die Modell-Backbone-Kombinationen mit Florence 2 den Gruppen mit Grounding DINO in nahezu allen Metriken überlegen sind. Die Gruppen 1 und 2 liefern konsistent bessere Leistungen, wobei keine der beiden eindeutig besser als die andere ist. Die Unterschiede zwischen beiden Konfigurationen fallen gering aus und deuten darauf hin, dass beide Florence 2-Varianten sehr gut für die Segmentierung mit offenen Prompteingaben geeignet sind. Die manuelle Segmentierung der Gruppe 5 liefert, wie auch bei den Bounding Boxes, solide, aber den anderen Gruppen unterlegene Werte. Bei der Segmentierung sind sich die Average Precision und Average Recall-Werte erneut sehr ähnlich. Wie zuvor bei den Bounding Boxes liegt dies vermutlich daran, dass nur ein Objekt pro Prompt segmentiert werden sollte und es keine richtigen Objektklassen gab.

5.2.4 Beispielbilder

Zur Veranschaulichung werden in diesem Kapitel ausgewählte Segmentierungsbeispiele aus den einzelnen Gruppen dargestellt. Dabei ist zu beachten, dass hauptsächlich Ergebnisse ausgewählt wurden, die besonders hervorstechen (positiv sowie negativ). Es sind also lediglich einzelne Fälle, die keine allgemeingültige Aussage über die durchschnittliche Gruppenleistung treffen können, sondern nur Stärken und Schwächen der Gruppen aufzeigen sollen. Alle in diesem Kapitel gezeigten Beispielbilder stammen Ursprünglich aus dem COCO 2017-Datensatz (Lin et al. 2015).

Gruppe 1

Für Gruppe 1 wurden aufgrund der auffällig kleinen AP_s- und AR_s-Werte für die Bounding Boxes drei Beispielbilder mit kleinen Objekten ausgewählt.

Abbildung 5.5: Beispielbild aus Gruppe 1 mit der Modell-Backbone-Kombination Florence 2 + Hiera-large. Zu sehen ist ein roter Pickup-Truck auf verschneiter Straße. Gesuchtes Objekt war ein „fire hydrant“. Prompt: “Water fountain”. Das falsche Objekt wurde lokalisiert.

Abbildung 5.6: Beispielbild aus Gruppe 1 mit Florence 2 + Hiera-large. Gezeigt ist ein blauer Zug vor einem Tunnel. Gesuchtes Objekt war die rot leuchtende Ampel links neben dem Zug. Prompt: “Red light”. Stattdessen wurde ein Teil des Zuges erkannt.

In den Ergebnissen von Gruppe 1 ist bei kleinen Objekten aufgefallen, dass etwas häufiger Prompts verwendet wurden, welche für die Modelle gegebenenfalls schwerer semantisch einzuordnen sind als in den anderen Gruppen. Ein Beispiel dafür ist Abbildung 5.5. Hier sollte eigentlich der Hydrant im Hintergrund segmentiert werden, als Prompt wurde aber „Water Fountain“ eingegeben. Ein anderes Beispiel war ein Bild, auf dem eine Pflanze segmentiert werden sollte, wo als Prompt „Vase“ eingegeben wurde. Außerdem ist aufgefallen, dass keiner der Teilnehmer:innen aus Gruppe 1 das richtige Objekt für das Bild in Abbildung 5.6 segmentieren konnte. Hier wurde entweder das falsche oder zusätzlich ein anderes Objekt ausgewählt. Hier sollte eigentlich die Ampel links neben dem Zug segmentiert werden, es wurde aber entweder nur das rote Licht der Ampel oder das rote Licht am Zug segmentiert.

Abbildung 5.7: Beispielbild aus Gruppe 1 mit Florence 2 + Hiera-large. Eine Fußballszene mit mehreren Spielern. Gesuchtes Objekt war der Fußball. Prompt: “Blue football with white points”. Das richtige Objekt wurde erkannt.

In Abbildung 5.7 wurde das richtige Objekt (der Fußball) zwar erkannt, allerdings zeigt sich bei genauerer Betrachtung, dass die Bounding Box etwas unpräzise ist und deutlich enger um den Fußball hätte sein können. Dabei ist zu beachten, dass in dem Prompt das Wort „white“ falsch geschrieben wurde. Ob das einen Einfluss auf die Genauigkeit der Bounding Box hatte, ist schwer zu beurteilen. Solche ungenauen Bounding Boxes wurden bei anderen Bildern dieser Gruppe ebenfalls häufiger bei kleineren Objekten beobachtet. Die gewählten Beispiele spiegeln insgesamt wider, wie es zu so niedrigen AP- und AR-Werten bei kleinen Objekten in Gruppe 1 kommen konnte.

Abbildung 5.8: Beispielbild aus Gruppe 1 mit der Modell-Backbone-Kombination Florence 2 + Hiera-large. Dargestellt ist eine Küche mit verschiedenen Objekten. Prompt: "It's small and seems like a green plant". Das gesuchte Objekt wurde erkannt.

Abbildung 5.9: Beispielbild aus Gruppe 1 mit Florence 2 + Hiera-large. Gezeigt wird ein Skatepark. Prompt: "It's a skateboard. It flies in the air because someone is jumping". Die Variante segmentierte das richtige Objekt.

In den Abbildungen 5.8 und 5.9 wurde jeweils das korrekte Objekt erkannt. Interessant ist daran, dass die eingegebenen Prompts relativ lang sind und die Modell-Backbone-Kombination offensichtlich in der Lage dazu ist, auch lange komplexe Eingaben zuverlässig zu interpretieren. In beiden Fällen enthielten die Prompts mehrere beschreibende Elemente, was besonders bei unerfahrenen Nutzenden häufig vorkommt. Insbesondere im Hinblick auf praktische Fälle zeigt sich somit eine robuste und vor allem nutzerfreundliche Anwendung.

Abbildung 5.10: Beispielbild aus Gruppe 1 mit der Variante Florence 2 + Hiera-large. Prompt: „No idea". Objekt wurde korrekt erkannt.

Abbildung 5.11: Beispielbild aus Gruppe1 mit Florence 2 + Hiera-large. Prompt: „Bank". Das gesuchte Objekt wurde nicht erkannt.

In Abbildung 5.10 ist zu sehen, dass trotz eines sinnfreien Prompts („No idea") das richtige Objekt (der gelbe Frisbee) erkannt wurde. Der Grund dafür könnte sein, dass sich das Zielobjekt deutlich im Bildzentrum befindet, was auch zuvor bei anderen Beispielen zu beobachten war.
In Abbildung 5.11 wurde anstelle eines englischen ein deutscher Prompt verwendet. Der Prompt „Bank" führte nicht zur Erkennung des richtigen Objekts, während Teilnehmer, die „Bench" eingegeben hatten, korrekte Ergebnisse erhielten. Offensichtlich arbeiten die Modelle besser mit englischsprachigen als mit deutschen Eingaben.

Gruppe 2

Gruppe 2 arbeitete mit der Modell-Backbone-Kombination Florence 2 und Hiera-small. Die folgenden beiden Beispielbilder zeigen die Ergebnisse der Modelle auf Prompteingaben mit unterschiedlicher Groß- und Kleinschreibung.

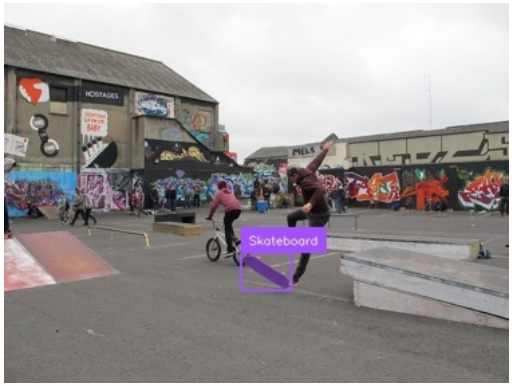

Abbildung 5.12: Beispielbild aus Gruppe 2 mit der Variante Florence 2 + Hiera-small. Prompt: „Skateboard“. Objekt wurde korrekt erkannt.

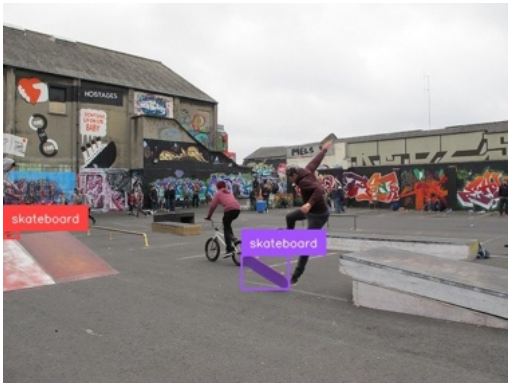

Abbildung 5.13: Beispielbild aus Gruppe 2 mit Florence 2 + Hiera-small. Prompt: „skateboard“. Das gesuchte Objekt wurde erkannt, zusätzlich aber noch ein Weiteres.

In Abbildung 5.12 wurde ein einzelnes Skateboard richtig erkannt, wohingegen in Abbildung 5.13, wo der Prompt kleingeschrieben wurde, ein weiteres Skateboard erkannt wurde. Theoretisch haben die Modelle mit der kleingeschriebenen Prompteingabe besser gearbeitet, da das zusätzlich erkannte Objekt ebenfalls ein Skateboard ist. Obwohl ein solcher Effekt nur vereinzelt beobachtet wurde und vergleichbare Fälle in der Studie kaum vorkamen, könnte die unterschiedliche Groß- und Kleinschreibung einen Einfluss auf die Interpretation haben. Außerdem fällt auf, dass bei keinem der Bilder die Räder des Skateboards mit segmentiert wurden.

Abbildung 5.14: Beispielbild aus Gruppe 2 mit der Variante Florence 2 + Hiera-small. Prompt: „Something Red inside“. Objekt wurde korrekt erkannt.

Abbildung 5.15: Beispielbild aus Gruppe 2 mit Florence 2 + Hiera-small. Prompt: „Boat“. Das gesuchte Objekt wurde nicht erkannt.

Auch Gruppe 2 scheint semantisch komplexere Prompts korrekt interpretieren zu können. In Abbildung 5.14 wurde der etwas kompliziertere Prompt „Something Red inside“ mit dem richtigen Objekt verknüpft. In Abbildung 5.15 wurde wieder wie bei Gruppe 1 ein sinnloser Prompt mit einem im Zentrum liegenden Objekt in Verbindung gebracht. Es handelt sich zwar nicht um das

gesuchte Objekt, zeigt aber trotzdem, dass auch diese Modell-Backbone-Kombination große Modelle im Zentrum der Bilder priorisiert. Die Segmentierung ist hier bei allen Objekten relativ gut.

Gruppe 3

In Gruppe 3 kam die Modell-Backbone-Kombination Grounding DINO mit Hiera-large zum Einsatz. Folgende Bildbeispiele zeigen sowohl gelungene als auch weniger gute Ergebnisse.

Abbildung 5.16: Beispielbild aus Gruppe 3 mit der Variante Grounding DINO 2 + Hiera-large. Prompt: „can of tomato paste with a white sticker on it“. Objekt wurde korrekt erkannt.

Abbildung 5.17: Beispielbild aus Gruppe 3 mit Grounding DINO + Hiera-large. Prompt: „bank“. Objekt wurde nicht erkannt.

In Abbildung 5.16 ist zu erkennen, dass auch Grounding DINO mit Hiera-large in der Lage dazu ist, längere, komplexere Prompts zu verstehen. Die Dose wurde hier richtig erkannt. Worin die Modell-Backbone-Kombination aber anscheinend nicht gut ist, ist der Umgang mit deutschen Begriffen. In Abbildung 5.17 wurde die Bank trotz des einfachen Prompts „bank“ nicht erkannt. Diese Beobachtung wurde bereits in Gruppe 2 gemacht.

Abbildung 5.18: Beispielbild aus Gruppe 3 mit der Variante Grounding DINO + Hiera-large. Prompt: „bagpack“. Objekt wurde nicht erkannt.

Abbildung 5.19: Beispielbild aus Gruppe 3 mit Grounding DINO + Hiera-large. Prompt: „fristby“. Objekt wurde nicht erkannt.

In den Abbildungen 5.18 und 5.19 fällt auf, dass Rechtschreibfehler wie hier „bagpack“ oder „fristby“ statt „backpack“ oder „frisbee“ leicht zu Fehlschlägen führen können. In beiden Fällen wurden große Objekte, die im Zentrum des Bildes standen und mit dem Prompt nichts übereinstimmen, erkannt. Es werden also auch hier tendenziell zentrale Objekte priorisiert. Auch in Gruppe 3 wirkt die Segmentierungsgenauigkeit relativ hoch.

Gruppe 4

Gruppe 4 hat mit Grounding DINO kombiniert mit Hiera-small gearbeitet. Die ausgewählten Beispielbilder sollen Schwächen und Stärken verdeutlichen.

Abbildung 5.20: Beispielbild aus Gruppe 4 mit der Variante Grounding DINO + Hiera-small. Prompt: „can on table“. Objekt wurde nicht erkannt.

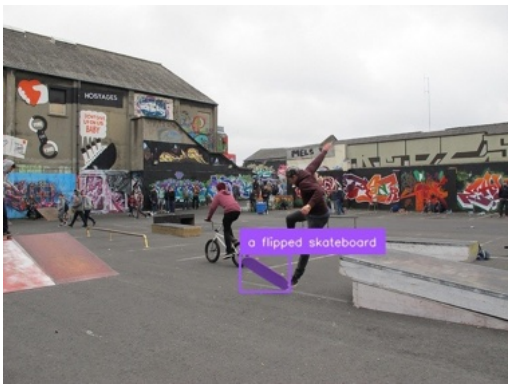

Abbildung 5.21: Beispielbild aus Gruppe 3 mit Grounding DINO + Hiera-small. Prompt: „a flipped skateboard“. Objekt wurde erkannt.

Abbildung 5.20 zeigt ein Beispiel, in dem eine etwas längere, eigentlich aber noch einfache Eingabe zu einem falschen Ergebnis geführt hat. Es wurde anstelle der Dose auf dem Tisch, der Tisch segmentiert. Grund dafür ist wahrscheinlich, dass der Tisch relativ zentral und groß auf dem Bild abgebildet ist, und dass das Wort „table“ selber auch in dem Prompt enthalten ist. Im Gegensatz dazu wurde das Objekt in Abbildung 5.21 mit dem Prompt „a flipped skateboard“ korrekt erkannt, was zeigt, dass auch diese Konfiguration teilweise längere Formulierungen verarbeiten kann. Im Vergleich zu den Florence 2 Kombinationen wirkt es so, als würde sie hierbei insgesamt schwächer abschneiden.

Abbildung 5.22: Beispielbild aus Gruppe 4 mit der Variante Grounding DINO + Hiera-small. Prompt: „toilet“. Objekt wurde erkannt.

Abbildung 5.23: Beispielbild aus Gruppe 4 mit Grounding DINO + Hiera-small. Prompt: „red light“. Objekt wurde nicht erkannt.

Gruppe 4 war die Einzige, bei der man in einigen wenigen Bildern eindeutige Mängel bei der Segmentierungsgenauigkeit in mehreren Beispielen sehen konnte. Ein Beispiel dafür ist in Abbildung 5.22. Die Toilette wurde hier zwar richtig lokalisiert, am Spülkasten fehlen jedoch einzelne Teile bei der Segmentierung.
In Abbildung 5.23 wird deutlich, dass auch Gruppe 4, wenn die Modell-Backbone-Kombination kein Objekt findet, das mit dem Prompt in Zusammenhang steht, einfach auf ein zentrales Objekt ausweicht.

Hier wurde bei der Prompt-Eingabe „red light“ der zentral im Bild stehende Zug ausgewählt. Auch bei ähnlichen Eingaben wurde in Gruppe 3 und 4 auf diesem Bild der Zug ausgewählt. Kombinationen mit Florence 2 waren in der Lage, bei der Eingabe „red light“ zumindest etwas Rotes auf dem Bild zu lokalisieren. Allerdings haben die Gruppen 3 und 4 im Gegensatz zu den Kombinationen mit Florence 2 das richtige Objekt bei der Eingabe „traffic light“ gewählt. Dies deutet darauf hin, dass die Konfigurationen mit Grounding DINO zwar größere Schwächen bei freien, vageren Texteingaben zeigen, dafür jedoch bei präziseren Prompts wie „traffic light“ relativ zuverlässig reagieren können.

Gruppe 5

Gruppe 5 arbeitete ohne Grounded SAM 2 und erstellte sowohl Bounding Boxes als auch Segmentierungen manuell. Die Beispielergebnisse stellen typische Ungenauigkeiten relativ gut dar.

Abbildung 5.24: Beispielbild aus Gruppe 5 mit manueller Segmentierung. Die Aufgabe lautete:„Segment the frisbee by first drawing a tight bounding box around it. Then place points along its edge. Finally, refine the segmentation using the brush and eraser tools.“

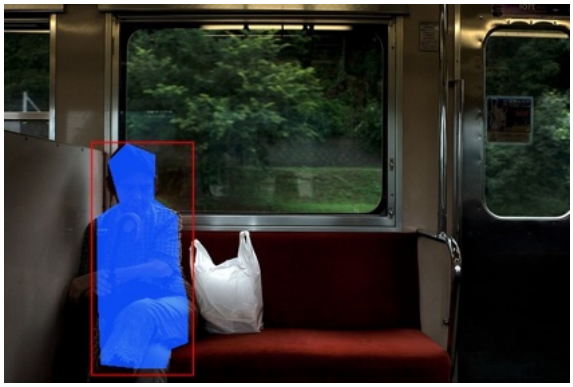

Abbildung 5.25: Beispielbild aus Gruppe 5 mit manueller Segmentierung. Die Aufgabe lautete: „Segment the woman by first drawing a tight bounding box around it. Then place points along its edge. Finally, refine the segmentation using the brush and eraser tools.“

In Abbildung 5.24 wurde das Objekt, sowie in fast allen Segmentierungsaufgaben der Gruppe, korrekt identifiziert, allerdings fällt auf, dass die Bounding Box zu groß und die Segmentierungsmaske etwas ungenau ist. Bei der Segmentierung wurden Teile der Hand überdeckt und auch die Kanten sind etwas unscharf.
In Abbildung 5.25 in der eine Frau segmentiert werden sollte, umschließt die Bounding-Box nicht das gesamte Objekt. Zudem ist die Segmentierung sehr ungenau.

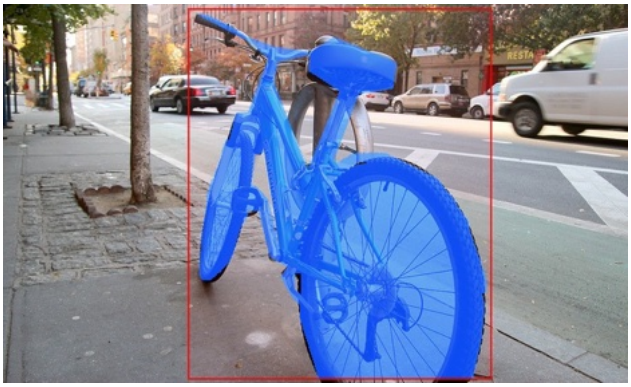

Abbildung 5.26: Beispielbild aus Gruppe 5 mit manueller Segmentierung. Die Aufgabe lautete: „ Segment the bicycle by first drawing a tight bounding box around it. Then place points along its edge. Finally, refine the segmentation using the brush and eraser tools.“

Auch in Abbildung 5.26 wurde das richtige Objekt (das Fahrrad) gewählt. Hier ist auch sehr gut zu erkennen, dass trotz wahrscheinlich etwas längerer Bearbeitungszeit, Details schwierig manuell zu segmentieren sind. Kleinere Bereiche wurden ausgelassen oder es wurden Teile vom Hintergrund mit segmentiert. Gerade bei der Segmentierung von Außenkanten stößt die manuelle Methode an ihre Grenzen.

5.2.5 Auswertung der Fragebogenergebnisse

Im Folgenden sind die Ergebnisse aus dem Fragebogen, den die Studienteilnehmer:innen nach der Annotation der Bilder ausgefüllt haben, dargestellt. Die ursprünglichen Fragen wurden in englischer Sprache (siehe 5.1.2) gestellt, wurden aber für die Darstellung in diesem Abschnitt übersetzt.
Die Multiple-Choice-Fragen werden grafisch in Skalen mit markierten Gruppendurchschnittswerten zusammengefasst.
Da die Freitextfragen nicht in Diagrammen darstellbar sind, werden relevante Antworten im Text erwähnt. Bei diesen Fragen wurden insgesamt allerdings nur vereinzelt aussagekräftige Antworten gesammelt.

Die Antworten der ersten Frage: „**Würden Sie das Tool anderen weiterempfehlen**“ sind in Tabelle 5.6 zusammengefasst.

Gruppe	Backbone	Modell	J/V/N
1	Florence 2	Hiera-large	5/2/0
2		Hiera-Small	8/0/0
3	Grounding Dino	Hiera-large	7/2/0
4		Hiera-small	4/3/1
5	/	/	2/4/2

Tabelle 5.6: Gruppenweise Antworten auf die Frage Würden Sie das Tool anderen weiterempfehlen?. Antwortenanzahl aufgeteilt nach J = Ja, V = Vielleicht und N = Nein.

In den Gruppen 1 und 2 (Die mit Florence 2 arbeiten) sprachen sich die meisten Teilnehmenden für eine Weiterempfehlung aus. Auch in den Gruppen, in denen Grounding DINO verwendet wurde, wurde das Tool meistens empfohlen, wobei in Gruppe 4 erstmals auch eine Person „Nein“ wählte. In Gruppe 5, die ohne Grounded SAM 2 arbeitete, fiel das Meinungsbild gemischter aus. Neben zwei positiven und 4 neutralen gab es auch zwei negative Rückmeldungen. Die modellgestützten Varianten wurden also öfter weiterempfohlen als das Tool zur manuellen Segmentierung.

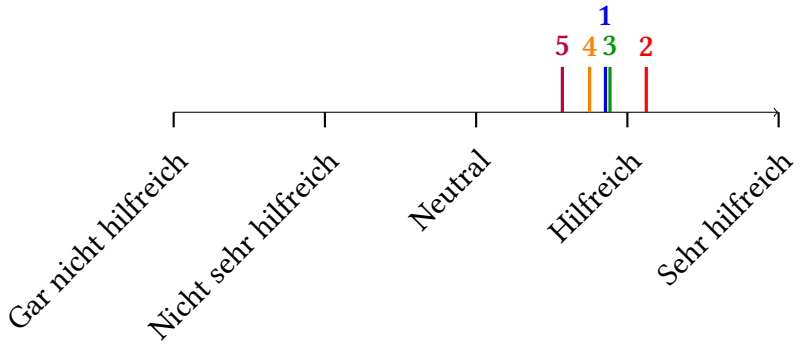

Abbildung 5.27: Durchschnittswerte der Gruppen zu der Frage: „Wie hilfreich fanden Sie das Segmentierungstool?“

In Abbildung 5.27 sind die Gruppenmittelwerte für die Frage: „**Wie hilfreich fanden Sie das Segmentierungstool?**“
dargestellt. Die Durchschnittswerte der Gruppen liegen alle relativ nah aneinander und auch innerhalb der Gruppen waren die Antworten eher homogen. Gruppe 5 wurde wie erwartet als etwas weniger hilfreich als die anderen Gruppen eingestuft. Der Unterschied ist jedoch minimal und könnte zufällig sein, da auch die Antworten in Gruppe 5 nicht signifikant abfallen.

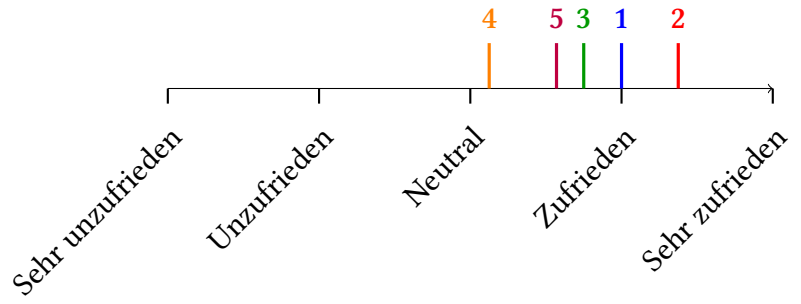

Abbildung 5.28: Durchschnittswerte der Gruppen zu der Frage: „Wie zufrieden waren Sie mit den Ergebnissen des Segmentierungstools?“

Die nächste Frage: „**Wie zufrieden waren sie mit den Ergebnissen des Segmentierungstools?**“ wird in Abbildung 5.28 dargestellt. Die Frage ist etwas redundant zur vorherigen und es lassen sich auch ähnliche Trends feststellen. Die Durchschnittswerte liegen auch hier in einem ähnlichen Bereich . Die Abstände der Gruppen sind jedoch etwas größer als bei der vorherigen Frage und hier schneidet Gruppe 4 am schlechtesten ab. Am besten sind die Umfrageergebnisse von Gruppe 2 und darauf folgt Gruppe 1.
Die Ergebnisse der nächsten Frage „**Wie gut wurden die rechteckigen Boxen (Bounding Boxes) um die Objekte gemalt?**“ werden in Abbildung 5.29 gezeigt.

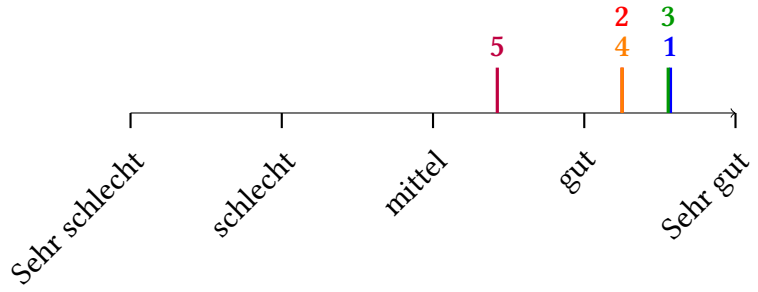

Abbildung 5.29: Durchschnittswerte der Gruppen zu der Frage: „Wie gut wurden die rechteckigen Boxen (Bounding Boxes) um die Objekte gemalt?“

Da nur ein Teil der Teilnehmenden Erfahrung mit Annotationstools (siehe Tabelle 5.1) hat, war es für die meisten wahrscheinlich schwierig zu beurteilen, wie gut oder schlecht die Bounding Boxes sind. Trotzdem sind die Ergebnisse relativ erwartungsgemäß ausgefallen. Gruppe 5 hat hier die schlechtesten Umfrageergebnisse und liegt zwischen mittel und gut, während alle anderen Gruppen zwischen gut und sehr gut liegen. Zu der Frage „**Haben Sie etwas Bestimmtes an den Bounding Boxes bemerkt?**“ waren die meisten Antworten auf bestimmte Bilder bezogen, die größtenteils schon in Abschnitt 5.2.4 betrachtet wurden. In Gruppe drei wurde von einem der Teilnehmenden geschrieben, dass die Bounding-Box manchmal Teile mit einbindet, die sie nicht soll, oder andersrum Teile des Objektes vergisst. In Gruppe 4 wurde von einer Person geschrieben, dass, wenn ein falsches Objekt gewählt wurde, meistens der Mensch selber schuld daran sei (mit fehlerhafter/ungenauer Prompteingabe). Die Aussagen lassen sich natürlich schwer mit anderen Gruppen vergleichen und spiegeln nur individuelle Wahrnehmungen von Einzelpersonen wider, sind jedoch dennoch allgemein interessant in Bezug auf Grounded SAM 2.

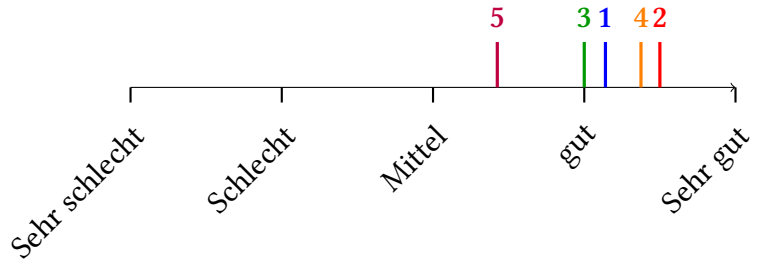

Abbildung 5.30: Durchschnittswerte der Gruppen zu der Frage: „Wie gut wurden die Objekte innerhalb der Bounding Boxes segmentiert?“

In Abbildung 5.30 sind die Ergebnisse für die Frage „**Wie gut wurden die Objekte innerhalb der Bounding Boxes segmentiert?**“ dargestellt. Auch hier ist das Ergebnis trotz der geringen Erfahrung mit ähnlichen Anwendungen nicht besonders überraschend ausgefallen. Gruppe 5 hat wieder die schlechteste Bewertung zwischen mittel und gut bekommen, während Gruppe 2 am nächsten von allen Gruppen an sehr gut liegt. Die Gruppen 3, 1 und 4 sind relativ nah an Gruppe 2.
Zu der Frage „**Haben Sie etwas Bestimmtes an der Segmentierung innerhalb der Bounding Boxes bemerkt?**“ wurde etwas mehr geschrieben als zu den Bounding Boxes. In Gruppe 1 wurde beschrieben, dass es Probleme damit gab, Objekte zu segmentieren, die Löcher hatten. Als Beispiel wurden die Speichen auf einem Bild mit einem Fahrrad genannt, hier wurde der Hintergrund hinter den Speichen ebenfalls segmentiert. Eine Person in Gruppe 2 und eine in Gruppe 4 beobachteten, dass die Segmentierungsqualität abnahm, wenn sich die zu segmentierenden Objekte farblich wenig von

ihrer Umgebung unterschieden. In Gruppe 3 und 4 wurde außerdem beobachtet, dass manchmal das falsche Objekt gewählt wurde. In der fünften Gruppe fand eine Person, dass es schwierig war, kleine Objekte zu segmentieren. Wie auch bei den Bounding Boxes lässt sich daraus nur schwer etwas zu bestimmten Modell-Backbone-Kombinationen ableiten. Die erhöhte Schwierigkeit bei kleinen Objekten in Gruppe 5 spiegelt sich jedoch ebenfalls in den AP- und AR-Werten aus Abschnitt 5 wider.

5.2.6 Benötigte Zeit

Neben der Segmentierungs- sowie Bounding Box-Qualität war auch die benötigte Zeit für den Annotationsteil der Studie ein wichtiger Faktor, um die dritte Forschungsfrage beantworten zu können. In den Gruppen 1-4 wurde nur die Zeit gemessen, die die Teilnehmenden für die Prompteingaben und das Modell für die Annotationen gebraucht hat, nicht die Zeit, in der sich die Probanden das Ergebnis angeschaut haben. Bei Gruppe 5 wurde die Zeit gemessen, die gebraucht wurde, um die Bounding Boxes sowie die Segmentierungsmasken zu erstellen.

Gruppe	Backbone	Modell	Zeit in Sekunden
1	Florence 2	Hiera-large	144
2		Hiera-Small	184
3	Grounding Dino	Hiera-large	272
4		Hiera-small	242
5	/	/	595

Tabelle 5.7: Benötigte Zeit der Segmentierung in den verschiedenen Gruppen in Sekunden

In Tabelle 5.7 sind die Durchschnittszeiten der jeweiligen Gruppen dargestellt. Die beiden Gruppen, die mit Florence 2 gearbeitet haben, erreichen die kürzesten Bearbeitungszeiten. Danach folgen Gruppe 3 und 4 mit Grounding DINO. Gruppe 5 schneidet mit Abstand am schlechtesten ab. Hier brauchten die Teilnehmenden im Schnitt mehr als doppelt so lange wie die in den anderen Gruppen. Zwischen den Hiera-Varianten unterscheiden sich die Werte weniger und müssen nicht unbedingt eine Überlegenheit der Konfiguration bezüglich der Zeit darstellen. Bei den Gruppen 1 und 2 mit Florence 2 ist Hiera-large und bei den Gruppen 3 und 4 ist Grounding DINO mit Hiera-small minnimal besser.

5.2.7 Fazit zur Usability-Studie

Die Usability-Studie zeigt, dass Grounded SAM 2 im Vergleich zur manuellen Segmentierung eine signifikante Zeitersparnis bietet. Grounded SAM 2 liefert schnellere und oft präzisere Ergebnisse. Insbesondere wurden Objektkanten mit Grounded SAM 2 sauberer segmentiert und die generierten Bounding Boxes lagen oft genauer um das Zielobjekt. Allerdings liefert die manuelle Methode,laut den AP- und AR-Werten, bei mittleren und großen Objekten eine vergleichbare Leistung zu den Grounding-DINO basierten Kombinationen. Auch wenn die Segmentierung mit der Hand aufwändig ist, gelingt es in der Regel, das richtige Objekt zu identifizieren. Die verschiedenen Modellvarianten zeigen ebenfalls Unterschiede. Florence 2 erreichte deutlich höhere Werte über alle Metriken hinweg und schien bei Betrachtung der Bilder, besonders bei längeren Prompteingaben, besonders gut abzuschneiden. Aufällig war bei allen Kombinationen, dass bei Unsicherheiten, unabhängig vom eigentlichen Prompt, häufig ein auffälliges Objekt im Bild segmentiert wurde. Insgesamt bietet Grounded SAM 2 mit geeignetem Modell und Backbone einen klaren Mehrwert für viele Anwendungsfälle. Die Segmentierung durch Grounded SAM 2 wurde von Teilnehmenden als effizienter und hilfreicher empfunden als die manuelle Methode.

6 Diskussion

6.1 Ergebnisdiskussion im Kontext der Forschungsfragen

In diesem Abschnitt werden die Ergebnisse der durchgeführten Experimente sowie der Usability-Studie im Kontext der zuvor gestellten Forschungsfragen diskutiert.

Die Ergebnisse aus den Experimenten (vgl. Abschnitt 4 und der Usability-Studie (vgl. Abschnitt 5) zeigen deutlich, dass alle untersuchten Modell-Backbone-Kombinationen, also Florence 2 und Grounding DINO jeweils kombiniert mit Hiera-large sowie Hiera-small, grundsätzlich in der Lage waren, Objekte anhand von textbasierten Prompts meistens zuverlässig zu erkennen und zu segmentieren.

6.1.1 Forschungsfrage 1:

„*Wie wirken sich die in Grounded SAM 2 eingesetzten Modell-Backbone-Kombinationen, Florence 2 bzw. Grounding DINO, jeweils mit Hiera-small oder Hiera-large, auf die Segmentierungsgenauigkeit und Rechenzeit aus?*“

Die erste Frage beschäftigt sich mit dem Vergleich der getesteten Modell-Backbone-Kombinationen hinsichtlich Genauigkeit und Effizienz. Florence 2 erzielte in nahezu allen Metriken, bei Bounding Boxes sowie bei den Masken, bessere Ergebnisse als Grounding DINO. Insbesondere bei komplexeren und längeren Prompts, wie sie im RefCOCO+ und RefCOCO-Datensatz genutzt werden, hatte Grounding DINO häufiger Schwierigkeiten, das korrekte Objekt zu lokalisieren als die Kombinationen mit Florence 2. Bezüglich der Rechenzeit schneidet Florence 2 deutlich schlechter ab als Grounding DINO. Besonders in Kombination mit dem Hiera-large-Backbone braucht Florence 2 deutlich mehr Zeit zur Verarbeitung der Datensätze. Grounding DINO mit Hiera-small war durchgängig die schnellste Konfiguration und lieferte dabei akzeptable Genauigkeit.

Der Einfluss der getesteten Backbones fiel insgesamt geringer aus. Zwar zeigten sich die Kombinationen mit Hiera-large in den meisten Auswertungen leicht überlegen, der Vorsprung gegenüber

Hiera-small war jedoch sehr klein. Einige Ergebnisse deuten auf eine etwas höhere Robustheit von Hiera-small bei kleinen Objekten hin. Aufgrund der geringen Anzahl an Teilnehmenden und der gering ausfallenden Unterschiede lässt sich daraus aber keine allgemeine Aussage ableiten. Bezüglich der anderen Objektgrößen war Hiera-large in den meisten Tests auf den Datensätzen entweder genauso gut oder minimal besser als Hiera-small. Die Wahl des Backbones scheint generell keinen signifikanten Einfluss auf die Qualität der Bounding Boxes zu haben. Auffällig war außerdem, dass alle Kombinationen Schwierigkeiten mit der korrekten Interpretation von Positionsangaben wie „links“ und „rechts“ hatten. Die RefCOCO-Ergebnisse deuten an, dass Florence 2 etwas besser mit solchen Prompteingaben umgehen kann.

Insgesamt hat also die Modellwahl einen größeren Einfluss auf das Ergebnis als die Backbone-Wahl. Je nach Fragestellung kann ein Anwender mit einer schnelleren und ungenaueren (Grounding DINO mit Hiera-small) oder aber mit einer zeitaufwendigeren und dafür genaueren Kombination (Florence 2 mit Hiera-large/Hiera-small) arbeiten. Wobei der Zeitunterschied zwischen Hiera-small und Hiera-large jeweils deutlich, der Genauigkeitsunterschied aber eher gering ausfällt. Der deutlich höhere Zeitaufwand bei Hiera-large dürfte also in der Regel den geringen Gewinn an Genauigkeit nicht aufwiegen.

6.1.2 Forschungsfrage 2

„Welche Unterschiede zeigen sich in der Segmentierungsgenauigkeit für kleine, mittlere und große Objekte zwischen den verschiedenen Modell-Backbone-Kombinationen innerhalb von Grounded SAM 2?“

Um differenzierter auf Unterschiede in der Leistung der Systemkonfigurationen einzugehen, betrachtet dieser Abschnitt die Segmentierungsleistung von kleinen, mittleren und großen Objekten genauer.

Kleine Objekte

Die Ergebnisse der Experimente und der Usability-Studie zeigen, dass kleine Objekte für alle Modell-Backbone-Kombinationen größere Herausforderungen darstellen. Florence 2 in Kombination mit Hiera-small erzielte auf dem COCO-Datensatz bezüglich der Bounding-Boxes bei kleinen Objekten die besten Werte. Bei der Segmentierung lagen die Ergebnisse auf einem vergleichbaren Niveau, wobei Hiera-large geringfügig besser abschnitt. Die anderen Systemvarianten erreichten ähnliche Genauigkeiten. In der Usability-Studie bestätigte sich diese Tendenz weitgehend, zeigt aber insbesondere schwächere Werte für Florence 2 mit Hiera-large auf. Eine Ursache könnte sein, dass kleine Objekte entweder mit dem Prompt nicht präzise genug oder eindeutig beschrieben wurden und dass Hiera-small vielleicht tatsächlich besser für kleine Objekte geeignet ist als Hiera-large.

Mittlere Objeke

Bei mittleren Objekten liefern Florence 2-basierte Kombinationen, sowohl im COCO-Datensatz als auch in der Studie, insgesamt die besten Werte. Hier erreichen Hiera-large und Hiera-small gute Ergebnisse. Die Unterschiede zwischen den Backbones fallen in den COCO-Datensatz-Ergebnissen sehr gering aus. Sollen mittelgroße Objekte in Bildern gefunden werden, ist die Wahl des Modells (Grounding DINO oder Florence 2) also relevanter als die des Backbones (Hiera-large oder Hiera-small).

Große Objekte

Bei großen Objekten zeigen alle Modelle die besten Resultate. Im COCO-Datensatz und besonders in der Usability-Studie zeigt sich, dass Grounded SAM 2 bei ausreichend großen und gut sichtbaren

Objekten besonders zuverlässig arbeitet. Florence 2 erzielt hier mit beiden Backbones durchgängig die höchsten Werte, aber auch Grounding DINO erreicht solide Resultate. Die großen Unterschiede zwischen Grounding DINO und Florence 2 in der Usability-Studie könnten an den weniger standardisierten Prompts liegen. Das Modell kann unter realistischeren und variableren Bedingungen die relevanten Merkmale großer Objekte präziser erfassen und segmentieren, während bei kleinen Objekten die Vorteile durch die begrenzte Bildinformationen weniger stark ins Gewicht fallen. Die Unterschiede zwischen den Hier-Varianten lagen meist nur im einstelligen Prozentbereich.

Empfehlung zur Modell-Backbone-Wahl

Verschiedene Kombinationen liefern in Abhängigkeit von der Objektgröße und je nachdem, ob eine Segmentierungsmaske oder nur eine Bounding Box erstellt werden soll, das beste Ergebnis. Ein Anwender muss je nach Aufgabe und vorliegenden Informationen zur Größe der Objekte die am besten geeignete Kombination wählen. Hat der Anwender keine Informationen zur gesuchten Größe, kann es sinnvoll sein, Bilder mit mehreren Systemvarianten von Grounded SAM 2 zu untersuchen.

6.1.3 Forschungsfrage 3

„*Wie wirkt sich die Nutzung von Grounded SAM 2 im Vergleich zur manuellen Segmentierung auf die Verarbeitungszeit und Effizienz von Segmentierungsaufgaben aus?*“

Inwiefern sich die Nutzung von Grounded SAM 2 im Vergleich zur manuellen Segmentierung auf Verarbeitungszeit und Effizienz auswirkt, wurde anhand der Usability-Studie genauer untersucht. Dabei haben sich klare Vorteile bei dem automatisierten Verfahren gezeigt. Sowohl in Bezug auf Bearbeitungsdauer als auch auf die Segmentierungsgenauigkeit schnitt Grounded SAM 2 vor allem in Kombination mit Florence 2 besser ab als die manuelle Methode. Die durchschnittliche Bearbeitungszeit bei der manuellen Segmentierung war mehr als doppelt so hoch wie die der anderen Modell-Backbone-Kombinationen. Auch die Qualität der Segmentierung sowie der Bounding Boxes war bei Florence 2 höher. Die Ergebnisse der manuellen Segmentierung lagen aber sehr nah an denen von den Kombinationen mit Grounding DINO. Bei genauerer Betrachtung fiel auf, dass es mit der manuellen Methode schwerfällt, saubere Kanten oder präzise zum Objekt passende Bounding Boxes zu erstellen. Auch bei der subjektiven Einschätzung der Teilnehmenden wurden die automatischen Tools insgesamt als hilfreicher und die Ergebnisse als zufriedenstellender eingestuft. Trotzdem ist zu betonen, dass auch die manuelle Methode, insbesondere wenn ausreichend Zeit und Erfahrung vorhanden sind, solide Ergebnisse liefert. Für den Praxiseinsatz ist jedoch offensichtlich die Kombination aus höherer Geschwindigkeit und Präzision von Grounded SAM 2 überlegen.

6.2 Vergleich mit bestehenden Arbeiten

Um die Ergebnisse dieser Arbeit mit bestehenden Ansätzen einzuordnen, wurden Florence 2 (Xiao et al. 2023)und LAVT (Language-Aware Vision Transformer for Referring Image Segmentation) (Yang et al. 2022) ausgewählt. Die Modelle werden zum Teil auf den gleichen Datensätzen getestet und nutzen vergleichbare Metriken zur Auswertung. Sie sind beide in der Lage dazu, Text und Bild als Prompt zu einer Maske zu verarbeiten.
LAVT erreicht auf dem RefCOCO+ Datensatz eine mIoU von 62,14% und liegt damit deutlich über den in dieser Arbeit gemessenen Werten. Dieser Unterschied lässt sich dadurch erklären, dass LAVT im Gegensatz zu Grounded SAM 2 nicht Zero-Shot getestet wurde, sondern auf Datensätzen wie

RefCOCO+ trainiert wurde. Das verdeutlicht, wie viel Potenzial Modelle haben können, wenn sie noch spezifischer an Aufgaben angepasst werden.
Florence 2 wurde im Gegensatz dazu wie die hier eingesetzten Modelle ohne Fine-tuning auf dem RefCOCO-Datensatz ausgewertet und erreicht einen mIoU von 35,8%. Damit liegt Florence 2 unter den hier getesteten Modellen. Auch auf dem COCO 2017-Datensatz, auf dem nur die Bounding Boxes ausgewertet wurden, erreicht Florence 2 AP-Werte von 37,5%. Im Vergleich zeigt sich, dass das Verknüpfen von Florence 2 mit SAM 2 eine deutliche Leistungssteigerung bewirken kann.

7 Fazit und Ausblick

7.1 Zusammenfassung der Ergebnisse

In der vorliegenden Arbeit werden Kombinationen aus Florence 2, Grounding DINO sowie Hiera-large und Hiera-small innerhalb von Grounded SAM 2 und einer manuellen Methode miteinander verglichen. Als wesentliches Ergebnis lässt sich festhalten, dass die KI-gestützte Auswertung der Daten der manuellen überlegen ist, aber es bei den verschiedenen Modell- und Backbone-Kombinationen von der Fragestellung abhängt, welche hier die besten Ergebnisse liefert. So ist Grounding DINO mit Hiera-small am schnellsten, liefert aber oft die schlechtesten Ergebnisse. Florence 2 schneidet bei längeren Prompteingaben mit beiden Hiera-Modellen besser ab als Grounding DINO, ist aber auch um einiges langsamer. Für die Segmentierung kleiner Objekte liefert Florence 2 mit Hiera-large die besten Ergebnisse, während für die Detektion von kleinen Objekten (BBoxen) Florence 2 mit Hiera-small bessere Werte erreicht. Je nach Problemstellung sollte ein Anwender also die geeignete Kombination einsetzen.

7.2 Zukünftige Arbeit

Neben dem Experimentieren mit weiteren Systemvarianten von Grounded SAM 2, zum Beispiel DINO X und weiteren Hiera-Modellen wie Hiera-base und Hiera-tiny, wäre es außerdem interessant, eine Studie durchzuführen, in der alle Teilnehmenden sämtliche Modell-Backbone-Kombinationen testen. Dadurch ließen sich die Studienergebnisse noch besser miteinander vergleichen und es würden subjektive Unterschiede in der Bewertung der Modell-Backbone-Kombinationen verringert werden.

7.2.1 Potentielle Verbesserungen von Grounded SAM 2

Im Bezug auf Grounded SAM 2 wäre es sinnvoll, gezielt auf die Schwächen bei der Erkennung von kleinen Objekten einzugehen. Dafür könnten zum Beispiel gezielt Trainingsdaten mit Fokus auf kleine Instanzen ergänzt werden. Weitere Verbesserungen wären möglich durch robustere Behandlung der Erkennung von Wörtern trotz Rechtschreibfehlern sowie Verarbeitung von Positionsangaben wie „right“, „left“, „top“ oder „bottom“. Außerdem kann es sinnvoll sein, auszuprobieren, SAM 2

noch mit anderen Bounding Box generierenden Modellen zu verknüpfen. Beispielsweise könnte MDETR (Kamath et al. 2021) mit SAM 2 zusammenarbeiten.

7.2.2 Erweiterung Webanwendung

Die bestehende Webanwendung könnte erweitert werden, sodass Nutzende nicht nur textbasierte Prompts eingeben, sondern beispielsweise auch mit einem Punkt markieren können, welches Objekt sie von Grounded SAM 2 segmentiert haben möchten. Da Grounded SAM 2 ebenfalls mit Videodateien arbeiten kann, wäre außerdem eine Einbindung von einem Videoannotationstool in die Anwendung denkbar.

Anhang

A

Zusätzliches Material

A.1 Bilder der Webanwendung

Abbildung A.1: Die Startseite der Webanwendung. Es kann zwischen selber segmentieren und Grounded SAM 2 zum Segmentieren sowie der Studie ausgewählt werden.

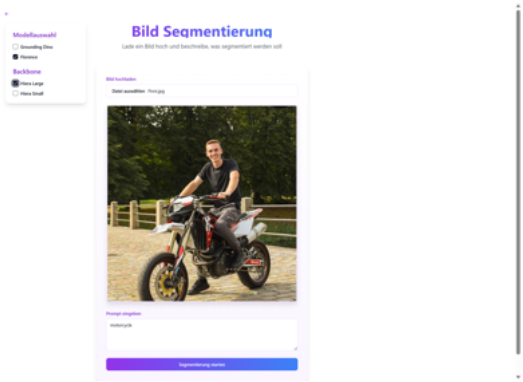

Abbildung A.2: Seite um Objekte mithilfe von Grounded SAM 2 und den verschiedenen Backbones beziehungsweise Modellen zu Segmentieren.

Abbildung A.3: Ergebnisseite der Segmentierung durch Grounded SAM 2. Hier können außerdem die JSON-Datei zur BBox sowie zu der Segmentierungsmaske heruntergeladen werden.

Abbildung A.4: Auf der Seite, kann man selber eine BBox erstellen und Punkte für die Segmentierung wählen.

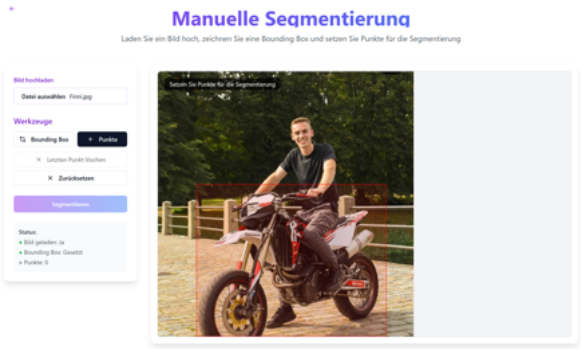

Abbildung A.5: Selber erstellte BBox

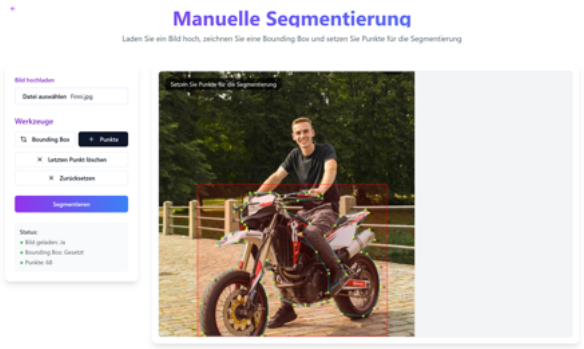

Abbildung A.6: Selber erstellte Umrandung für die Segmentierung mithilfe von gesetzten Punkten

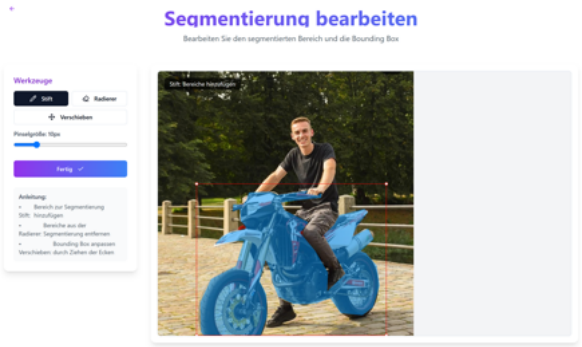

Abbildung A.7: Um die Segmentierung zu bearbeiten kann man mit einem Pinsel oder Radierer weitere Anpassungen machen und auch die BBox noch verschieben.

A.1.1 Screenshots der Webanwendung aus der Studie

Abbildung A.8: Beispiel aus dem ersten Fragebogen der Studie

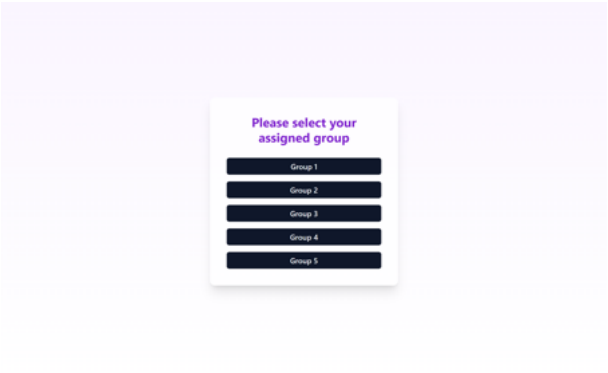

Abbildung A.9: Gruppenauswahl in der Studie. Die Gruppen wurden vorher zugewiesen.

Abbildung A.10: Erklärendes Video zu Bounding Boxes, Segmentierung und wie das jeweilige Tool genutzt werden kann.

Abbildung A.11: Aufgabenbeispiel der Gruppe 1-4

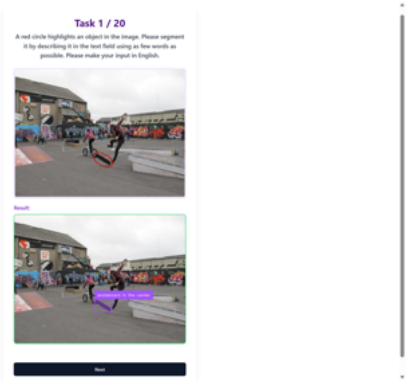

Abbildung A.12: Ergebnisbeispiel der Gruppen 1-4

Abbildung A.13: Aufgabenbeispiel Gruppe 5

A.2 Verwendete Bilder in der Studie

Abbildung A.14: Bild aus der Studie. Hier sollte der Frisbee segmentiert werden. Bild stammt aus dem COCO-Datensatz (Lin et al. 2015)

Abbildung A.15: Bild aus der Studie. Hier sollte der Hydrant im Hintergrund segmentiert werden. Bild stammt aus dem COCO-Datensatz (Lin et al. 2015)

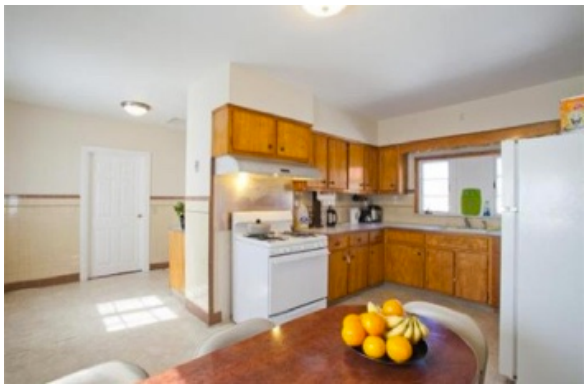

Abbildung A.16: Bild aus der Studie hier sollte die kleine Pflanze hinter der Wand segmentiert werden. Bild stammt aus dem COCO-Datensatz (Lin et al. 2015)

Abbildung A.17: Bild aus der Studie. Hier sollte das Skateboard segmentiert werden. Bild stammt aus dem COCO-Datensatz (Lin et al. 2015)

Abbildung A.18: Bild aus der Studie. Hier sollte einmal der Bus und in einer anderen Aufgabe der Mann der auf dem Bus ist segmentiert werden. Bild stammt aus dem COCO-Datensatz (Lin et al. 2015)

Abbildung A.19: Bild aus der Studie. Hier sollte das Fahrrad segmentiert werden. Bild stammt aus dem COCO-Datensatz (Lin et al. 2015)

Abbildung A.20: Bild aus der Studie. Hier sollte die Schirme rechts segmentiert werden. Bild stammt aus dem COCO-Datensatz (Lin et al. 2015)

Abbildung A.21: Bild aus der Studie. Hier sollte in einer Aufgabe die Giraffe und in einer anderen der Wellensittich segmentiert werden. Bild stammt aus dem COCO-Datensatz (Lin et al. 2015)

Abbildung A.24: Bild aus der Studie. Hier sollte der Koffer von der Person rechts segmentiert werden. Bild stammt aus dem COCO-Datensatz (Lin et al. 2015)

Abbildung A.25: Bild aus der Studie. Hier sollte das Schild im Hintergrund segmentiert werden. Bild stammt aus dem COCO-Datensatz (Lin et al. 2015)

Abbildung A.22: Bild aus der Studie. Hier sollte der Rucksack von der Person in blau segmentiert werden. Bild stammt aus dem COCO-Datensatz (Lin et al. 2015)

Abbildung A.23: Bild aus der Studie. Hier sollte der die Person segmentiert werden. Bild stammt aus dem COCO-Datensatz (Lin et al. 2015)

Abbildung A.28: Bild aus der Studie. Hier sollte die Toilette segmentiert werden. Bild stammt aus dem COCO-Datensatz (Lin et al. 2015)

Abbildung A.29: Bild aus der Studie. Hier sollte der Fußball segmentiert werden. Bild stammt aus dem COCO-Datensatz (Lin et al. 2015)

Abbildung A.26: Bild aus der Studie. Hier sollte die Tastatur segmentiert werden. Bild stammt aus dem COCO-Datensatz (Lin et al. 2015)

Abbildung A.27: Auf der Seite, kann man selber eine BBox erstellen und Punkte für die Segmentierung wählen.

Abbildung A.30: Bild aus der Studie. Hier sollte die Bank segmentiert werden. Bild stammt aus dem COCO-Datensatz (Lin et al. 2015)

Abbildung A.31: Bild aus der Studie. Hier sollte das Rote Licht links neben der Bahn segmentiert werden. Bild stammt aus dem COCO-Datensatz (Lin et al. 2015)

A.3 Hinweis zur Textbearbeitung

Für die Überprüfung der Grammatik und Rechtschreibung wurde in dieser Arbeit das in Overleaf integrierte, KI-gestützte Tool Writefull verwendet.

Literaturverzeichnis

Josh Achiam et al. 2024. GPT-4 Technical Report. arXiv: 2303.08774 [cs.CL]. (Siehe Seite 4).

Tadas Baltrušaitis, Chaitanya Ahuja und Louis-Philippe Morency. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41 (2): 423–443. (Siehe Seite 4).

Alexey Bochkovskiy, Chien-Yao Wang und Hong-Yuan Mark Liao. 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv: 2004.10934 [cs.CV]. (Siehe Seite 13).

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Lorenzo Baraldi, Marcella Cornia und Rita Cucchiara. 2024. The Revolution of Multimodal Large Language Models: A Survey. arXiv: 2402.12451 [cs.CV]. (Siehe Seiten 1 f.).

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski und Armand Joulin. 2021. Emerging Properties in Self-Supervised Vision Transformers, arXiv: 2104.14294 [cs.CV]. (Siehe Seite 16).

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov und Rohit Girdhar. 2022. Masked-attention Mask Transformer for Universal Image Segmentation, arXiv: 2112.01527 [cs.CV]. (Siehe Seite 16).

Kaiming He, Georgia Gkioxari, Piotr Dollár und Ross Girshick. 2018. Mask R-CNN. arXiv: 1703.06870 [cs.CV]. (Siehe Seiten 6, 13).

Sheng He, Rina Bao, Jingpeng Li, Jeffrey Stout, Atle Bjornerud, P. Ellen Grant und Yangming Ou. 2023. Computer-Vision Benchmark Segment-Anything Model (SAM) in Medical Images: Accuracy in 12 Datasets. arXiv: 2304.09324 [eess.IV]. (Siehe Seite 17).

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song und Furu Wei. 2023. Language Is Not All You Need: Aligning Perception with Language Models. arXiv: 2302.14045 [cs.CL]. (Siehe Seite 17).

Wei Ji, Jingjing Li, Qi Bi, Tingwei Liu, Wenbo Li und Li Cheng. 2024. Segment Anything Is Not Always Perfect: An Investigation of SAM on Different Real-world Applications. *Machine Intelligence Research* 21, Nr. 4 (April): 617–630. (Siehe Seite 17).

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra und Nicolas Carion. 2021. MDETR – Modulated Detection for End-to-End Multi-Modal Understanding. arXiv: 2104.12763 [cs.CV]. (Siehe Seiten 16, 55).

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten und Tamara Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* herausgegeben von Alessandro Moschitti, Bo Pang und Walter Daelemans. Doha, Qatar: Association for Computational Linguistics, Oktober. (Siehe Seiten i, 2, 11 f., 22, 26).

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár und Ross Girshick. 2023. Segment Anything. arXiv: 2304.02643 [cs.CV]. (Siehe Seite 9).

Youngwan Lee und Jongyoul Park. 2020. CenterMask : Real-Time Anchor-Free Instance Segmentation. arXiv: 1911.06667 [cs.CV]. (Siehe Seite 13).

Junnan Li, Dongxu Li, Caiming Xiong und Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. arXiv: 2201.12086 [cs.CV]. (Siehe Seite 16).

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology* 22 (140): 5–55. (Siehe Seite 33).

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick und Piotr Dollár. 2015. Microsoft COCO: Common Objects in Context. arXiv: 1405.0312 [cs.CV]. (Siehe Seiten i, 2, 6, 10 f., 15, 18, 22, 26, 32, 39, 60 ff.).

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu und Lei Zhang. 2024. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection, arXiv: 2303.05499 [cs.CV]. (Siehe Seiten i, 2, 8).

Lovable, Inc. 2025. Lovable. AI-powered full-stack development platform, Zugriff am 16.06.2025. https://lovable.dev. (Siehe Seite 30).

Thibaut Lucas. 2023. COCO Evaluation Metrics Explained. Zugriff am 2025-01-02. (Siehe Seiten 13 ff.).

Chun Luo, Jing Zhang, Xinglin Chen, Yinhao Tang, Xiechuan Weng und Fan Xu. 2021. UCATR: Based on CNN and Transformer Encoding and Cross-Attention Decoding for Lesion Segmentation of Acute Ischemic Stroke in Non-contrast Computed Tomography Images. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),* 3565–3568. (Siehe Seite 5).

Shaoni Mukherjee. 2024. SAM 2: Meta's Next-Gen Model for Video and Image Segmentation. Zugriff am 2024-03-19. https://www.digitalocean.com/community/tutorials/sam-2-metas-next-gen-model-for-video-and-image-segmentation. (Siehe Seite 10).

Fuseini Mumuni und Alhassan Mumuni. 2024. Segment Anything Model for automated image data annotation: empirical studies using text prompts from Grounding DINO. arXiv: 2406.19057 [cs.CV]. (Siehe Seite 17).

Jinlai Ning, Haoyan Guan und Michael Spratling. 2023. Rethinking the Backbone Architecture for Tiny Object Detection. In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications,* 103–114. SCITEPRESS - Science / Technology Publications. (Siehe Seite 22).

Rafael Padilla, Wesley L. Passos, Thadeu L. B. Dias, Sergio L. Netto und Eduardo A. B. da Silva. 2021. A Comparative Analysis of Object Detection Metrics with a Companion Open-Source Toolkit. *Electronics* 10 (3): 279. (Siehe Seite 14).

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár und Christoph Feichtenhofer. 2024. SAM 2: Segment Anything in Images and Videos. arXiv: 2408.00714 [cs.CV]. (Siehe Seiten 2, 9 f.).

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang und Lei Zhang. 2024. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. arXiv: 2401.14159 [cs.CV]. (Siehe Seiten 2, 10).

Tianhe Ren und Shuo Shen. 2025. Grounded SAM 2 Repository. https://github.com/IDEA-Research/Grounded-SAM-2. Zugriff am 2024-11-06. (Siehe Seiten i, 2, 7 f., 10).

Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, Jitendra Malik, Yanghao Li und Christoph Feichtenhofer. 2023. Hiera: A Hierarchical Vision Transformer without the Bells-and-Whistles. arXiv: 2306.00989 [cs.CV]. (Siehe Seiten i, 2, 9).

Sachinsoni. 2024. Cross-Attention in Transformer. *Medium,* (siehe Seite 5).

Cristian Santini, Etienne Posthumus, Mary Ann Tan, Oleksandra Bruns, Tabea Tietz und Harald Sack. 2023. Multimodal Search on Iconclass using Vision-Language Pre-Trained Models. arXiv: 2306.16529 [cs.IR]. (Siehe Seite 17).

Richard Szeliski. 2022. Computer Vision: Algorithms and Applications. 2. Aufl. 925. Texts in Computer Science. Second Edition. Springer Cham. (Siehe Seiten 1, 4, 6, 9, 17).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser und Illia Polosukhin. 2023. Attention Is All You Need. arXiv: 1706.03762 [cs.CL]. (Siehe Seite 5).

Shaoru Wang, Yongchao Gong, Junliang Xing, Lichao Huang, Chang Huang und Weiming Hu. 2019. RDSNet: A New Deep Architecture for Reciprocal Object Detection and Instance Segmentation. arXiv: 1912.05070 [cs.CV]. (Siehe Seite 13).

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu und Lu Yuan. 2023. Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks. arXiv: 2311.06242 [cs.CV]. (Siehe Seiten i, 2, 7 f., 13, 17, 52).

Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao und Philip H. S. Torr. 2022. LAVT: Language-Aware Vision Transformer for Referring Image Segmentation. arXiv: 2112.02244 [cs.CV]. (Siehe Seite 52).

Shanliang Yao, Runwei Guan, Xiaoyu Huang, Zhuoxiao Li, Xiangyu Sha, Yong Yue, Eng Gee Lim, Hyungjoon Seo, Ka Lok Man, Xiaohui Zhu und Yutao Yue. 2024. Radar-Camera Fusion for Object Detection and Semantic Segmentation in Autonomous Driving: A Comprehensive Review. *IEEE Transactions on Intelligent Vehicles* 9, Nr. 1 (Januar): 2094–2128. (Siehe Seite 4).

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg und Tamara L. Berg. 2016. Modeling Context in Referring Expressions. arXiv: 1608.00272 [cs.CV]. (Siehe Seite 11).

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou und Pengchuan Zhang. 2021. Florence: A New Foundation Model for Computer Vision. arXiv: 2111.11432 [cs.CV]. (Siehe Seite 16).

Chunhui Zhang, Li Liu, Yawen Cui, Guanjie Huang, Weilin Lin, Yiqian Yang und Yuehong Hu. 2023. A Comprehensive Survey on Segment Anything Model for Vision and Beyond. arXiv: 2305.08196 [cs.CV]. (Siehe Seite 17).

Lu Zhang, Yang Wang, Jiaogen Zhou, Chenbo Zhang, Yinglu Zhang, Jihong Guan, Yatao Bian und Shuigeng Zhou. 2022. Hierarchical Few-Shot Object Detection: Problem, Benchmark and Method. arXiv: 2210.03940 [cs.CV]. (Siehe Seite 13).

Yuxuan Zhang, Tianheng Cheng, Lianghui Zhu, Rui Hu, Lei Liu, Heng Liu, Longjin Ran, Xiaoxin Chen, Wenyu Liu und Xinggang Wang. 2025. EVF-SAM: Early Vision-Language Fusion for Text-Prompted Segment Anything Model. arXiv: 2406.20076 [cs.CV]. (Siehe Seiten 13, 17).

Hao Zhou, Yao He, Xiaoxiao Cui und Zhi Xie. 2024. AGSAM: Agent-Guided Segment Anything Model for Automatic Segmentation in Few-Shot Scenarios. *Bioengineering* 11 (5): 447. (Siehe Seiten i, 17).

Tianfei Zhou, Wang Xia, Fei Zhang, Boyu Chang, Wenguan Wang, Ye Yuan, Ender Konukoglu und Daniel Cremers. 2024. Image Segmentation in Foundation Model Era: A Survey. arXiv: 2408.12957 [cs.CV]. (Siehe Seite 17).