Содержание

1 Введение **4**

2 Обзорно-постановочный раздел работы **8**

2.1 Основные понятия и определения 8

2.2 Постановка задачи 10

2.3 Высокоточная сегментация 11

2.4 Модели семейства SAM и сегментация по подсказкам 13

2.5 Сегментация на основе CLIP и DINO 16

2.6 Обзор наборов данных для сегментации тонких и визуально сложных объектов 18

3 Разработка датасета TAOMR **21**

3.1 Мотивация создания датасета 21

3.2 Состав и структура датасета 21

3.3 Источники изображений и стратегия отбора 22

3.4 Пайплайн аннотирования данных 23

3.5 Статистика и визуализация данных 24

3.6 Контрольная выборка TAOMR-Bench 27

3.7 Заключение 27

4 Метод Talk2SAM **28**

4.1 Мотивация 28

4.2 Общая архитектура метода Talk2SAM 28

4.3 Построение семантической карты релевантности 30

4.4 Интеграция семантической карты в декодер SAM-HQ 33

4.5 Обучение модели 34

4.6 Особенности и преимущества метода 35

5 Вычислительные эксперименты **37**

5.1 Результаты на датасете ThinObject5K 38

5.2 Результаты на датасете BIG 39

5.3 Результаты на DIS5K 40

5.4 Результаты на датасете TAOMR 41

5.5 Разбор ошибок и ограничений 43

6 Заключение **44**

Аннотация

Сегментация объектов на изображениях является одной из приоритетных задач компьютерного зрения и играет важную роль в приложениях, связанных с автономными системами, робототехникой и анализом окружающей среды. Особую сложность для алгоритмов представляют тонкие, вытянутые и слабо выраженные объекты, такие как провода, шланги, лужи и иные препятствия, характерные для городской и парковой среды. Их нечеткие границы, сложная текстура и высокая внутриклассовая вариативность затрудняют автоматическую обработку традиционными методами.

В данной работе рассматривается подход к семантической сегментации таких объектов на основе текстовых описаний. Для этого разработан специализированный датасет **TAOMR** (Traversable Area Objects for Mobile Robotics) [63], включающий более 8 тысяч изображений с аннотациями десяти категорий. Разметка выполнена с использованием гибридного метода, сочетающего автоматическое выделение объектов и ручную корректировку.

Основу предложенного решения составляет метод **Talk2SAM** [64], который объединяет возможности архитектуры SAM-HQ [44] и мультимодальных моделей CLIP [21] и DINOv2 [20]. Текстовые описания объектов трансформируются в эмбеддинги и проецируются в визуальное пространство, что позволяет уточнять маски сегментации с учётом семантического контекста.

Проведённые эксперименты на бенчмарках ThinObject5K [19], DIS5K [27], BIG [13] и TAOMR подтверждают эффективность Talk2SAM, особенно в условиях слабой контрастности и сложной структуры объектов, демонстрируя потенциал интеграции языковых моделей в задачи визуального анализа.

1 Введение

Семантическая сегментация изображений представляет собой одну из фундаментальных задач компьютерного зрения, направленную на разметку каждого пикселя изображения в соответствии с принадлежностью к определённому классу объектов или фону. Успешное решение этой задачи позволяет не только обнаружить наличие объектов на изображении, но и определить их точные границы и форму. В последние годы сегментация стала краеугольным камнем для широкого круга прикладных задач: от автономного вождения и управления беспилотными транспортными средствами до медицинской диагностики, промышленного контроля качества, систем видеонаблюдения и мобильной робототехники.

Особенно важна высокая точность сегментации в задачах, где необходимо учитывать мельчайшие детали сцены — например, в случае движения робота по городской среде, где даже небольшое препятствие (шланг, провод, яма, лужа) может повлиять на траекторию или безопасность перемещения. При этом сложность задачи многократно возрастает в условиях неструктурированной сцены, наличия визуального шума, перекрытий и частично видимых объектов. Таким образом, надёжная сегментация становится критически важной компонентой для построения адаптивных и безопасных автономных систем.

Одной из наиболее проблемных категорий в сегментации являются объекты, обладающие тонкой, изогнутой или нестандартной геометрией. К таким объектам можно отнести провода, велосипедные спицы, трещины, шланги, лужи и другие элементы дорожной сцены. Эти объекты часто занимают небольшую площадь, обладают слабовыраженными границами, низким контрастом с фоном, и могут быть частично перекрыты другими объектами. В результате, даже современные архитектуры глубокого обучения — включая продвинутые модели семейства Segment Anything Model (SAM) [43] — демонстрируют нестабильные результаты при попытке сегментировать подобные элементы.

На этом фоне активно развиваются подходы, основанные на использовании мультимодальных моделей, объединяющих визуальные и языковые представления. В так называемой парадигме vision-language взаимодействия текст выступает в роли семантического запроса, описывающего нужный объект, а система возвращает соответствующую маску или рамку. Это направление находит применение в задачах open-vocabulary сегментации, интерактивной аннотации, визуального поиска и авто-

матической разметки данных. Ключевым преимуществом такого подхода является возможность "человеко-подобного"управления процессом сегментации: пользователь может формулировать запрос в свободной форме — например, «тонкий шланг слева на дорожке» — и получать релевантный результат даже в случае отсутствия точного соответствия в тренировочных данных.

В настоящей работе рассматривается задача семантической сегментации объектов сложной формы по текстовому описанию. Для решения этой задачи был собран специализированный датасет **TAOMR** (Traversable Area Objects for Mobile Robotics), включающий более 8 тысяч изображений с аннотированными объектами, представляющими потенциальные препятствия для мобильных роботов. Среди таких объектов — лужи, провода, дорожные неровности, отходы и другие малозаметные элементы, критичные для безопасного передвижения.

На основе языково-визуального подхода в работе предложен метод **Talk2SAM**, направленный на уточнение сегментации объектов сложной формы с использованием текстовых подсказок. Метод сочетает архитектуру SAM-HQ [44], известную своей способностью к генерации высококачественных масок, с текстовыми эмбеддингами, полученными из модели CLIP [21]. Для согласования пространств признаков используется преобразование текстового описания в пространство DINOv2 [20], что позволяет интегрировать семантическую информацию в процесс визуальной сегментации.

Целью данной работы является разработка, описание и экспериментальное исследование метода Talk2SAM, а также создание и структурирование нового датасета TAOMR для задач сегментации объектов сложной формы в контексте мобильной робототехники. Оценка эффективности метода проводится на ряде открытых бенчмарков и демонстрирует его превосходство над существующими подходами по точности выделения тонких и структурно сложных объектов.

Задачи работы:

В соответствии с поставленной целью были сформулированы следующие основные задачи, решаемые в ходе выполнения исследования:

- *Разработка и сбор специализированного датасета TAOMR*, содержащего изображения объектов дорожной сцены, обладающих тонкой или структурно сложной геометрией. Особое внимание уделяется объектам, представляющим потенциальную опасность для мобильных роботов — лужам, кабелям, проводам, ямам и другим подобным элементам. Для обеспечения разнообразия и репрезен-

тативности данных применялся гибридный подход: автоматическая генерация аннотаций с последующей ручной проверкой и доразметкой.

- *Проектирование и реализация метода Talk2SAM*, сочетающего визуальные признаки модели SAM-HQ с текстовыми эмбеддингами CLIP, трансформируемыми в пространство DINOv2. Необходимо было адаптировать архитектуру к специфике задачи сегментации по текстовому описанию и обеспечить высокую точность сегментации объектов сложной формы в условиях ограниченного визуального контекста и слабых геометрических признаков.

- *Проведение серии вычислительных экспериментов* на существующих общедоступных датасетах, предназначенных для оценки качества сегментации тонких и трудно различимых объектов. В качестве тестовых выборок были выбраны наборы данных ThinObject5K, DIS5K и BIG, обладающие разнообразием сцен и высоким качеством разметки.

- *Сравнительный анализ предложенного метода с существующими решениями*, включая базовые архитектуры семейства SAM, а также модели, использующие визуальные и текстовые подсказки. Сравнение проводится по метрикам mIoU и mbIoU, с акцентом на точность выделения границ объектов и устойчивость к фоновому шуму.

- *Интерпретация результатов и формулирование направлений дальнейших исследований*, включая возможности масштабирования метода на более сложные сцены, интеграцию с системами локализации и управления, а также потенциальные пути повышения качества карт релевантности и генерации масок.

Структура работы:

Работа состоит из пяти основных глав:

- *Глава 1* содержит постановку задачи и обзор современных подходов к сегментации объектов. Особое внимание уделяется трудностям, возникающим при работе с тонкими, вытянутыми и визуально неоднозначными объектами, а также методам, основанным на совместном использовании текстовой и визуальной информации.

- *Глава 2* посвящена разработке датасета **TAOMR**, предназначенного для сегментации объектов, представляющих потенциальную угрозу в задачах мобильной робототехники. Подробно описаны источники данных, стратегия аннотации и внутренняя структура датасета.

- *Глава 3* описывает разработанный метод **Talk2SAM**, направленный на высокоточную сегментацию объектов сложной формы по текстовому описанию. Приведено архитектурное устройство, принципы семантического сопоставления, а также особенности обучения модели.

- *Глава 4* содержит результаты экспериментального исследования, включая описание используемых метрик, настройки тестирования, количественные результаты, а также визуальные примеры сегментации и сравнительный анализ.

- *Глава 5* — **Заключение** — подводит итоги работы, формулирует основные выводы и обозначает направления для дальнейших исследований.

2 Обзорно-постановочный раздел работы

Данный раздел посвящён обзору современных подходов к задаче сегментации изображений, в частности — методов, использующих пространственные и семантические подсказки. Основное внимание уделяется моделям семейства SAM и их модификациям, а также языково-визуальным архитектурам, интегрирующим текстовые запросы, таким как CLIP и DINO. В обзоре рассматриваются сильные и слабые стороны ключевых решений, используемых в задачах open-vocabulary и zero-shot сегментации. Представленный анализ служит обоснованием выбора компонентов предложенного метода Talk2SAM и подводит к формализации исследуемой задачи.

2.1 Основные понятия и определения

В данном разделе приводятся ключевые понятия и определения, используемые в работе, а также обозначения, необходимые для формализации задач и методов сегментации.

- **Семантическая сегментация (semantic segmentation)** — задача компьютерного зрения, заключающаяся в пиксельной классификации изображения, при которой каждый пиксель относится к определённому классу объекта.

- **Объекты сложной формы** — объекты, имеющие вытянутую, неоднородную или фрагментированную структуру, часто с тонкими границами (например, провода, шланги, лужи), которые затрудняют выделение с помощью традиционных алгоритмов сегментации.

- **Open-vocabulary сегментация** — разновидность сегментации, в которой модель не ограничена фиксированным числом предопределённых классов, а способна обрабатывать произвольные текстовые описания объектов. Это требует совмещения языковых и визуальных признаков.

- **Текстовая подсказка (text prompt)** — текстовое описание, указывающее, какой объект следует сегментировать на изображении. Например: “wire on the road”.

- **Карта релевантности (simmilarity map)** — двумерное распределение, отражающее степень соответствия между текстовым запросом и различными об-

ластями изображения. Генерируется на основе сравнения эмбеддингов из языковой и визуальной моделей.

- **Talk2SAM** — предложенный в данной работе метод сегментации по текстовому запросу, использующий модель SAM-HQ в комбинации с эмбеддингами CLIP и признаками DINOv2 для построения карты релевантности и точной генерации масок объектов.

- **TAOMR (Traversable Area Objects for Mobile Robotics)** — специализированный набор данных, разработанный в рамках данной работы, содержащий изображения дорожных сцен с объектами, потенциально опасными для передвижения мобильных роботов.

- **mIoU (mean Intersection over Union)** — стандартная метрика оценки качества сегментации, рассчитываемая как среднее значение отношения площади пересечения предсказанной и истинной маски к площади их объединения по всем изображениям:

$$mIoU = \frac{1}{|D|} \sum_{(I,T)\in D} \frac{|\hat{M} \cap M_{gt}|}{|\hat{M} \cup M_{gt}|} \tag{2.1}$$

- **bIoU (boundary IoU)** [18] — модифицированная версия IoU, чувствительная к качеству предсказания границ объектов. Вычисляется аналогично IoU, но только в пределах полосы фиксированной ширины вдоль контура истинной и предсказанной масок:

$$bIoU = \frac{|\hat{B} \cap B_{gt}|}{|\hat{B} \cup B_{gt}|} \tag{2.2}$$

где $\hat{B}$ и B_{gt} — бинарные маски границ, полученные из соответствующих масок сегментации с помощью морфологических операций.

- **SAM-HQ (Segment Anything Model — High Quality)** — модификация модели SAM, включающая улучшенный декодер и адаптивную обработку признаков, позволяющая достигать высокой точности на задачах сегментации сложных и тонких объектов.

- **CLIP (Contrastive Language-Image Pre-training)** — языково-визуальная модель, обученная на сопоставлении изображений и текстов, позволяющая извлекать эмбеддинги и осуществлять семантический поиск или сопоставление.

- **DINOv2** — мощный визуальный encoder, обученный в стиле self-supervised learning, обеспечивающий высокую выразительность признаков, используемых для построения карты релевантности в методе Talk2SAM.

2.2 Постановка задачи

Рассматривается задача семантической сегментации объектов сложной формы в изображении на основе текстового запроса. Пусть $\mathcal{I}$ — множество изображений, каждое изображение $I \in \mathcal{I}$ содержит объекты из множества классов $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$, представляющих тонкие и структурно сложные объекты (например, “wire”, “hose”, “puddle” и др.).

Каждому изображению I соответствует текстовый запрос T, описывающий целевой класс объекта, и истинная маска сегментации M_T^{gt}, определяющая пиксельную область, соответствующую этому классу. Требуется построить функцию сегментации

$$\Phi : \mathcal{I} \times \mathcal{T} \to \mathcal{M},$$

которая по изображению I и текстовому запросу T генерирует маску $\hat{M} = \Phi(I, T)$, приближающую истинную маску M_T^{gt}.

Качество сегментации оценивается с использованием следующих метрик:

- **IoU (Intersection over Union)** — мера перекрытия между предсказанной и истинной масками:
$$IoU(\hat{M}, M^{\text{gt}}) = \frac{|\hat{M} \cap M^{\text{gt}}|}{|\hat{M} \cup M^{\text{gt}}|};$$
- **mean IoU (mIoU)** — среднее значение IoU по всем изображениям:
$$mIoU = \frac{1}{|\mathcal{D}|} \sum_{(I,T)\in\mathcal{D}} IoU(\Phi(I, T), M_T^{\text{gt}});$$
- **boundary IoU (bIoU)** — модифицированная метрика, уделяющая большее внимание точности предсказания границ объекта. Она рассчитывается аналогично IoU, но только в узкой полосе вдоль контура объекта:
$$bIoU = \frac{|\hat{B} \cap B^{\text{gt}}|}{|\hat{B} \cup B^{\text{gt}}|},$$
где $\hat{B}$ и B^{gt} — бинарные маски, представляющие границы предсказанной и истинной масок, полученные через морфологические операции (например, градиент или сужение).

Метрика bIoU особенно важна в задачах сегментации объектов тонкой структуры, где стандартный IoU может не учитывать точные ошибки на границах, критичные для восприятия.

Особенность данной постановки состоит в том, что класс объекта не выбирается из конечного множества меток, а задаётся произвольным текстовым описанием, что требует использования моделей, способных к семантическому сопоставлению языка и визуального контекста.

Задача усугубляется следующими факторами:

- Объекты могут быть частично перекрыты другими;
- Геометрия масок может быть фрагментированной или состоять из узких длинных контуров;
- Семантические различия между близкими классами (например, “hose” vs “wire”) могут быть слабо выражены визуально.

Таким образом, задача требует построения модели, обладающей следующими свойствами:

1. Интеграция текстового и визуального признакового пространства;
2. Высокая пространственная разрешающая способность на тонких структурах;
3. Обобщаемость на произвольные текстовые категории.

В работе данная задача решается с использованием метода Talk2SAM, в котором текстовые эмбеддинги используются как семантические подсказки для генерации масок на основе архитектуры SAM-HQ.

2.3 Высокоточная сегментация

Классические подходы к сегментации высокого качества, как правило, ориентированы на конкретные задачи в закрытом мире — такие как instance-сегментация изображений и видео [16, 31, 34, 23, 32], семантическая сегментация [7, 9, 11, 17] или паноптическая сегментация [37]. Эти методы демонстрируют высокую точность, но требуют предварительного обучения на фиксированных наборах классов и плохо адаптируются к новым категориям без дополнительной настройки.

Для повышения качества сегментации нередко применяются методы постобработки. Одним из таких подходов являются условные случайные поля (Conditional Random Fields, CRF) [1], позволяющие уточнять границы масок на основе локальных цветовых и текстурных признаков. Другие методы используют эвристические стратегии, такие как последовательное расширение областей (region growing) [10]. Однако такие способы зависят от низкоуровневых признаков изображения и не учитывают глобальный семантический контекст, что ограничивает их эффективность в случае крупных ошибок сегментации.

Альтернативой выступают каскадные архитектуры, применяющие несколько этапов итеративного уточнения масок с использованием глубоких нейросетей [13, 26]. Несмотря на улучшение качества сегментации, такие модели увеличивают вычислительную нагрузку и склонны к переобучению на конкретных наборах данных.

Значительный прогресс в направлении высококачественной универсальной сегментации был достигнут с появлением модели **SAM-HQ** [44]. В отличие от моделей, применяющих внешние блоки уточнения, SAM-HQ сохраняет zero-shot способности оригинальной модели Segment Anything (SAM), но модифицирует её архитектуру таким образом, чтобы повышать детализацию масок на этапе предсказания, а не постобработки. В частности, в SAM-HQ предлагается модуль эффективного обучения токенов, который позволяет точнее описывать границы объектов и восстанавливать мелкие детали без увеличения числа параметров. Архитектура модели приведена на рис. 1.

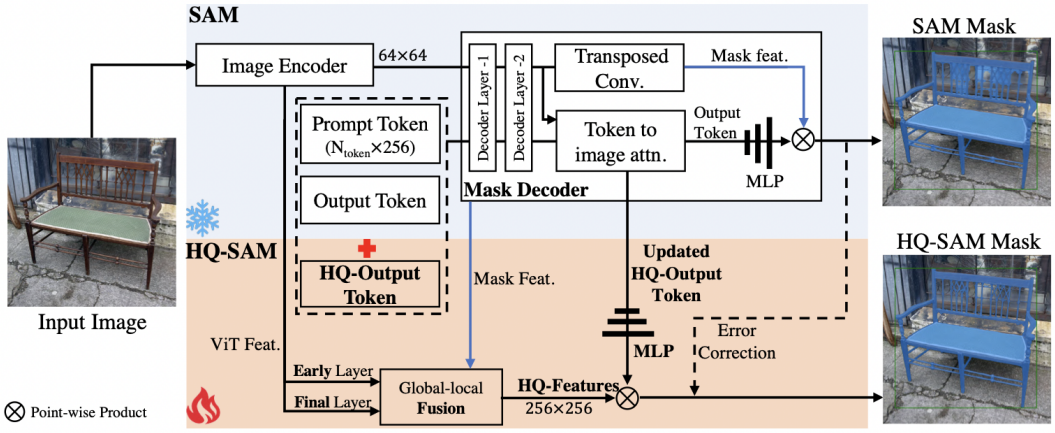

Рис. 1: Архитектура SAM-HQ

В отличие от методов с отдельными блоками refinement, SAM-HQ напрямую

генерирует высококачественную маску, повторно используя энкодер и декодер оригинального SAM. Такой подход сохраняет универсальность и способность обобщения на новые задачи без переобучения, что делает его особенно перспективным в контексте построения zero-shot систем, ориентированных на разнообразные сцены и типы объектов.

Таким образом, современные архитектуры сегментации высокого качества стремятся минимизировать зависимость от жёстко заданных классов и этапов уточнения, обеспечивая точное и устойчивое выделение объектов в широком спектре условий.

2.4 Модели семейства SAM и сегментация по подсказкам

Появление модели **Segment Anything Model (SAM)** [43] стало важным этапом в развитии универсальных систем сегментации изображений. В отличие от традиционных моделей, обученных на ограниченном наборе классов, SAM демонстрирует способность выделять объекты произвольных категорий без дополнительного обучения, используя только простые и интуитивно понятные пространственные подсказки — такие как ключевые точки, прямоугольные рамки или грубые маски. Эта архитектура основана на мощном визуальном энкодере на базе Vision Transformer (ViT) [12] и компактном декодере, обрабатывающем embedded-промпты, что обеспечивает высокую эффективность и гибкость. Благодаря использованию обучающего набора из 11 миллионов изображений и более 1 миллиарда масок, SAM показывает исключительную обобщающую способность и zero-shot производительность на множестве задач сегментации.

Благодаря своей универсальности, SAM получила широкое распространение в различных прикладных и исследовательских направлениях. Модель успешно применяется в медицинской визуализации [39, 47, 46, 45], где точность сегментации имеет критическое значение, а также в задачах выделения трудноразличимых объектов, включая замаскированные [47] и малозаметные структуры. Кроме того, SAM используется как базовый компонент для построения более сложных систем: от мультимодального редактирования изображений [38] и генеративных моделей, до семантической сегментации [39] и отслеживания объектов на видео [48, 42]. Этот спектр применений подчёркивает гибкость архитектуры и её способность адаптироваться к новым доменам и типам входных данных.

Несмотря на выдающиеся возможности, SAM не лишена ограничений. Наиболее

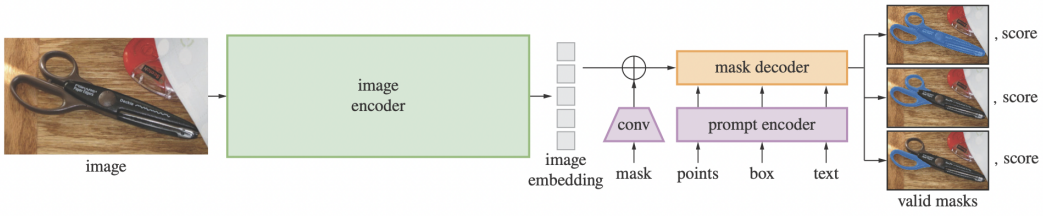

Рис. 2: Архитектура Segment Anything Model (SAM)

заметные трудности возникают при работе с объектами сложной геометрии — например, с тонкими, вытянутыми, пересекающимися или слабо выраженными структурами, такими как провода, шланги, лужи и ямы. В подобных случаях стандартный механизм генерации масок может не обеспечивать достаточную детализацию, особенно в условиях низкого контраста или сложного фона. Ограниченная точность границ масок снижает применимость SAM в задачах, где требуется пиксельная точность и устойчивость к структурной неоднозначности.

Для повышения качества сегментации таких объектов были предложены усовершенствованные версии оригинальной модели, включая SAM-HQ, PA-SAM [57], RobustSAM [59] и HRSAM [56]. Эти модели ориентированы на генерацию более точных и устойчивых масок, не теряя при этом универсальности и способности к zero-shot сегментации. Улучшения достигаются за счёт внедрения дополнительных слоёв обработки, модификации токенов внимания или усиления выходных признаков. Такие модификации позволяют существенно повысить чувствительность к объектам со сложной формой и размытыми границами, что критически важно для робототехнических и навигационных систем, работающих в естественной среде.

Тем не менее, даже усовершенствованные модели на базе SAM могут испытывать затруднения при сегментации семантически близких или визуально неоднозначных объектов в отсутствие чётких пространственных подсказок. В случаях, когда объекты имеют схожую текстуру, форму или цветовую гамму, точность масок может снижаться, особенно при использовании минимальных промптов (например, одной точки).

Одним из направлений, позволяющих преодолеть эти ограничения, стало использование дополнительных модальностей или источников информации. Так, модель **VRP-SAM (Visual Reference Prompt SAM)** [62] предлагает выделять объекты на основе сопоставления с референсным изображением, содержащим аналогичный объект. Такой подход позволяет уточнять сегментацию в условиях визуальной

неопределённости, особенно когда отсутствует чёткая геометрическая информация. Однако практическое применение VRP-SAM ограничено необходимостью предоставления заранее размеченного и подходящего эталонного изображения.

Дальнейшее развитие получила интеграция SAM с языковыми моделями. В частности, **SAM-CLIP** [60] сочетает визуальные признаки из SAM с семантическими эмбеддингами из модели CLIP, позволяя проводить сегментацию объектов по текстовым описаниям. Такой механизм обеспечивает управление сегментацией на основе высокоуровневой семантики, что особенно полезно в задачах open-vocabulary. Тем не менее, качество результата сильно зависит от точности текстового эмбеддинга и полноты описания.

Модель **CAT-SAM** [52] предлагает механизм адаптации базовой архитектуры SAM к новым категориям с использованием лишь нескольких размеченных примеров. В отличие от подходов, основанных на текстовых или категориальных подсказках, CAT-SAM реализует визуальную адаптацию через модуль условной настройки, который позволяет эффективно подстраивать маски под заданный класс при помощи визуального прототипа. Это делает модель особенно ценной для применения в условиях ограниченной аннотированной выборки. Тем не менее, метод требует наличия хотя бы нескольких размеченных изображений для каждой новой категории, что ограничивает его применение в полностью zero-shot сценариях.

Таким образом, развитие семейства моделей SAM направлено на устранение ограничений базовой архитектуры, связанных с точностью, устойчивостью и семантическим управлением. Каждое из предложенных решений расширяет область применимости SAM в различных условиях, однако ни одно из них не является универсальным. Модели **SAM-HQ** и **HRSAM** повышают структурную детализацию, **VRP-SAM** усиливает семантическую устойчивость к неоднозначности, **SAM-CLIP** позволяет управлять сегментацией с помощью текстовых запросов, а **CAT-SAM** предлагает эффективную адаптацию к новым классам при помощи немногих примеров. В рамках настоящей работы архитектура **SAM-HQ** выбрана в качестве базовой, поскольку демонстрирует оптимальный баланс между качеством масок и универсальностью при ограниченных вычислительных затратах.

2.5 Сегментация на основе CLIP и DINO

За последние годы модели мультимодального предварительного обучения, такие как CLIP, получили широкое распространение благодаря своей способности к обобщению и переносу знаний между модальностями. CLIP [21], обученный на больших объёмах пар изображение–текст с использованием контрастивной функции потерь, демонстрирует выдающиеся результаты в задачах классификации и поиска изображений в zero-shot сценариях. Однако глобальная природа сопоставления CLIP, направленная на выравнивание целых изображений и текстов, приводит к ограниченной точности в задачах, требующих плотных предсказаний на уровне пикселей. Это делает применение CLIP в задачах локализации и сегментации менее эффективным [30].

Одной из первых моделей, использующих CLIP для генерации сегментационных масок на основе текстового запроса, стала **CLIPSeg** [30]. Она применяет декодер к эмбеддингам текста и изображения, получая бинарную маску объекта, соответствующего запросу. CLIPSeg показывает хорошие результаты в open-vocabulary сценариях, но ограничена по разрешению выходных масок и слабо справляется с сегментацией тонких объектов.

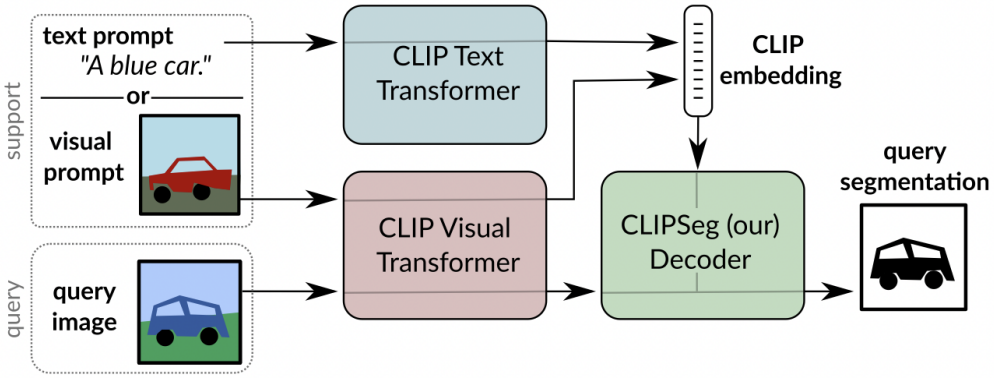

Рис. 3: Архитектура CLIPSeg

Для улучшения пространственной привязки были предложены модели, такие как **DenseCLIP** [25], преобразующие текстовые эмбеддинги в плотные карты признаков. Этот подход позволяет улучшить семантическое сопоставление между текстом и изображением, но требует значительных вычислительных ресурсов.

Более универсальный способ совмещения текстовых и визуальных признаков ре-

ализован в **X-Decoder** [36], который решает задачи сегментации, классификации и captioning в едином фреймворке. Несмотря на гибкость, X-Decoder уступает в точности при выделении объектов с детальной структурой, особенно в условиях неоднозначной сцены.

Параллельно с развитием мультимодальных моделей происходил прогресс в области самообучающихся визуальных представлений. Модельное семейство **DINO** [20] и особенно **DINOv2** [51] обучается без использования разметки и демонстрирует высокую способность к локализации объектов за счёт свойств self-attention в Vision Transformer. В частности, наблюдается корреляция между активациями внимания и регионами переднего плана, что позволяет использовать DINO в задачах бесконтурной (unsupervised) сегментации объектов [22, 49, 35].

Недавние работы предложили объединять сильные стороны CLIP и DINO для построения open-vocabulary сегментационных систем. Среди них — ReCo [33], ProxyCLIP [58], CLIP-DINOiser [53], FOSSIL [54], где визуальные и текстовые пространства согласуются через обучение либо без обучения (training-free), либо с использованием модулей проекции. Однако многие из этих методов требуют дополнительных визуальных энкодеров или сложной настройки на целевом домене.

Особое место среди них занимает модель **Talk2DINO** [61], сочетающая мультимодальные возможности CLIP и сильные локализационные свойства DINOv2. Talk2DINO напрямую отображает текстовые эмбеддинги CLIP в визуальное пространство DINOv2, формируя карту семантической релевантности. Такая карта позволяет точно определить расположение интересующего объекта на изображении, заданного в виде текстового запроса, и затем использовать её для генерации маски сегментации (см. рис. 4). В отличие от ProxyCLIP и схожих методов, Talk2DINO не требует нескольких визуальных энкодеров и достигает высокого качества без дообучения на внешних источниках данных.

Именно Talk2DINO используется в данной работе для построения карты релевантности между изображением и текстом, благодаря высокой точности семантического сопоставления, эффективности и пригодности к интеграции в более сложные архитектуры сегментации. Таким образом, комбинация CLIP и DINO открывает новые возможности в задачах семантической сегментации с текстовым управлением. CLIP обеспечивает мощное семантическое обобщение, а DINO — точную локализацию. Их объединение в рамках таких моделей, как Talk2DINO, позволяет добиться баланса между универсальностью, качеством масок и гибкостью применения в open-

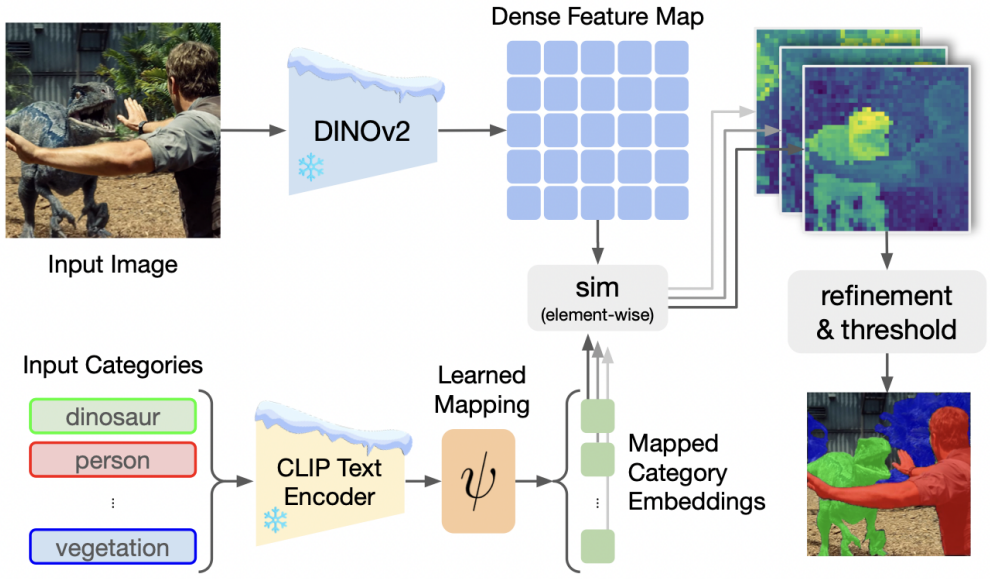

Рис. 4: Архитектура Talk2DINO

vocabulary сценариях.

2.6 Обзор наборов данных для сегментации тонких и визуально сложных объектов

Для оценки эффективности метода сегментации, ориентированного на выделение объектов сложной формы, были выбраны как специализированные, так и универсальные датасеты. Это позволяет не только протестировать модель на предмет обработки вытянутых, тонких и визуально неоднозначных структур, но и оценить её устойчивость к разнообразию визуальных условий. Датасеты варьируются по сложности сцен, количеству классов, качеству аннотаций и типу объектов, что делает их применение особенно релевантным для задач...

Ключевым источником данных является **ThinObject5K** — специализированный набор изображений, созданный для исследования сегментации тонких и вытянутых объектов, таких как провода, шланги, трещины, ветви деревьев и другие малозаметные элементы. Свыше 5 700 изображений с высокоточными масками позволяют провести детальную валидацию способности модели к точному выделению границ объектов с малым поперечным сечением. Удобной особенностью является наличие текстовых описаний в именах файлов, что упрощает ...

Ещё одним важным набором является **DIS5K** — датасет с ручной аннотаци-

ей объектов в условиях сложных фонов и с нечеткими границами. Маски создавались с особой точностью, без использования промежуточных меток неопределённости (`void`), что делает его эталоном в задачах плотной и точной сегментации. Объекты здесь зачастую трудноотделимы от фона, обладают сложной формой и не всегда явно выделяются, что особенно ценно для тестирования мультимодальных архитектур.

Датасет **BIG** (Boundary Inference Groundtruth) содержит всего 150 изображений, однако отличается беспрецедентно высоким качеством аннотаций. Он используется в качестве строго бенчмарка для проверки точности масок и способности моделей выделять сложные границы. Его высокая разрешающая способность и разнообразие объектов делают его важным дополнением к более объёмным наборам.

Для оценки способности модели к обобщению также были включены более общие датасеты. **FSS-1000** [14] предоставляет по 10 изображений на каждую из 1 000 категорий и ориентирован на задачу few-shot сегментации. Маски в этом наборе охватывают широкий спектр объектов — от простых до высоко структурированных, включая сложные по форме и слабо выраженные на фоне элементы. Он особенно важен при оценке моделей в условиях ограниченного количества обучающих примеров.

Крупномасштабный набор **DUTS** [2] содержит более 15 000 изображений и является одним из наиболее обширных среди датасетов сегментации салиентных объектов. Его сцены включают разнообразные композиции, от простых до визуально насыщенных. Маски, размеченные вручную, дают хорошую основу для анализа устойчивости модели к изменению условий и сложности визуальной сцены.

Более компактный по размеру, но концептуально ценный **ECSSD** [5] включает 1 000 изображений, в которых представлены визуально сложные и насыщенные сцены с высокой семантической нагрузкой. Объекты могут быть частично перекрыты или иметь размытые границы. Разметка была проведена несколькими экспертами, что повышает её достоверность и делает датасет подходящим для оценки качества масок в неоднозначных условиях.

Наконец, **MSRA-10K** [3] остаётся классическим датасетом, широко применяемым в задачах базовой сегментации. Он содержит 10 000 изображений с относительно простыми аннотациями и визуально отчётливыми объектами. Несмотря на это, он полезен для первичной калибровки моделей и валидации на базовом уровне, прежде чем переходить к более сложным задачам.

Для оценки качества сегментации с использованием языковых подсказок были выбраны три набора данных: **ThinObject5K**, **DIS5K** и **BIG**. Особенность этих

Таблица 1: Объём описанных наборов данных

Датасет	**Количество изображений**
ThinObject5K	5 700+
DIS5K	5 000+
BIG	150
FSS-1000	10 000
DUTS	15 000+
ECSSD	1 000
MSRA-10K	10 000

датасетов заключается в том, что к каждому изображению можно ассоциировать уникальный текстовый запрос, соответствующий целевому объекту, что делает их пригодными для open-vocabulary сегментации. Примеры изображений и соответствующих запросов приведены на рис. 5.

Таким образом, совокупность выбранных датасетов охватывает широкий спектр условий и типов объектов — от структурно сложных до визуально незаметных. Это обеспечивает всестороннюю проверку модели на способность точно сегментировать неоднозначные и нетривиальные по геометрии элементы в разнообразных контекстах.

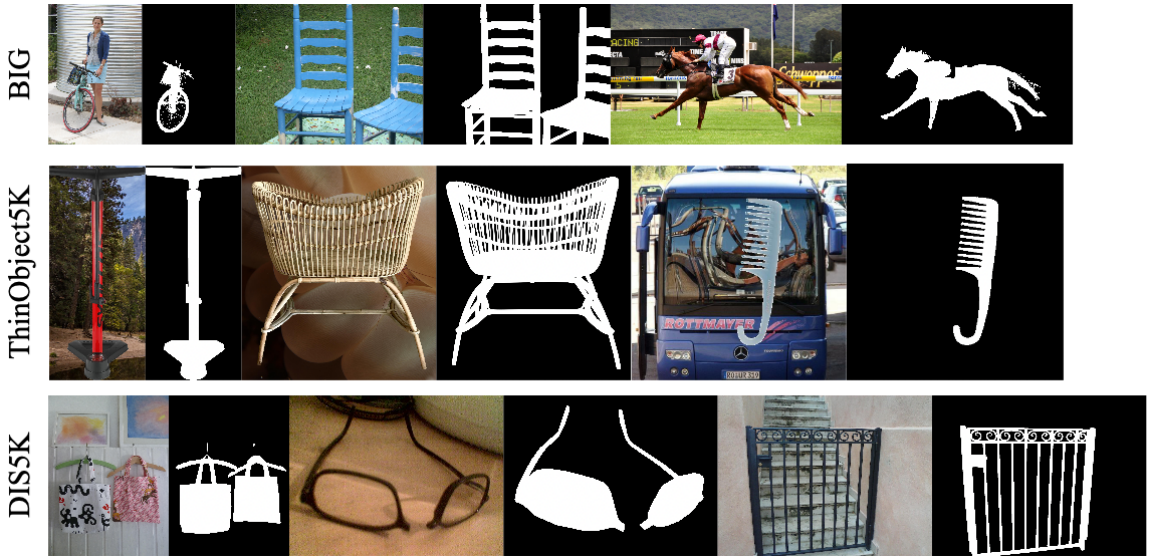

Рис. 5: Примеры изображений из наборов данных ThinObject5K, DIS5K и BIG с соответствующими текстовыми запросами

3 Разработка датасета TAOMR

3.1 Мотивация создания датасета

Разработка надежных алгоритмов восприятия для мобильных роботов требует специализированных наборов данных, охватывающих визуально сложные и малозаметные объекты, способные нарушить безопасность автономного движения. Существующие общие датасеты (такие как COCO [4], ADE20K [8], Cityscapes [6]) не включают ключевые категории объектов-препятствий, встречающихся в реальных условиях: провода, шланги, лужи, отходы, выбоины. В то же время, такие объекты имеют критическое значение в задачах локальной навигации в парках, на тротуарах и в неструктурированной городской среде.

Датасет **TAOMR** был создан с целью заполнения этого пробела. Он включает сцены с характерной структурой городской и парковой среды и аннотирован высокоточными масками тонких, структурно неоднородных объектов, затрудняющих движение мобильных платформ. Кроме прямого фокуса на препятствия, в состав разметки включены и фрагменты проходимых поверхностей, что позволяет использовать TAOMR для комплексной сегментации проходимости.

3.2 Состав и структура датасета

TAOMR содержит **8 823 изображения** с детализированной пиксельной разметкой **75 479 масок**, распределённых по **10 категориям**, охватывающим как препятствия, так и фоновые элементы. Все категории представляют собой визуально различимые объекты, критичные для задач безопасной навигации мобильных роботов в городских и природных условиях. Классы в датасете организованы следующим образом:

- **Препятствия:** *wire, hose, fire hose, puddle, animal waste, bump*;
- **Фоновая сцена:** *road, sidewalk, terrain, vegetation.*

Такое деление отражает реальные условия движения автономных мобильных платформ, где важно не только точно обнаруживать потенциальные препятствия, но и различать проходимые области. Отсутствие абстрактных или обобщённых категорий позволяет использовать датасет как для задач сегментации, так и для последующего использования в навигационных и планировочных модулях.

3.3 Источники изображений и стратегия отбора

Формирование датасета TAOMR основывалось на сборе визуально релевантных и разнообразных изображений, отражающих реальные сценарии, в которых объекты могут представлять опасность для навигации мобильных роботов. Основным источником изображений выступили поисковые системы **Яндекс** и **Baidu**, обладающие функциональностью *поиска по изображению* — механизмом, позволяющим находить визуально схожие изображения на основе загруженного примера.

Для каждого целевого класса вручную подбирались **эталонные изображения**, демонстрирующие характерные признаки соответствующего объекта (например, тонкий провод, лужа, шланг, дорожный дефект). Эти изображения загружались в поиск по изображению, после чего автоматически собирались визуально схожие результаты. Такой подход показал высокую эффективность по сравнению с текстовым поиском, поскольку последнему зачастую свойственно возвращать нерелевантные или художественные изображения, не пригодные для целей аннотации.

Полученные изображения проходили **ручную фильтрацию** с учётом следующих критериев:

- наличие нужного объекта на изображении;
- сцена с реальной перспективой, характерной для мобильных роботов;
- разнообразие условий освещения, текстур, фонов.

Помимо изображений из поисковых систем, в состав датасета были включены изображения из двух открытых источников — **NPO** [15] и **InstSnowRoadDetection** [41]. Эти данные были вручную доразмечены в соответствии с категоризацией TAOMR, включая детальную сегментацию объектов и фоновых областей.

Таким образом, выбранная стратегия сбора изображений позволила получить разнообразный и сбалансированный датасет, отражающий широкий спектр реальных ситуаций, с которыми может столкнуться мобильный робот при передвижении в городской или природной среде.

3.4 Пайплайн аннотирования данных

Процесс разметки изображений в TAOMR реализован в виде полуавтоматического пайплайна, объединяющего современные модели сегментации с процедурами постобработки и ручной проверки. Общая схема аннотирования и этапов обработки представлена на рисунке 6.

1. **Обнаружение объектов.** Для получения ограничивающих рамок (bounding boxes) объектов-целей использовалась модель Grounding DINO [55], способная находить объекты по текстовому описанию.

2. **Сегментация объектов.** Внутри полученных рамок применялась модель SAM-HQ для генерации детализированных масок объектов, особенно эффективная при работе с вытянутыми или неоднородными структурами.

3. **Сегментация фоновых классов.** Для классов «дорожная поверхность», «тротуар», «растительность» и «грунт» применялась модель OneFormer [40], обученная на датасетах Mapillary и ADE20K. После предсказания классы сопоставлялись с категоризацией TAOMR с помощью процедуры remapping.

4. **Морфологическая фильтрация и слияние.** На этом этапе маски проходили пространственную постобработку для устранения артефактов. Объекты, занимающие слишком малую площадь, автоматически удалялись как нерелевантные. Для устранения разрывов внутри масок применялись морфологические операции (замыкание, заполнение дыр), что позволяло улучшить целостность сегментированных областей. Также выполнялось объединение пересекающихся фрагментов одного класса и проверка на логические конфликты между различными категориями.

5. **Ручная корректировка.** Для изображений из внешних источников выполнялась ручная верификация и коррекция масок в интерфейсе CVAT с целью повышения точности аннотаций.

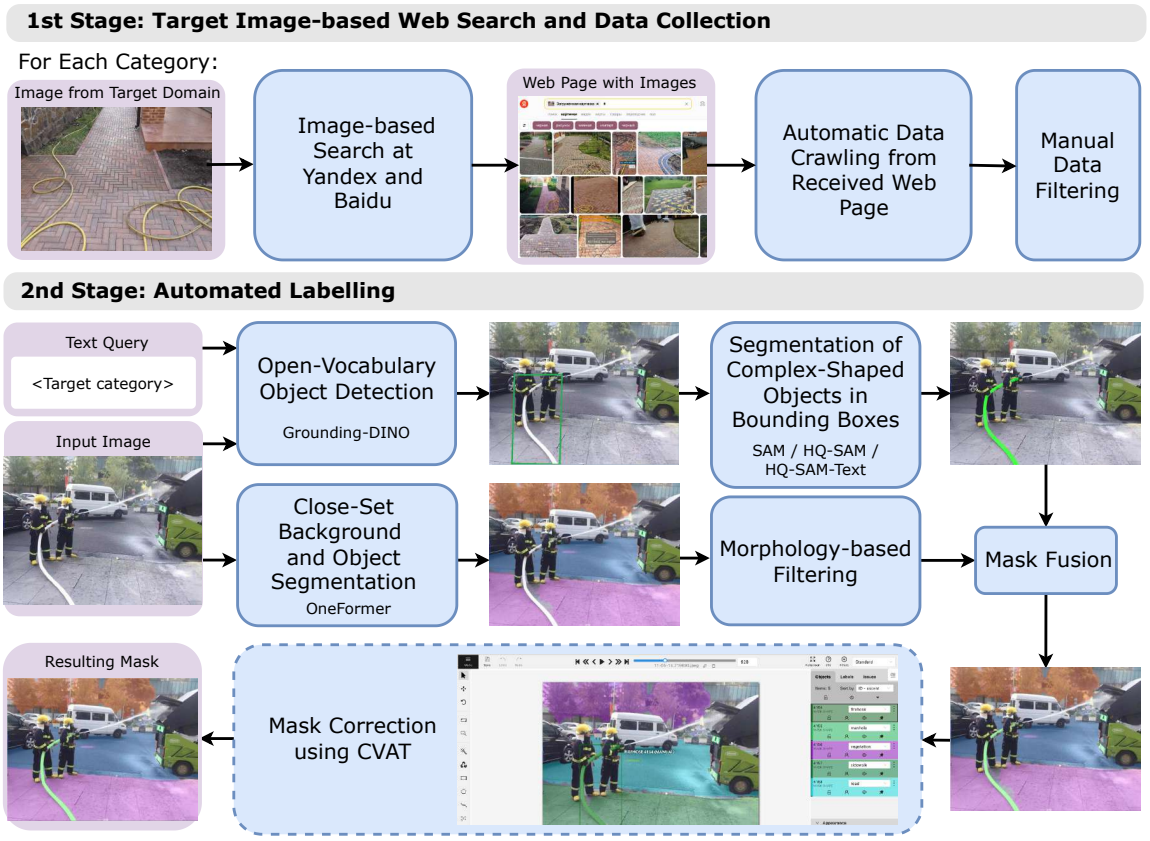

Рис. 6: Пайплайн сбора и аннотации данных TAOMR

3.5 Статистика и визуализация данных

Итоговая версия датасета TAOMR включает сбалансированный по категориям набор изображений и соответствующих масок. Всего в датасете представлено **8 823 изображения** и **75 479 масок**, что в среднем составляет около 8–9 объектов на кадр. Некоторые категории, такие как *road bump* и *waste*, характеризуются высокой плотностью объектов: на одном изображении может присутствовать сразу несколько экземпляров одного класса. Это важно для повышения устойчивости моделей к множественным экземплярам объектов, а также для улучшения генерализации в условиях сложной городской среды.

В таблице 2 представлено подробное распределение изображений и масок по категориям и источникам данных, включая поисковые системы, специализированные наборы и дополнительные внешние источники.

Таблица 2: Распределение изображений и масок по источникам и категориям в TAOMR

Категория	Яндекс, Baidu	NPO	InstSnow	Итого
Fire hose	272 / 387	0 / 0	0 / 0	272 / 387
Hose	269 / 328	0 / 0	0 / 0	269 / 328
Wire	220 / 231	0 / 0	0 / 0	220 / 231
Animal waste	239 / 1455	0 / 0	0 / 0	239 / 1455
Puddle	273 / 520	0 / 0	0 / 0	273 / 520
Road bump	2059 / 3174	2749 / 13964	84 / 216	5202 / 20005
Sidewalk	1802 / 1802	2706 / 2706	243 / 243	4751 / 4751
Terrain	1753 / 1753	2839 / 2839	320 / 320	4920 / 4920
Vegetation	1711 / 1711	4981 / 4981	442 / 442	7339 / 7339
Road	2444 / 2444	5000 / 5000	442 / 442	8126 / 8126
Other negative	0 / 0	4571 / 10527	36 / 130	4609 / 10659
Other positive	0 / 0	3087 / 11788	437 / 4227	3718 / 16758
Всего	**3141 / 13805**	**5000 / 51805**	**442 / 8455**	**8823 / 75479**

Кроме того, на рисунке 7 приведены примеры изображений и соответствующих аннотаций, демонстрирующие разнообразие сцен и объектов, охваченных в датасете. Такая визуализация позволяет наглядно оценить уровень детализации и точность разметки, достигнутые в результате предложенного пайплайна аннотирования.

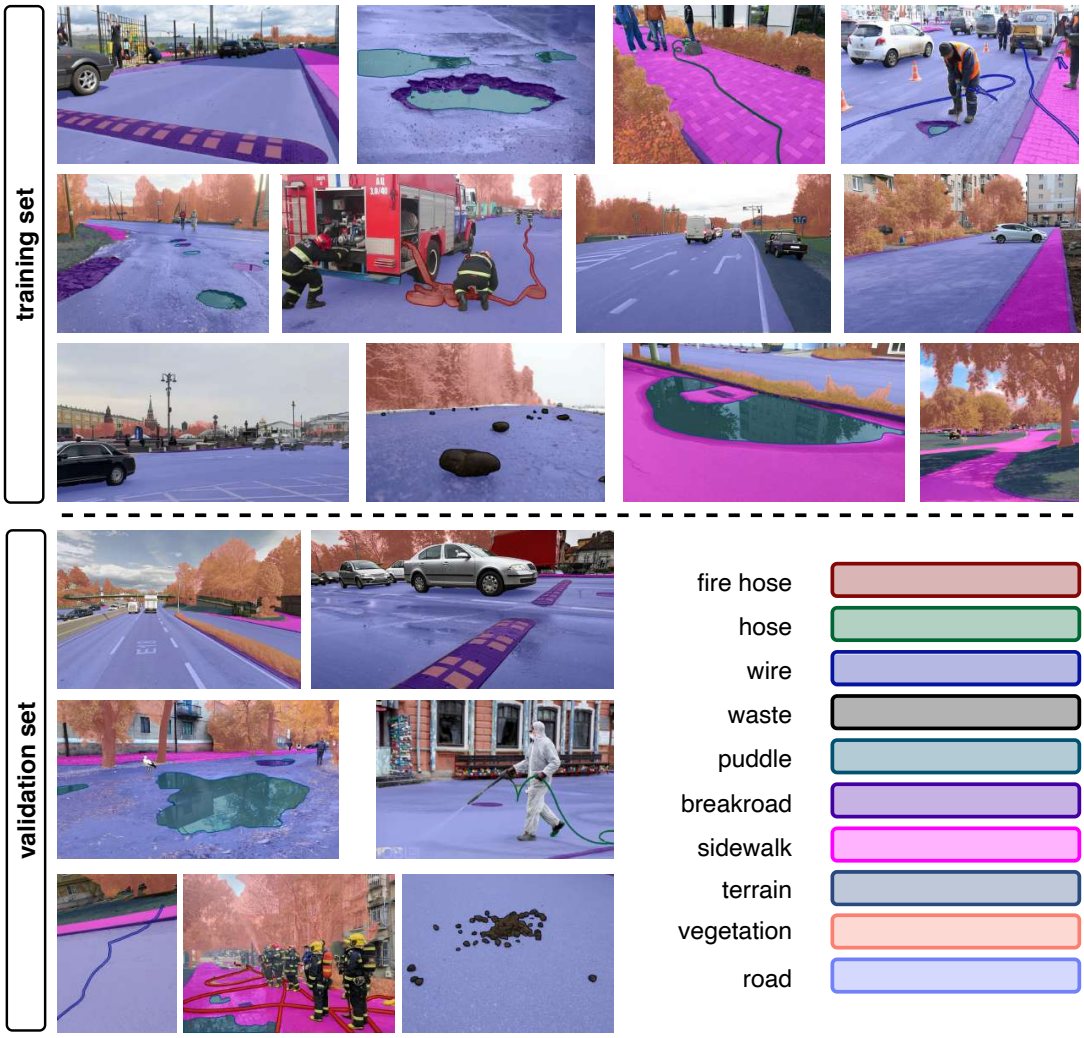

Рис. 7: Примеры изображений и аннотаций из датасета TAOMR

3.6 Контрольная выборка TAOMR-Bench

Для тестирования алгоритмов сегментации в условиях, приближенных к реальному миру, был сформирован отдельный поднабор TAOMR-Bench. Он включает 529 изображений, сбалансированных по категориям, прошедших дополнительную валидацию и ручную правку.

Таблица 3: Распределение масок по категориям в TAOMR-Bench

Категория	Обучающая выборка	Валидационная выборка
Fire hose	59	16
Hose	44	11
Wire	96	24
Waste	508	144
Puddle	226	43
Road bump	1106	271
Sidewalk	867	193
Terrain	896	267
Vegetation	1640	425
Road	760	157
Всего масок	**6202**	**1551**

3.7 Заключение

TAOMR представляет собой полноценный высококачественный датасет, предназначенный для сегментации объектов на проходимых поверхностях в реальных условиях городской и парковой среды. Он демонстрирует значительное преимущество в покрытии тонких и визуально сложных препятствий и может служить бенчмарком в задачах семантической, инстанс- и паноптической сегментации для мобильной робототехники.

4 Метод Talk2SAM

4.1 Мотивация

Сегментация объектов со сложной геометрией (например, тонких протяжённых структур вроде проводов или сеток, а также перекрывающихся и визуально схожих объектов) представляет существенную проблему для современных сегментационных моделей. В частности, SAM при всей своей универсальности заметно теряет качество на объектах с тонкими границами и мелкими деталями. Даже усовершенствованная версия SAM-HQ, предлагающая более точные границы масок, остаётся ограниченной типом вводимых подсказок: точек, рамок или грубых масок. Такие чисто геометрические подсказки часто недостаточны, чтобы однозначно выделить цельный объект в сценах с несколькими похожими или пересекающимися объектами. Например, одна ограничивающая рамка, охватывающая две переплетённые структуры (две провода или части велосипеда), не даёт явного сигнала, *какой* из объектов нужно сегментировать. В результате существующие методы либо пропускают тонкие части объектов, либо объединяют в одну маску несколько близко расположенных объектов. Это мотивирует включение семантической информации более высокого уровня в процесс сегментации. В методе **Talk2SAM** предлагается решать проблему тонких и неоднозначных объектов с помощью **текстовых подсказок**, которые несут описание целевого объекта и тем самым дополняют визуальные признаки сегментации недостающим контекстом.

4.2 Общая архитектура метода Talk2SAM

Модель **Talk2SAM** интегрирует семантические текстовые подсказки в контур интерактивной сегментации, позволяя уточнять маски объектов даже в сложных или неоднозначных сценах. В отличие от стандартных подходов SAM, использующих только пространственные подсказки (например, точки или рамки), Talk2SAM обогащает процесс сегментации высокоуровневым языковым описанием объекта. Это особенно полезно в ситуациях, когда требуется различать визуально схожие или перекрывающиеся объекты.

Общая концепция подхода показана на рис. 8. Модель принимает на вход изображение и текстовый запрос, соответствующий целевому объекту, например, `"Person"` или `"Motorbike"`. Также может быть задана пространственная подсказка — рамка

(box prompt), ограничивающая область интереса. Далее оба типа информации — визуальная и текстовая — обрабатываются параллельно и объединяются в едином пространстве признаков на этапе семантического сопоставления.

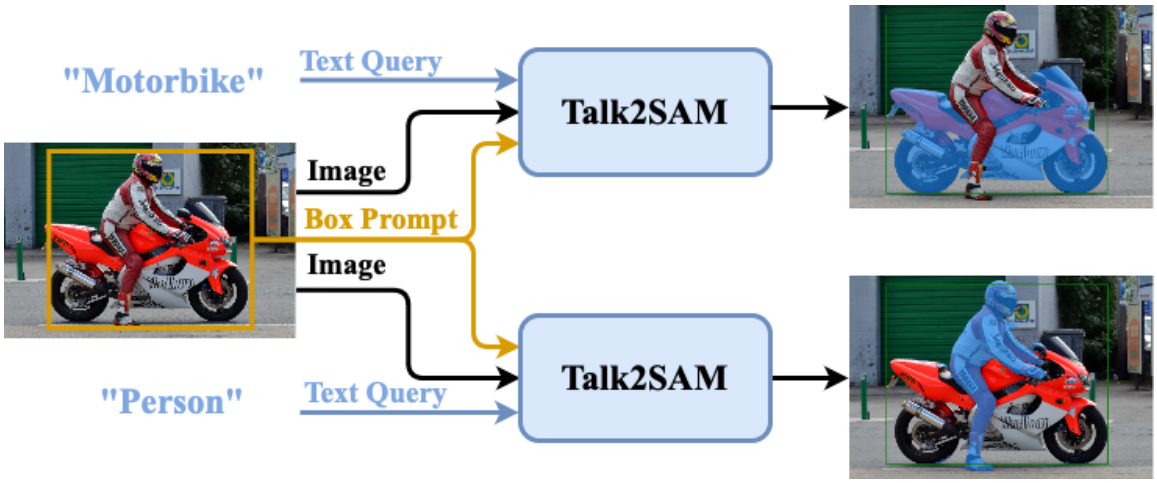

Рис. 8: Общая схема метода Talk2SAM. Модель получает изображение, текстовый запрос и опциональную box-подсказку. В зависимости от текста модель сегментирует разные объекты в пределах одной области изображения. Такой механизм позволяет точно различать перекрывающиеся или визуально схожие объекты за счёт семантической интерпретации запроса.

На рисунке изображены два последовательных вызова модели: для текстового запроса `"Motorbike"` и для `"Person"`. При этом изображение и box prompt остаются одинаковыми, однако различие в семантическом описании приводит к тому, что модель выделяет разные объекты в пределах одной и той же области. Таким образом, Talk2SAM реализует *семантически управляемую* сегментацию, позволяя пользователю явно указывать, какой именно объект следует сегментировать, даже если они визуально перекрываются. Такая способность особенно важна для задач сегментации в сложных сценах, где традиционные пространственные подсказки не позволяют однозначно определить объект интереса.

На рис. 9 представлена детальная архитектура метода Talk2SAM, показывающая все ключевые компоненты модели и схему их взаимодействия. Изображение, текст и опциональная рамка-промпт проходят отдельные этапы обработки, прежде чем быть объединёнными в финальном сегментационном блоке.

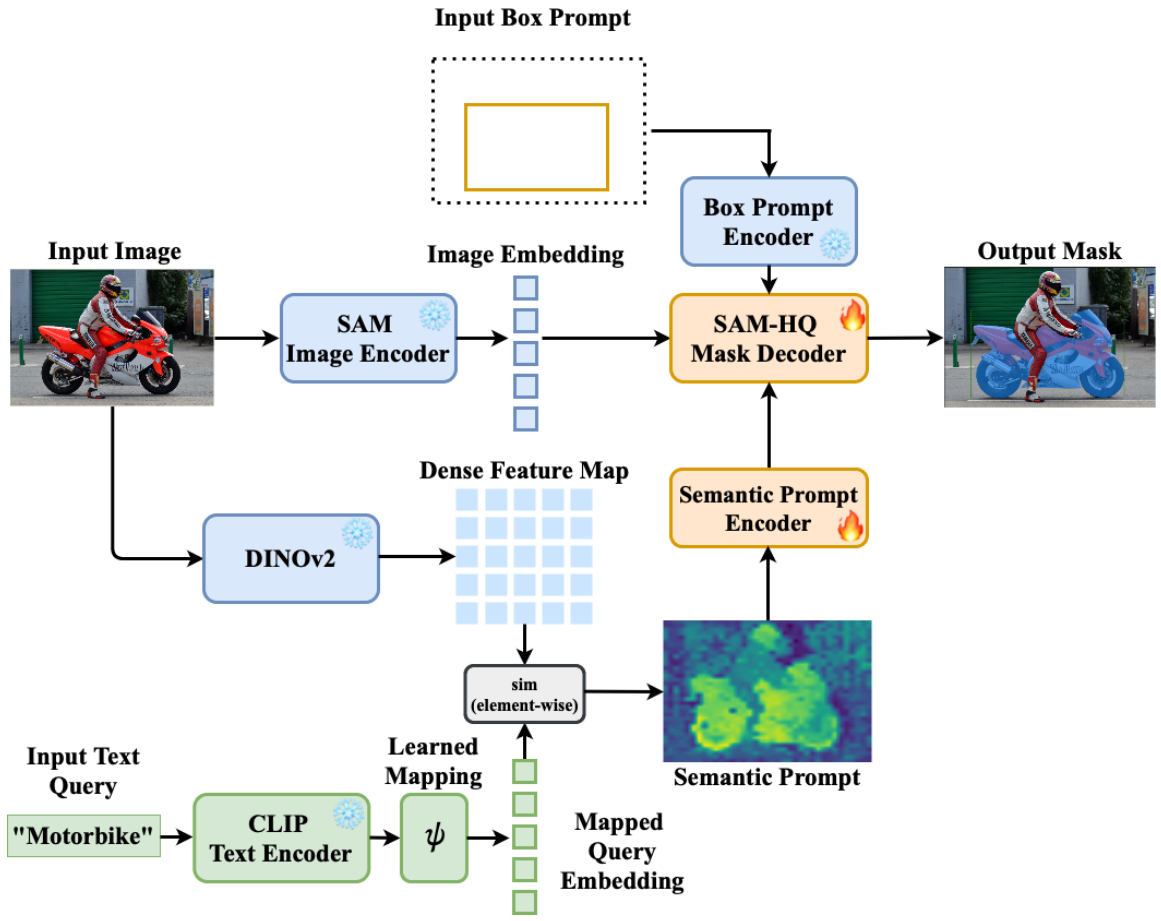

Рис. 9: Детализированная архитектура метода Talk2SAM. Модель включает два потока обработки: визуальный (через SAM и DINOv2) и текстовый (через CLIP), которые объединяются на этапе построения карты релевантности. Эта карта затем кодируется и подаётся в декодер SAM-HQ вместе с признаками изображения и геометрическими подсказками. Таким образом достигается точное выделение объектов, соответствующих текстовому описанию.

4.3 Построение семантической карты релевантности

Основой метода является вычисление **семантической карты релевантности**, показывающей, какие области изображения соответствуют заданному текстовому запросу. Этот этап осуществляется следующим образом.

Шаг 1: Текстовое кодирование. Пользователь задаёт текстовую категорию или краткое описание целевого объекта t (без развернутого предложения, достаточно

одного слова или фразы, например «провод»). Этот текст поступает в текстовый энкодер модели CLIP, который преобразует запрос в векторное представление $t \in \mathbb{R}^{D_t}$ фиксированной размерности. Полученный семантический эмбеддинг содержит высокоуровневую информацию о признаках объекта, описанного в тексте.

Шаг 2: Проекция в пространство DINOv2. Прямое сравнение CLIP-эмбеддинга с признаками изображения затруднительно из-за различий в пространстве представлений. Поэтому в Talk2SAM применяется обучаемое нелинейное отображение $\psi(\cdot)$, которое **приводит текстовый эмбеддинг к пространству визуальных признаков DINOv2**. Это отображение состоит из двух аффинных слоёв с промежуточной активацией tanh (по аналогии с подходом Talk2DINO) и определяется формулой:

$$\psi(t) \;=\; W_b^T \,\tanh\!\big(W_a^T\, t + b_a\big) \;+\; b_b\,, \tag{4.1}$$

где W_a, W_b — обучаемые матрицы преобразования, а b_a, b_b — соответствующие смещения. В результате применения $\psi(t)$ исходный текстовый признак отображается в вектор $\psi(t) \in \mathbb{R}^{D_v}$, лежащий в том же призначном пространстве D_v, что и визуальные признаки DINOv2. Такая «подгонка» обеспечивает лучшее соответствие между текстовыми и визуальными представлениями одной сцены.

Шаг 3: Выделение регионов в визуальных признаках. Одновременно с текстовым потоком, изображение обрабатывается зрительным энкодером DINOv2 (ViT-L/14), который известен способностью сегментировать изображение на семантически осмысленные части с помощью механизмов самовнимания. В последнем слое трансформера DINOv2 формируется N карт внимания $A_i \in \mathbb{R}^{(H/P)\times(W/P)}$, где N — число голов самовнимания, H и W — размеры признакового тензора изображения, а P — размер патча (в нашем случае $P = 16$). Каждая карта A_i соответствует одной голове и выделяет определённый семантический регион изображения (напрямую интерпретируется как маска внимания, указывающая, на какие патчи обращает внимание данная голова). На основе карты A_i вычисляется **вектор признаков региона** $v_{A_i} \in \mathbb{R}^{D_v}$ как взвешенное среднее всех признаков патчей $v[h, w]$ с весами из матрицы внимания:

$$v_{A_i} \;=\; \sum_{h,w} \mathrm{softmax}(A_i)[h, w] \cdot v[h, w]\,, \tag{4.2}$$

где $v[h, w] \in \mathbb{R}^{D_v}$ — признак (выход последнего слоя энкодера DINOv2) патча, расположенного в позиции (h, w) сетки патчей, а $\mathrm{softmax}(A_i)[h, w]$ — нормализованное значение внимания данной головы для этого патча. Таким образом, для каждой го-

ловы самовнимания получаем свой вектор, представляющий определённую область изображения.

Шаг 4: Семантическое соответствие текста и регионов. Далее необходимо определить, какой из полученных региональных признаков наиболее соответствует текстовому запросу пользователя. Для этого вычисляется **косинусное сходство** между каждым вектором региона v_{A_i} и проецированным текстовым вектором $\psi(t)$:

$$\mathrm{sim}\big(v_{A_i},\, t\big) \;=\; \frac{v_{A_i} \cdot \psi(t)^T}{\|v_{A_i}\|\,\|\psi(t)\|}\,, \tag{4.3}$$

где точка обозначает скалярное произведение. Метрика sim принимает значения от -1 до 1 и показывает степень семантического соответствия регионального признака содержанию текстового описания. Затем из всех N значений сходства выбирается максимум:

$$s(t) \;=\; \max_{i=1,\ldots,N} \mathrm{sim}\big(v_{A_i},\, t\big)\,. \tag{4.4}$$

Пусть индекс j соответствует голове внимания, давшей наибольшее сходство $s(t) = \mathrm{sim}(v_{A_j}, t)$. Это означает, что j-я голова выделила область изображения, наиболее релевантную заданному тексту.

Шаг 5: Формирование карты релевантности. Используя найденное соответствие, Talk2SAM генерирует плотную карту релевантности размером $(H/P) \times (W/P)$, указывающую, какие части изображения относятся к искомому объекту. Естественным выбором является взять карту внимания A_j для головы j с максимальным ответом и трактовать её как семантическую карту: элементы $\mathrm{softmax}(A_j)[h, w]$ служат мерами принадлежности патча (h, w) к объекту, описанному текстом. Эта грубая карта (разрешением порядка 37×37 для ViT-L/14 на входном изображении 592×592) затем **билинейно масштабируется** до размера исходного изображения (например, 256×256 признаков, сопоставимых по размеру с выходным разрешением SAM-HQ). В результате получается **семантическая карта релевантности** $M_t(x, y)$, где для каждого положения (x, y) отражается степень соответствия пикселя объекту, указанному в тексте. Данная карта выступает в роли дополнительного входного признакового слоя, задающего априорное семантическое выделение нужной области на изображении.

4.4 Интеграция семантической карты в декодер SAM-HQ

На завершающем этапе сформированная семантическая карта используется для улучшения сегментации в модели SAM-HQ. В оригинальной архитектуре SAM-HQ предусмотрен так называемый *Mask Prompt Encoder* — легкий сверточный модуль, способный принимать бинарную маску в качестве подсказки и преобразовывать её в эмбеддинг для дальнейшего использования декодером. Однако в стандартном SAM-HQ этот модуль имеет фиксированные (необучаемые) веса и рассчитан на простой двоичный вход (например, грубая маска-подсказка или результат предыдущей итерации сегментации). В Talk2SAM предлагается **заменить** или перепрофилировать этот модуль под новую задачу: обработку непрерывной семантической карты M_t.

Конкретно, семантическая карта релевантности сначала нормализуется и подаётся в **семантический энкодер подсказки** (Semantic Prompt Encoder) — небольшой сверточный нейросетевой блок, архитектурно аналогичный mask encoder в SAM-HQ, но **полностью обучаемый** на наших данных. Этот энкодер извлекает из карты компактное представление (эмбеддинг) e_M, которое содержательно отражает форму и положение целевого объекта, заданного текстом. Далее эмбеддинг e_M поступает в декодер SAM-HQ **наравне с другими подсказками**, такими как эмбеддинги точек, рамок или исходного изображения. Таким образом, декодер SAM-HQ получает дополнительный сигнал и выполняет **совместное рассуждение** на основе двух видов информации: геометрической (от традиционных промптов, ограничивающих область поиска) и семантической (от текстово-наведённой карты релевантности). В результате сегментация становится чувствительной к содержательному описанию объекта: модель вычленяет именно тот объект, который соответствует запросу, даже если рядом находятся другие похожие объекты или тонкие пересекающиеся детали.

В отличие от исходного бинарного mask encoder, который не изменялся в процессе обучения SAM-HQ, наш семантический энкодер подсказки обучается совместно с остальными новыми компонентами Talk2SAM. Это позволяет ему адаптироваться к особенностям непрерывных карт релевантности и эффективно кодировать их в пригодный для декодера формат. По сути, вместо фиксированного усреднения маски в embedding (как было в SAM-HQ), модель учится выделять из семантической карты именно ту информацию, которая улучшает сегментацию целевого объекта.

4.5 Обучение модели

При обучении Talk2SAM основная часть крупных модулей сохраняется неизменной, а обучаются лишь добавленные слои и адаптеры. Такой подход снижает требуемое число параметров для оптимизации и уменьшает риск деградации уже обученных компонентов. В частности, **замороженными (необучаемыми)** остаются параметры зрительного энкодера DINOv2, а также backbone-энкодер исходного SAM-HQ (Vision Transformer, извлекающий признаки изображения и геометрических промптов). Это означает, что возможности выделения признаков изображения, накопленные в этих больших моделях, используются «как есть». Что касается модели CLIP, стратегия зависит от объёма доступных данных: для небольших датасетов сохраняют замороженным и её, тогда как при наличии большого корпуса можно дообучить только финальный слой текстового энкодера CLIP, чтобы слегка подстроить семантическое пространство под предметную область.

Обучаемыми компонентами Talk2SAM являются:

- Нелинейное отображение $\psi(\cdot)$ для проекции текстовых эмбеддингов в пространство DINOv2 (параметры W_a, b_a, W_b, b_b в формуле 4.1).
- Семантический энкодер подсказки (сверточная сеть, кодирующая карту релевантности M_t в эмбеддинг e_M).
- Декодер SAM-HQ, генерирующий маску высокого разрешения на основе объединенных визуально-геометрических и семантических признаков.
- (Опционально) Проекционный слой текстового энкодера CLIP – только в тех экспериментах, где дообучение языка допускается.

Обучение производят методом стохастического градиентного спуска (SGD). В экспериментах использовалась скорость обучения $\eta = 10^{-3}$, батч размером 4 и продолжительность порядка 30 эпох на одном GPU (NVIDIA Tesla V100). Для функции потерь взята **стандартная сумма SAM-HQ потерь**: бинарная кросс-энтропия (BCE) совместно с Dice-лоссом, оценивающие расхождение предсказанной маски с истинной по пиксельному перекрытию и по граничной точности. Эти компоненты ошибки сбалансированы так же, как в оригинальной работе по SAM-HQ , для одновременной оптимизации точности всего объекта и его границ.

Важно подчеркнуть, что добавление текстового контура практически не усложнило модель в плане числа параметров. Совокупно обучаемые параметры составляют

менее 1% от всех параметров объединённой системы (основная часть весов принадлежит замороженным CLIP, DINOv2 и SAM-HQ). Благодаря этому обучение Talk2SAM остаётся достаточно эффективным по памяти и времени: например, при батче 1 максимальное потребление памяти составляет около 10.7 ГБ VRAM, что лишь незначительно больше, чем 10.1 ГБ у исходного SAM-HQ. Данные факты подчёркивают, что предложенная интеграция текста не требует существенных ресурсов и может быть реализована на базе существующих моделей сегментации без их полного переобучения.

Сравнение числа обучаемых параметров в Talk2SAM и других конфигурациях представлено в таблице 4. Как видно, даже в случае частичного дообучения CLIP, объём оптимизируемых параметров остаётся ниже 1% от общего числа весов, что делает Talk2SAM удобной для обучения в условиях ограниченных ресурсов.

Таблица 4: Сравнение числа обучаемых и общих параметров моделей

Конфигурация модели	**Trainable**	**Total**	**Доля обучаемых**
SAM-HQ	5.65M	641M	0.88%
Talk2SAM (CLIP заморожен)	7.22M	1096M	0.66%
Talk2SAM (CLIP частично обучается)	10.64M	1096M	0.97%

4.6 Особенности и преимущества метода

Предложенный подход Talk2SAM привносит несколько важных преимуществ по сравнению с базовыми методами сегментации:

- **Текстовое управление сегментацией:** Пользователь может явно указать, *что именно* нужно сегментировать, словами описав целевой объект. Это позволяет системе ориентироваться на семантику объекта, а не только на его положение. В результате достигается *контролируемая сегментация*: в одной и той же области изображения можно по разным текстовым запросам получить разные маски (например, выделить либо “тонкую проволоку”, либо “толстый кабель” внутри одного участка сцены).
- **Разрешение неоднозначностей и перекрытий:** Talk2SAM успешно разделяет пересекающиеся или близко расположенные объекты, которые визуально похожи. Языковая подсказка помогает отличить их друг от друга по смыс-

ловым признакам. В традиционном SAM подобная ситуация могла приводить к слиянию объектов в одну маску или пропуску нужного объекта, тогда как здесь текст (например, «передний объект» vs «задний объект», или название конкретного предмета) служит дополнительным признаком для различения.

- **Улучшенная сегментация тонких структур:** Для объектов со сложной формой — тонких, извилистых, с мелкими деталями — добавление семантической карты позволяет модели фокусироваться на этих сложных областях. Даже если визуальные границы размыты или тонки, текстовое описание (например, «провод») направляет модель выделять соответствующую непрерывную структуру. В итоге границы масок получаются более полными и точными, особенно в метрике границ (boundary IoU).
- **Совместимость и минимальные вычислительные затраты:** Talk2SAM не требует тренировки модели «с нуля» и встраивается поверх уже существующего сегментатора. Благодаря тому, что изменяется лишь малое число параметров ($<1\%$), метод легко применять к разным вариантам SAM без существенных затрат. Архитектурных модификаций самого ядра SAM не требуется — текстовая подсказка добавляется модульно через дополнительный энкодер и не нарушает работу исходного декодера. Это значит, что *совместимость* с исходной моделью полная: при отсутствии текстового ввода система работает как обычный SAM-HQ, а при наличии текста — обогащает его результаты.

Таким образом, Talk2SAM демонстрирует, что интеграция естественного языка предоставляет эффективный механизм для точной сегментации сложных объектов. Метод сохраняет все сильные стороны SAM-HQ (универсальность, высокое разрешение масок) и дополняет их новым уровнем семантического понимания сцены. В контексте сегментации тонких и трудно различимых структур это позволяет достичь качества, недостижимого для одних лишь геометрических промптов.

5 Вычислительные эксперименты

В данном разделе представлены количественные и качественные результаты работы модели **Talk2SAM**, а также её сравнение с базовыми сегментаторами **SAM** и **SAM-HQ**. Все эксперименты проводились на фиксированных тестовых поднаборах четырёх датасетов: **ThinObject5K**, **DIS5K**, **BIG**, **TAOMR**. Для всех моделей использовались одинаковые входные изображения и текстовые запросы, извлечённые из имён файлов или аннотаций. В качестве визуального энкодера применялись различные версии Vision Transformer: **ViT-B**, **ViT-L** и **ViT-H**.

Важно отметить, что стратегия дообучения проекционного слоя текстового энкодера CLIP зависела от объёма обучающего корпуса. В экспериментах на сравнительно небольших наборах **BIG** и **TAOMR** данный слой оставался замороженным, чтобы избежать переобучения на ограниченных данных. В остальных случаях (на ThinObject5K и DIS5K) проекционный слой CLIP размораживался, что позволяло модели адаптировать семантическое пространство под конкретную предметную область.

Методика оценки

Качество сегментации оценивалось с использованием двух метрик:

- **mIoU (mean Intersection over Union)** — средняя доля пересечения предсказанной и истинной масок по отношению к их объединению.
- **mBIoU (mean Boundary IoU)** — модифицированная метрика, чувствительная к точности границ объектов.

5.1 Результаты на датасете ThinObject5K

Таблица 5: Сравнение качества сегментации на наборе изображений с тонкими структурами (ThinObject5K). Talk2SAM превосходит все базовые модели по всем метрикам.

Метод	mIoU	mBIoU	Время (мс)
SAM ViT-H	0.686	0.591	557
SAM-HQ ViT-H	0.870	0.759	560
Talk2SAM ViT-H	**0.929**	**0.842**	720
SAM-HQ ViT-B	0.811	0.695	142
Talk2SAM ViT-B	**0.921**	**0.830**	302
SAM-HQ ViT-L	0.853	0.737	329
Talk2SAM ViT-L	**0.926**	**0.839**	489

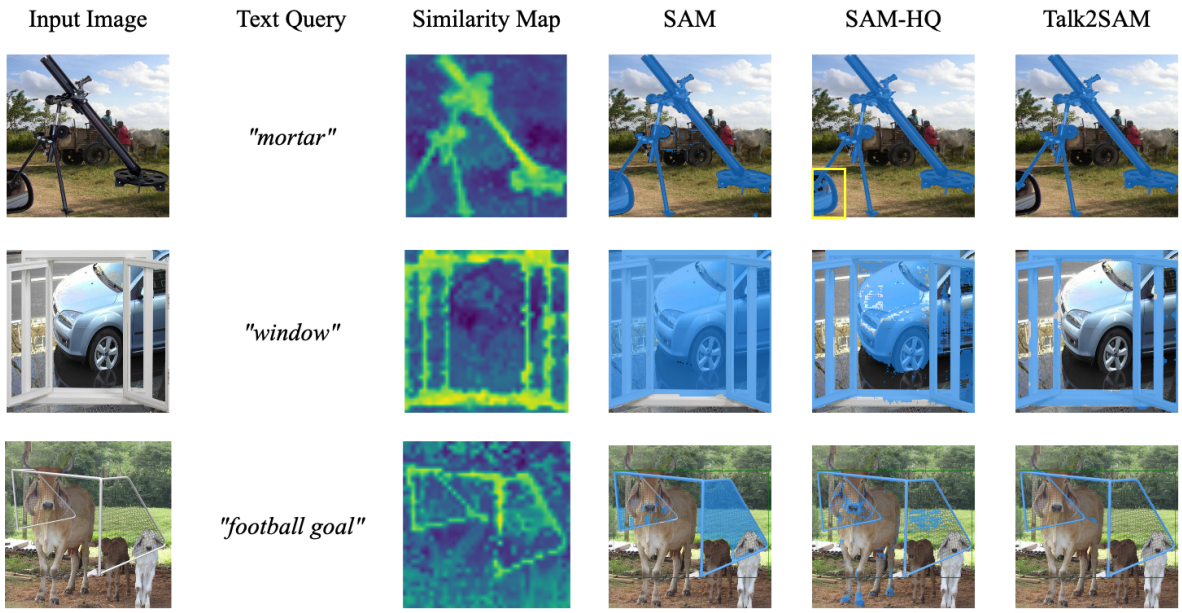

Рис. 10: Качественные результаты на изображениях с тонкими и вытянутыми структурами. В каждом ряду представлены: входное изображение, текстовый запрос, карта релевантности, маски SAM, SAM-HQ и Talk2SAM. Talk2SAM показывает наилучшее сохранение непрерывности и формы объекта.

На рисунке 10 представлены качественные результаты на изображениях с тонкими и вытянутыми структурами. В таких случаях даже небольшие ошибки в сегментации визуально сильно искажают результат. Модель Talk2SAM стабильно захваты-

вает полную структуру деликатных объектов — таких как металлические каркасы, окна и направляющие конструкции — в то время как SAM и SAM-HQ часто не сохраняют контекст или целостность формы. Это подчёркивает важность семантического приоритета, обеспечиваемого текстовым описанием, для точной и контекстно-информированной сегментации.

5.2 Результаты на датасете BIG

Таблица 6: Результаты сегментации на наборе BIG при использовании ViT-H. Talk2SAM демонстрирует уверенное улучшение по обеим метрикам.

Метод	**mIoU**	**mBIoU**
SAM	0.902	0.784
SAM-HQ	0.929	0.824
Talk2SAM	**0.947**	**0.866**

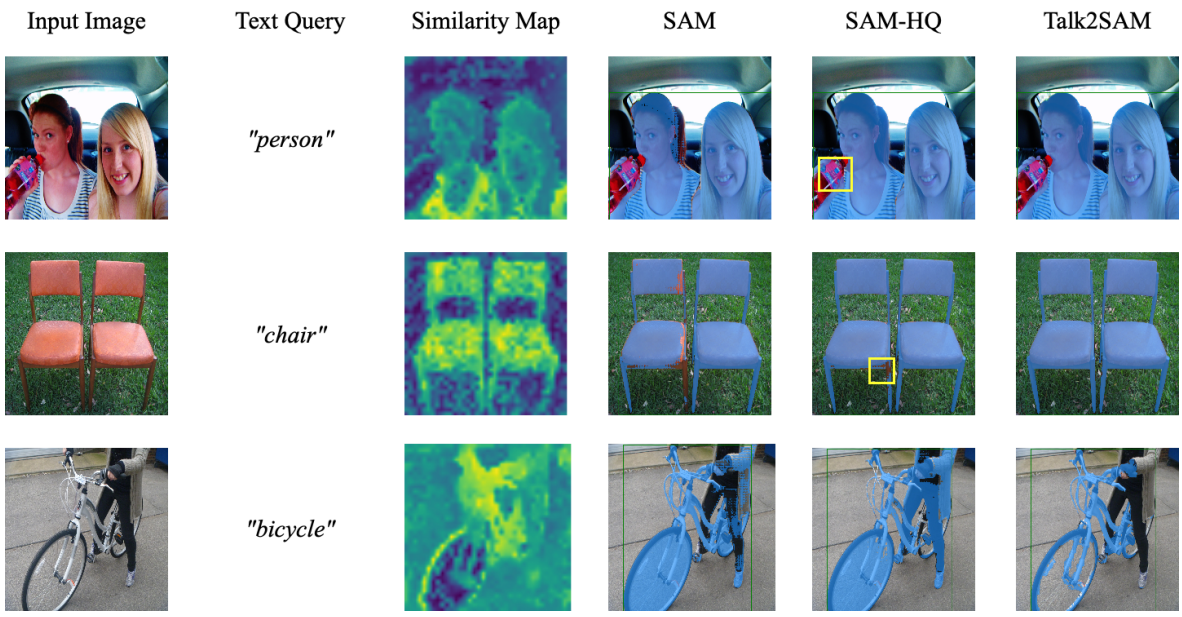

Рис. 11: Качественные результаты на наборе BIG. Каждый ряд содержит: входное изображение, текстовый запрос, карту релевантности (CLIP-DINOv2), а также предсказанные маски от моделей SAM, SAM-HQ и Talk2SAM. В то время как SAM и SAM-HQ нередко выдают неполные или неточные маски, Talk2SAM точно сегментирует нужные объекты, используя текстовое описание.

На рисунке 11 представлены примеры работы моделей на наборе изображений BIG. Для каждой сцены показаны результаты сегментации по текстовому запросу (*“person”*, *“chair”*, *“bicycle”*). Модель SAM часто демонстрирует неточную локализацию или частичное выделение объектов. SAM-HQ, в свою очередь, улучшает границы масок, но всё ещё испытывает сложности при наличии контекстных неоднозначностей, например, если рядом расположены несколько схожих объектов. Talk2SAM успешно преодолевает эти ограничения за счёт интеграции семантики текстового запроса: модель выделяет именно тот объект, который указан в описании. Особенно это заметно на примере с велосипедом, где Talk2SAM корректно сегментирует раму и колёса, отделяя их от человека и фона. Подобные результаты подчёркивают значимость языковой подсказки в сложных сценах с высокой насыщенностью.

5.3 Результаты на DIS5K

Рисунок 12 иллюстрирует работу моделей на визуально насыщенных и детализированных сценах из DIS5K. Talk2SAM стабильно обеспечивает более точную локализацию объектов благодаря использованию семантической информации, закодированной в текстовом запросе. Даже при наличии мелких или визуально схожих объектов модель демонстрирует способность удерживать границы и структуру сегментируемых областей. В то время как SAM и SAM-HQ часто теряют мелкие элементы или не могут корректно отделить перекрывающиеся части, Talk2SAM демонстрирует устойчивость за счёт дополнительной релевантностной карты, направляющей внимание декодера в нужные регионы изображения.

Таблица 7: Результаты на датасете DIS5K при использовании ViT-H. Talk2SAM показывает наилучшие значения по mIoU и mBIoU, несмотря на сложность сцен.

Метод	**mIoU**	**mBIoU**
SAM	0.573	0.497
SAM-HQ	0.723	0.660
Talk2SAM	**0.749**	**0.663**

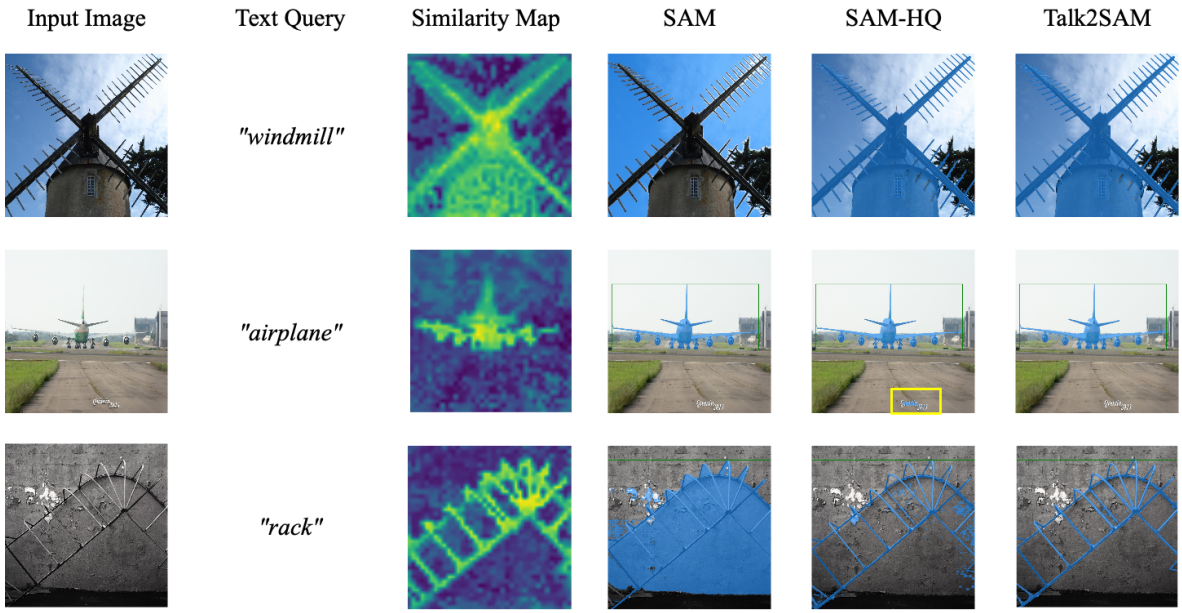

Рис. 12: Качественные результаты на наборе DIS5K. Каждый ряд содержит: исходное изображение, текстовый запрос, карту релевантности и маски, предсказанные моделями SAM, SAM-HQ и Talk2SAM. Talk2SAM сохраняет высокую точность сегментации даже в условиях сложных текстур и большого количества мелких элементов.

5.4 Результаты на датасете TAOMR

Особенностью датасета TAOMR является его направленность на **инстанс-сегментацию**, при которой необходимо выделять индивидуальные объекты даже в случае множественного присутствия экземпляров одного класса. Это предъявляет дополнительные требования к модели Talk2SAM, поскольку её релевантностная карта по умолчанию акцентирует *все* области изображения, соответствующие текстовому запросу. Такая особенность может приводить к ошибкам сегментации: например, выделению объектов, находящихся вне рамки (box prompt), что недопустимо в условиях инстанс-разметки.

Для преодоления этой проблемы в рамках Talk2SAM была реализована стратегия пространственного приоритезации: значения на релевантностной карте $M_t(x, y)$ внутри прямоугольной подсказки усиливаются в 2 раза, а за её пределами ослабляются в 10 раз. Такой подход позволяет **смягчить влияние фоновых объектов**, сохранив внимание модели внутри заданной области интереса и тем самым повысив точность инстанс-сегментации.

В таблице 8 представлены численные результаты на трёх категориях сложной формы: `wire`, `hose` и `firehose`. Видно, что Talk2SAM превосходит SAM-HQ по обе-

им метрикам, несмотря на сложности, связанные с вытянутой геометрией объектов и ограничениями рамочного контекста. Рисунок 13 визуализирует различия между подходами. Talk2SAM демонстрирует более целостное и корректное покрытие объектов даже при высоком уровне визуального шума и пересечении с другими структурами.

Таблица 8: Результаты на подмножестве TAOMR (ViT-H), категории сложной формы

Метод	**mIoU**	**mBIoU**
SAM-HQ	0.743	0.743
Talk2SAM	**0.772**	**0.771**

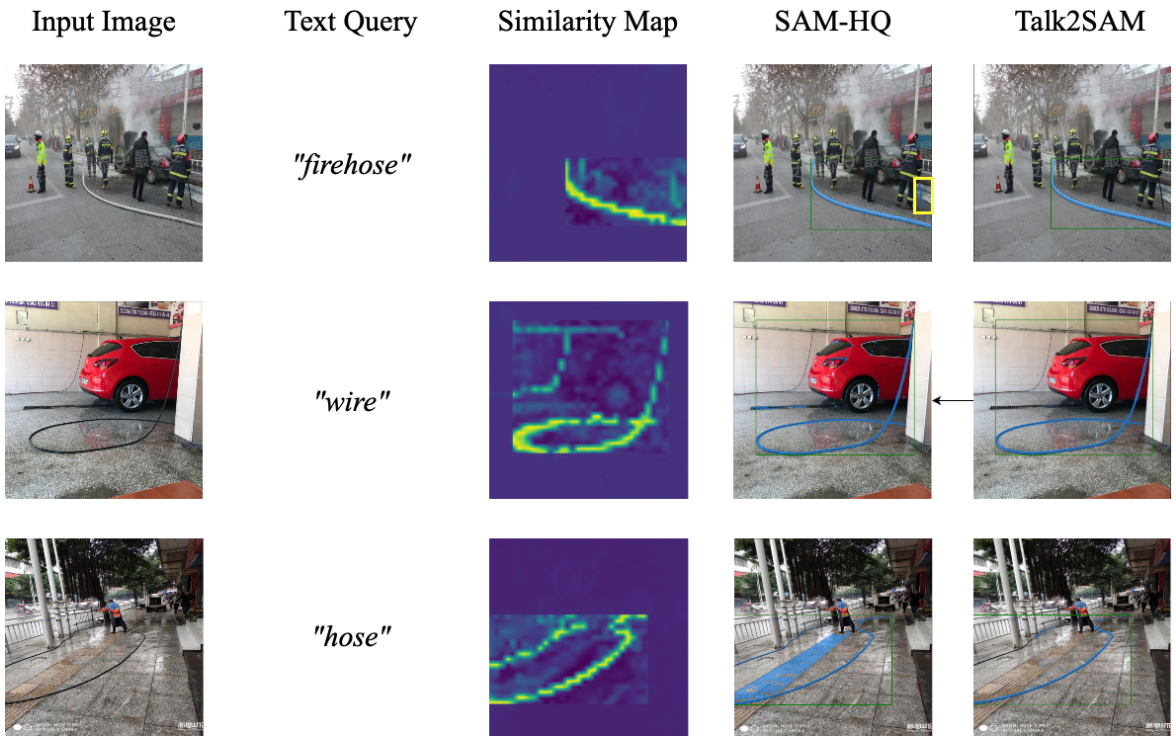

Рис. 13: Результаты сегментации на TAOMR. Слева направо: входное изображение, текстовый запрос, карта релевантности, предсказания SAM-HQ и Talk2SAM. Видно, как Talk2SAM выделяет целевые структуры более точно, несмотря на сложность форм.

5.5 Разбор ошибок и ограничений

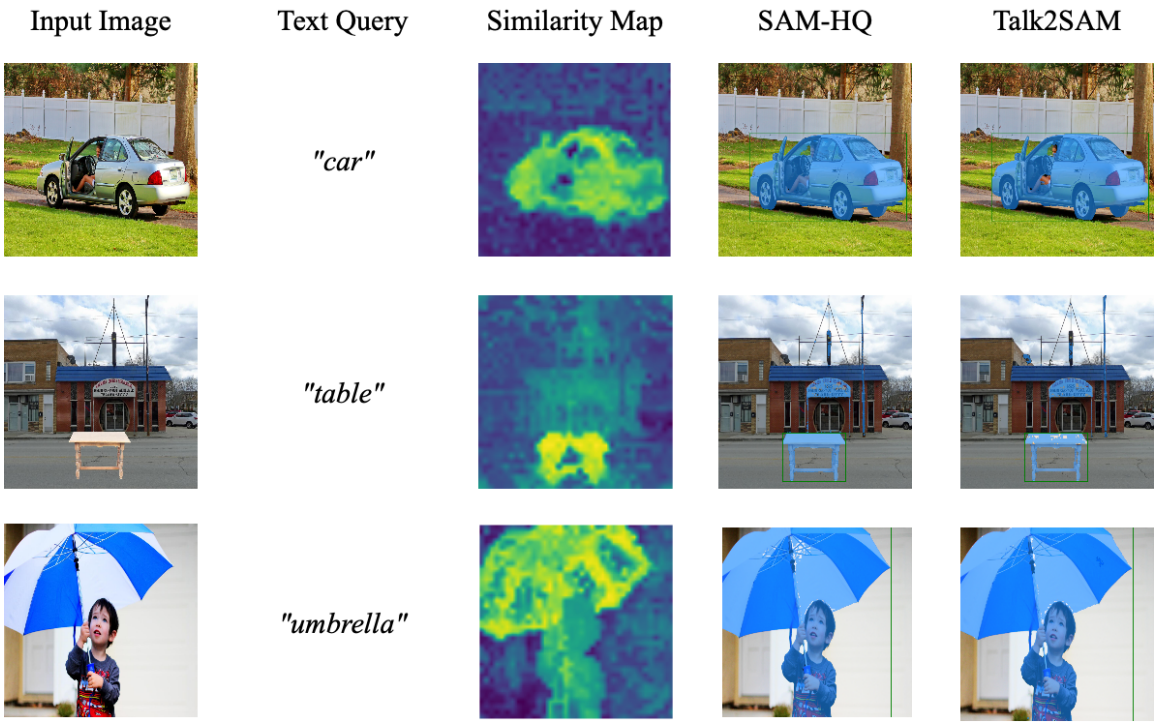

Рис. 14: Сложные случаи, где релевантностные карты соответствуют тексту, но итоговая сегментация оказывается некорректной. Каждая строка содержит: входное изображение, текстовый запрос, карту релевантности CLIP-DINOv2, и результаты SAM-HQ и Talk2SAM. Примеры демонстрируют, что одной только семантической локализации недостаточно — SAM-HQ может ошибаться при отклонении от привычных визуальных шаблонов. Talk2SAM частично компенсирует это за счёт использования текстовой семантики.

На рисунке 14 приведены примеры, в которых карта релевантности, построенная на основе признаков CLIP-DINOv2, корректно отражает положение объекта, соответствующего текстовому запросу (например, силуэт автомобиля или форму зонта). Однако финальная сегментация, особенно у SAM-HQ, остаётся далёкой от идеальной. Это указывает на архитектурное ограничение: даже при наличии точного семантического сигнала, декодер SAM-HQ не способен интерпретировать его без соответствующей адаптации.

Видно, что SAM-HQ унаследовал ряд слабых сторон базовой модели SAM — особенно в случае частично перекрытых или нечасто встречающихся объектов. Поскольку модель не имеет прямого доступа к семантическому содержанию релевантной карты, её refinement-механизмы оказываются недостаточными для надёжной идентификации целевого объекта.

Talk2SAM, в отличие от SAM-HQ, использует текстовые приоритеты, позволяя декодеру учитывать не только пространственные, но и семантические признаки. Это позволяет получить более согласованные и контекстно релевантные маски, даже при наличии неоднозначности, визуального шума или фоновых объектов схожей формы.

6 Заключение

В данной работе была предложена и подробно проанализирована система Talk2SAM — архитектура для интерактивной сегментации, основанная на интеграции языковых подсказок в мощную визуальную модель сегментации SAM-HQ. Основная идея заключалась в использовании семантической информации из текстовых запросов для генерации плотной карты релевантности, отражающей соответствие между текстовым описанием и визуальными признаками изображения. Эта карта затем использовалась как входной сигнал в pipeline сегментации, замещая или дополняя традиционные геометрические подсказки (такие как точки или bounding box).

Разработка Talk2SAM была мотивирована необходимостью сегментации сложных, тонких, неоднородных или частично перекрытых объектов, которые часто остаются вне зоны внимания классических сегментаторов. Особенно это важно в контексте робототехнических приложений и визуальных систем, где важна точная локализация потенциальных препятствий на изображении по семантическому описанию.

Основные результаты

Работа включает в себя следующие ключевые достижения:

- Предложена архитектура Talk2SAM, эффективно сочетающая энкодеры CLIP и DINOv2 для извлечения семантической карты релевантности между изображением и текстом.
- Разработан механизм внедрения карты релевантности в pipeline модели SAM-HQ через адаптированный Semantic Prompt Encoder, обучаемый отдельно от визуального энкодера.
- Обоснована возможность дообучения только небольшого числа параметров модели (менее 1%), при этом достигается значительный прирост точности по сравнению с SAM-HQ.

- Проведены всесторонние эксперименты на четырех независимых и разнообразных датасетах: ThinObject5K, DIS5K, BIG и TAOMR. Во всех случаях Talk2SAM показал лучшие результаты как по метрике mIoU, так и особенно по mBIoU, демонстрируя устойчивость к деталям и улучшение границ объектов.

- Подтверждена эффективность подхода в условиях ограниченного количества обучаемых параметров и умеренного объёма вычислений (на уровне SAM-HQ), что делает метод применимым в реальных условиях.

Визуальные эксперименты подтвердили способность Talk2SAM точно сегментировать тонкие и вытянутые объекты (например, провода, заборы, рамки), избегая типичных ошибок SAM, таких как захват фоновых областей, замыкание маски или прерывание границ. Особенно заметно это преимущество в случаях с множественными похожими объектами в сцене, где SAM и SAM-HQ часто путаются или не способны локализовать все объекты, соответствующие запросу.

Также были проанализированы неудачные случаи, показавшие, что даже при высококачественной карте релевантности сегментационный модуль может допускать ошибки. Основные причины заключаются в ограничениях архитектуры SAM-HQ: отсутствие семантической модальности и сложности с обработкой плотных карт в исходном prompt encoder. Дополнительно влияет низкое разрешение карты релевантности и ограничения CLIP в обработке редких или полисемантических терминов.

Вклад в разработку датасета TAOMR

Наряду с архитектурой Talk2SAM значительный вклад данной работы состоит в создании и систематизации нового специализированного набора данных **TAOMR** (Traversable Area Objects for Mobile Robotics). Данный датасет нацелен на сегментацию визуально сложных и опасных объектов в реальных уличных сценах, которые редко охватываются существующими публичными наборами, такими как COCO или Cityscapes. Объекты, такие как провода, шланги, дорожные неровности, лужи и мусор, представляют собой угрозу для мобильных роботов, но плохо распознаются существующими сегментационными системами.

TAOMR включает более 8800 изображений и более 75 тысяч аннотированных масок по 12 категориям. Изображения собраны из различных источников: веб-поиска по

картинкам, кадров с реальных мобильных платформ и сторонних датасетов. Для создания аннотаций применялась оригинальная полуавтоматическая разметка, включающая использование моделей Grounding-DINO и SAM-HQ с последующей ручной фильтрацией и уточнением в интерфейсе CVAT. Все маски приведены к формату COCO и снабжены текстовыми метками категорий.

Особое внимание было уделено созданию бенчмаркового подмножества **TAOMR-Bench**, содержащего сбалансированное распределение масок по категориям и предназначенного для объективного сравнения методов. Этот бенчмарк может быть использован как эталон для оценки моделей сегментации в условиях открытого мира и ограниченного количества визуальных подсказок.

Таким образом, TAOMR не только расширяет возможности исследований в области open-vocabulary сегментации и безопасной навигации, но и служит реальной платформой для оценки и совершенствования моделей, подобных Talk2SAM. В дальнейшем датасет может быть дополнен видеоаннотациями, атрибутивной информацией и интегрирован в симуляционные среды для более широкого тестирования и валидации методов сегментации в различных сценариях.

Направления дальнейших исследований

Работа открывает ряд перспективных направлений:

- **Обучение end-to-end**: хотя текущая архитектура разделена на модульную схему с фиксированными энкодерами, обучение всей модели совместно (включая DINOv2 и CLIP) может улучшить согласованность признаков и точность карты релевантности.
- **Иерархическая обработка карт релевантности**: внедрение multi-scale подхода может повысить точность сегментации объектов различного масштаба, особенно малых и слабо различимых на фоне.
- **Улучшение prompt encoder**: разработка специального энкодера подсказок, предназначенного для непрерывных плотных семантических карт, способного учитывать контекст и границы.
- **Интеграция обратной связи**: использование дополнительной информации от пользователя (например, уточняющих подсказок, уточняющих масок) может позволить итеративно улучшать сегментацию.

- **Применение в реальных сценариях**: перенос метода в задачи автономной навигации, промышленного контроля, медицинской визуализации, где требуется точная сегментация объектов по текстовому описанию.
- **Снижение зависимости от CLIP**: исследование альтернативных языково-визуальных моделей (например, BLIP [24], OWL-ViT [50]) может повысить устойчивость метода к редким и абстрактным понятиям.

Заключительные замечания

В целом, Talk2SAM демонстрирует потенциал методов открытой сегментации с управлением через язык. Работа показывает, что добавление семантической модальности в pipeline классических визуальных моделей приводит к ощутимому улучшению качества при незначительном увеличении сложности.

Предложенный подход не требует глобального переобучения, адаптируется к различным ViT-бэкбонам, и, что важно, позволяет удобно интегрировать семантические описания в существующие системы обработки изображений. Это делает Talk2SAM универсальным инструментом, применимым во множестве задач компьютерного зрения и человеко-машинного взаимодействия.

Список литературы

1. *Krähenbühl P.*, *Koltun V.* Efficient inference in fully connected crfs with gaussian edge potentials // Advances in neural information processing systems. — 2011. — Т. 24.
2. Saliency detection via graph-based manifold ranking / C. Yang, L. Zhang, H. Lu, X. Ruan, M.-H. Yang // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2013. — С. 3166—3173.
3. Global contrast based salient region detection / M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, S.-M. Hu // IEEE transactions on pattern analysis and machine intelligence. — 2014. — Т. 37, № 3. — С. 569—582.
4. Microsoft coco: Common objects in context / T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick // Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13. — Springer. 2014. — С. 740—755.
5. Hierarchical image saliency detection on extended CSSD / J. Shi, Q. Yan, L. Xu, J. Jia // IEEE transactions on pattern analysis and machine intelligence. — 2015. — Т. 38, № 4. — С. 717—729.
6. The cityscapes dataset for semantic urban scene understanding / M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2016. — С. 3213—3223.
7. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation / G. Lin, A. Milan, C. Shen, I. Reid // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2017. — С. 1925—1934.
8. Scene parsing through ade20k dataset / B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, A. Torralba // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2017. — С. 633—641.
9. Icnet for real-time semantic segmentation on high-resolution images / H. Zhao, X. Qi, X. Shen, J. Shi, J. Jia // Proceedings of the European conference on computer vision (ECCV). — 2018. — С. 405—420.

10. *Dias P. A.*, *Medeiros H.* Semantic segmentation refinement by monte carlo region growing of high confidence detections // Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part II 14. — Springer. 2019. — С. 131—146.

11. Gated-scnn: Gated shape cnns for semantic segmentation / T. Takikawa, D. Acuna, V. Jampani, S. Fidler // Proceedings of the IEEE/CVF international conference on computer vision. — 2019. — С. 5229—5238.

12. An image is worth 16x16 words: Transformers for image recognition at scale / A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly [и др.] // arXiv preprint arXiv:2010.11929. — 2020.

13. Cascadepsp: Toward class-agnostic and very high-resolution segmentation via global and local refinement / H. K. Cheng, J. Chung, Y.-W. Tai, C.-K. Tang // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2020. — С. 8890—8899.

14. Fss-1000: A 1000-class dataset for few-shot segmentation / X. Li, T. Wei, Y. P. Chen, Y.-W. Tai, C.-K. Tang // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2020. — С. 2869—2878.

15. *Litofcenko J.*, *Karner D.*, *Maier F.* Methods for classifying nonprofit organizations according to their field of activity: A report on semi-automated methods based on text // VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations. — 2020. — Т. 31, № 1. — С. 227—237.

16. Pointrend: Image segmentation as rendering / A. Kirillov, Y. Wu, K. He, R. Girshick // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2020. — С. 9799—9808.

17. Segfix: Model-agnostic boundary refinement for segmentation / Y. Yuan, J. Xie, X. Chen, J. Wang // Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16. — Springer. 2020. — С. 489—506.

18. Boundary IoU: Improving object-centric image segmentation evaluation / B. Cheng, R. Girshick, P. Dollár, A. C. Berg, A. Kirillov // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2021. — С. 15334—15342.

19. Deep interactive thin object selection / J. H. Liew, S. Cohen, B. Price, L. Mai, J. Feng // Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. — 2021. — С. 305—314.

20. Emerging properties in self-supervised vision transformers / M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin // Proceedings of the IEEE/CVF international conference on computer vision. — 2021. — С. 9650—9660.

21. Learning transferable visual models from natural language supervision / A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark [и др.] // International conference on machine learning. — PmLR. 2021. — С. 8748—8763.

22. Localizing objects with self-supervised transformers and no labels / O. Siméoni, G. Puy, H. V. Vo, S. Roburin, S. Gidaris, A. Bursuc, P. Pérez, R. Marlet, J. Ponce // arXiv preprint arXiv:2109.14279. — 2021.

23. Look closer to segment better: Boundary patch refinement for instance segmentation / C. Tang, H. Chen, X. Li, J. Li, Z. Zhang, X. Hu // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2021. — С. 13926—13935.

24. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation / J. Li, D. Li, C. Xiong, S. Hoi // International conference on machine learning. — PMLR. 2022. — С. 12888—12900.

25. Denseclip: Language-guided dense prediction with context-aware prompting / Y. Rao, W. Zhao, G. Chen, Y. Tang, Z. Zhu, G. Huang, J. Zhou, J. Lu // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2022. — С. 18082—18091.

26. High quality segmentation for ultra high-resolution images / T. Shen, Y. Zhang, L. Qi, J. Kuen, X. Xie, J. Wu, Z. Lin, J. Jia // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2022. — С. 1310—1319.

27. Highly accurate dichotomous image segmentation / X. Qin, H. Dai, X. Hu, D.-P. Fan, L. Shao, L. Van Gool // European Conference on Computer Vision. — Springer. 2022. — С. 38—56.

28. Language-driven semantic segmentation / B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, R. Ranftl // arXiv preprint arXiv:2201.03546. — 2022.

29. *Lüddecke T.*, *Ecker A. S.* Image segmentation using text and image prompts // CVPR. — 2022.

30. *Lüddecke T.*, *Ecker A.* Image segmentation using text and image prompts // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2022. — С. 7086—7096.

31. Mask transfiner for high-quality instance segmentation / L. Ke, M. Danelljan, X. Li, Y.-W. Tai, C.-K. Tang, F. Yu // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2022. — С. 4412—4421.

32. Patchdct: Patch refinement for high quality instance segmentation / Q. Wen, J. Yang, X. Yang, K. Liang // The Eleventh International Conference on Learning Representations. — 2022.

33. *Shin G.*, *Xie W.*, *Albanie S.* Reco: Retrieve and co-segment for zero-shot transfer // Advances in Neural Information Processing Systems. — 2022. — Т. 35. — С. 33754—33767.

34. Video mask transfiner for high-quality video instance segmentation / L. Ke, H. Ding, M. Danelljan, Y.-W. Tai, C.-K. Tang, F. Yu // European Conference on Computer Vision. — Springer. 2022. — С. 731—747.

35. Cut and learn for unsupervised object detection and instance segmentation / X. Wang, R. Girdhar, S. X. Yu, I. Misra // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2023. — С. 3124—3134.

36. Generalized decoding for pixel, image, and language / X. Zou, Z.-Y. Dou, J. Yang, Z. Gan, L. Li, C. Li, X. Dai, H. Behl, J. Wang, L. Yuan [и др.] // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2023. — С. 15116—15127.

37. *Geus D. de*, *Dubbelman G.* Intra-batch supervision for panoptic segmentation on high-resolution images // Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. — 2023. — С. 3165—3173.

38. Inpaint anything: Segment anything meets image inpainting / T. Yu, R. Feng, R. Feng, J. Liu, X. Jin, W. Zeng, Z. Chen // arXiv preprint arXiv:2304.06790. — 2023.

39. Input augmentation with sam: Boosting medical image segmentation with segmentation foundation model / Y. Zhang, T. Zhou, S. Wang, P. Liang, Y. Zhang, D. Z. Chen // International Conference on Medical Image Computing and Computer-Assisted Intervention. — Springer. 2023. — С. 129—139.

40. Oneformer: One transformer to rule universal image segmentation / J. Jain, J. Li, M. T. Chiu, A. Hassani, N. Orlov, H. Shi // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2023. — С. 2989—2998.

41. Roadway snow detection using dual-spectrum camera images and computer vision / X. He, Y. Wu, K. Zhang, X. Zhu, X. Yang // Transportation research record. — 2023. — Т. 2677, № 10. — С. 220—232.

42. Segment and track anything / Y. Cheng, L. Li, Y. Xu, X. Li, Z. Yang, W. Wang, Y. Yang // arXiv preprint arXiv:2305.06558. — 2023.

43. Segment Anything / A. Kirillov, E. Mintun, N. Ravi [и др.] // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). — 2023.

44. Segment anything in high quality / L. Ke, M. Ye, M. Danelljan, Y.-W. Tai, C.-K. Tang, F. Yu [и др.] // Advances in Neural Information Processing Systems. — 2023. — Т. 36. — С. 29914—29934.

45. Segment anything model (sam) meets glass: Mirror and transparent objects cannot be easily detected / D. Han, C. Zhang, Y. Qiao, M. Qamar, Y. Jung, S. Lee, S.-H. Bae, C. S. Hong // arXiv preprint arXiv:2305.00278. — 2023.

46. *Shen Q.*, *Yang X.*, *Wang X.* Anything-3d: Towards single-view anything reconstruction in the wild // arXiv preprint arXiv:2304.10261. — 2023.

47. *Tang L.*, *Xiao H.*, *Li B.* Can sam segment anything? when sam meets camouflaged object detection // arXiv preprint arXiv:2304.04709. — 2023.

48. Track anything: Segment anything meets videos / J. Yang, M. Gao, Z. Li, S. Gao, F. Wang, F. Zheng // arXiv preprint arXiv:2304.11968. — 2023.

49. Unsupervised object localization: Observing the background to discover objects / O. Siméoni, C. Sekkat, G. Puy, A. Vobeckỳ, É. Zablocki, P. Pérez // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2023. — С. 3176—3186.

50. Video owl-vit: Temporally-consistent open-world localization in video / G. Heigold, M. Minderer, A. Gritsenko, A. Bewley, D. Keysers, M. Lučić, F. Yu, T. Kipf // Proceedings of the IEEE/CVF International Conference on Computer Vision. — 2023. — C. 13802—13811.

51. Vision transformers need registers / T. Darcet, M. Oquab, J. Mairal, P. Bojanowski // arXiv preprint arXiv:2309.16588. — 2023.

52. Cat-sam: Conditional tuning for few-shot adaptation of segment anything model / A. Xiao, W. Xuan, H. Qi, Y. Xing, R. Ren, X. Zhang, L. Shao, S. Lu // European Conference on Computer Vision. — Springer. 2024. — C. 189—206.

53. CLIP-DINOiser: Teaching CLIP a few DINO tricks for open-vocabulary semantic segmentation / M. Wysoczańska, O. Siméoni, M. Ramamonjisoa, A. Bursuc, T. Trzciński, P. Pérez // European Conference on Computer Vision. — Springer. 2024. — C. 320—337.

54. Fossil: Free open-vocabulary semantic segmentation through synthetic references retrieval / L. Barsellotti, R. Amoroso, L. Baraldi, R. Cucchiara // Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. — 2024. — C. 1464—1473.

55. Grounding dino: Marrying dino with grounded pre-training for open-set object detection / S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su [и др.] // European Conference on Computer Vision. — Springer. 2024. — C. 38—55.

56. Hrsam: Efficiently segment anything in high-resolution images / Y. Huang, W. Lai, J. Ji, L. Cao, S. Zhang, R. Ji // arXiv preprint arXiv:2407.02109. — 2024.

57. Pa-sam: Prompt adapter sam for high-quality image segmentation / Z. Xie, B. Guan, W. Jiang, M. Yi, Y. Ding, H. Lu, L. Zhang // 2024 IEEE International Conference on Multimedia and Expo (ICME). — IEEE. 2024. — C. 1—6.

58. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation / M. Lan, C. Chen, Y. Ke, X. Wang, L. Feng, W. Zhang // European Conference on Computer Vision. — Springer. 2024. — C. 70—88.

59. RobustSAM: segment anything robustly on degraded images / W.-T. Chen, Y.-J. Vong, S.-Y. Kuo, S. Ma, J. Wang // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2024. — C. 4081—4091.

60. Sam-clip: Merging vision foundation models towards semantic and spatial understanding / H. Wang, P. K. A. Vasu, F. Faghri, R. Vemulapalli, M. Farajtabar, S. Mehta, M. Rastegari, O. Tuzel, H. Pouransari // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2024. — C. 3635—3647.

61. Talking to DINO: Bridging Self-Supervised Vision Backbones with Language for Open-Vocabulary Segmentation / L. Barsellotti, L. Bianchi, N. Messina, F. Carrara, M. Cornia, L. Baraldi, F. Falchi, R. Cucchiara // arXiv preprint arXiv:2411.19331. — 2024.

62. VRP-SAM: SAM with visual reference prompt / Y. Sun, J. Chen, S. Zhang, X. Zhang, Q. Chen, G. Zhang, E. Ding, J. Wang, Z. Li // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2024. — C. 23565—23574.

63. How to Segment Wires, Hoses, and Other Obstructive Objects in the Traversable Area of a Mobile Robot / L. Vetoshkin, O. Matykina, M. Kichik, U. Izmesteva, D. Yudin // Hoses, and Other Obstructive Objects in the Traversable Area of a Mobile Robot. — 2025.

64. *Vetoshkin L.*, *Yudin D.* Talk2SAM: Text-Guided Semantic Enhancement for Complex-Shaped Object Segmentation // arXiv preprint arXiv:2506.05396. — 2025.