УДК 004.8:004.896

ВИЗУАЛЬНОЕ ЦЕЛЕУКАЗАНИЕ КАК МЕХАНИЗМ ПРОСТРАНСТВЕННОГО ЗАЗЕМЛЕНИЯ VLA-МОДЕЛЕЙ В РОБОТИЗИРОВАННОЙ МАНИПУЛЯЦИИ

Рашитов И.Н.[1]

Магистрант

Московский технический университет связи и информатики

Россия, Москва

Аннотация. В статье рассматривается визуальное целеуказание как способ повышения пространственного заземления моделей класса «зрение – язык – действие» (Vision-Language-Action, VLA) в задачах роботизированной манипуляции. Показано, что языковая инструкция не всегда однозначно задает целевой объект, особенно при наличии похожих экземпляров, окклюзий и требований к точному позиционированию. Обоснована роль масок, рамок, ключевых точек, траекторных следов и кликовых указаний как интерфейса между восприятием сцены и VLA-политикой.

Ключевые слова: визуальное целеуказание, модели «зрение – язык – действие», VLA, роботизированная манипуляция, пространственное заземление, маски сегментации, ключевые точки, Segment Anything.

VISUAL PROMPTING AS A MECHANISM OF SPATIAL GROUNDING FOR VLA MODELS IN ROBOTIC MANIPULATION

Rashitov I.N.

Graduate Student

Moscow Technical University of Communications and Informatics

[1] Научный руководитель – к.т.н., доцент Воронов В.И.

Russia, Moscow

Abstract. The article examines visual prompting as a method for improving spatial grounding in Vision-Language-Action (VLA) models for robotic manipulation. It is shown that a language instruction does not always specify the target object unambiguously, especially in scenes with similar instances, occlusions, and strict positioning requirements. The role of masks, bounding boxes, keypoints, trajectory traces, and click-based cues is substantiated as an interface between scene perception and the VLA policy.

Keywords: visual prompting, Vision-Language-Action, VLA, robotic manipulation, spatial grounding, segmentation masks, keypoints, Segment Anything.

Введение

Модели класса «зрение – язык – действие» (Vision-Language-Action, VLA) связывают изображение сцены, естественно-языковую инструкцию и действие робота. Такая архитектура повышает гибкость управления манипулятором, но одновременно возлагает на исполнительную модель несколько разнородных задач: распознавание объекта, интерпретацию пространственных отношений, выбор целевого экземпляра и генерацию моторных команд. В сценах с похожими объектами или строгими требованиями к позиционированию языковое описание перестает быть достаточным интерфейсом управления. Современные обзоры VLA-моделей также связывают развитие этого класса систем с переходом от распознавания сцены к управлению действием в единой модели [9].

Визуальное целеуказание позволяет снизить эту неоднозначность за счет промежуточных представлений: масок, рамок, ключевых точек, кликовых указаний и визуальных следов. Они переводят семантическую цель из текста в геометрию сцены и делают ее наблюдаемой для исполнительной VLA-политики.

Цель и задачи исследования

Цель статьи состоит в обосновании визуального целеуказания как механизма пространственного заземления VLA-моделей. Объектом анализа являются робототехнические архитектуры, в которых визуальная подсказка включается в контур управления; предметом – маски, рамки, ключевые точки, кликовые указания и визуальные следы как интерфейс между семантикой инструкции и моторным действием. Новизна заключается в рассмотрении визуального целеуказания не как приема улучшения изображения, а как самостоятельного интерфейсного слоя между восприятием сцены и VLA-политикой.

Для достижения цели решаются задачи: раскрыть проблему пространственного заземления в VLA-управлении; сопоставить виды визуальных подсказок; определить их функции в контуре роботизированной манипуляции; описать архитектурное включение масок и ключевых точек в цикл управления.

Основная часть

Под пространственным заземлением в роботизированной манипуляции понимается связывание языкового намерения с конкретной областью физической сцены. Если в текстовой инструкции присутствует объект, система должна определить не только класс объекта, но и его экземпляр, положение, границы, доступность для захвата и отношение к целевой зоне. Для человека эти операции часто сливаются в единый акт восприятия. Для VLA-модели они образуют несколько вычислительных задач, каждая из которых может стать источником ошибки.

Классическая VLA-архитектура стремится решить эти задачи за один прямой проход модели: изображение и инструкция поступают на вход, а на выходе формируется действие. VP-VLA указывает на слабость такой схемы: одна модель одновременно выполняет интерпретацию инструкции,

пространственное заземление и низкоуровневое управление, что может снижать точность позиционирования [1]. Из этого следует инженерный вывод: чем выше требования к локализации объекта и целевой зоны, тем менее оправданно оставлять пространственную привязку полностью неявной.

Ограничение проявляется и на уровне пиксельного понимания сцены. PixelVLA фиксирует, что существующие VLA-модели испытывают трудности с пониманием сцены на пиксельном уровне и чрезмерно зависят от текстовых подсказок [2]. Для манипуляции это критично, поскольку физическое действие определяется не категорией объекта в целом, а его геометрией: контуром, центром, доступной стороной захвата и положением относительно других предметов. Ошибка в несколько сантиметров может быть незначительной для визуального описания, но критичной для захвата.

Сегментационные модели семейства Segment Anything создают основу для вынесения части пространственного заземления из VLA-политики в специализированный модуль восприятия. SAM был предложен как модель сегментации, управляемая подсказками и способная переноситься на новые распределения изображений [3]. Для робототехники это означает, что целевой объект может быть выделен не только как класс, но и как конкретная область сцены, пригодная для последующего целеуказания и проверки результата.

Сама по себе сегментация не решает задачу управления: она дает маску, рамку или идентификатор объекта, но исполнительная политика должна использовать эти данные. Поэтому визуальное целеуказание выступает интерфейсом, через который геометрическая информация передается VLA-политике в той же модальности, что и входное изображение.

Визуальные подсказки различаются по форме и функции. Ограничивающая рамка задает приблизительное положение объекта. Маска передает форму и площадь. Ключевая точка фиксирует центр или функционально значимую область. Траекторный след сообщает историю

движения. Клик оператора устраняет неоднозначность выбора экземпляра. Эти формы нельзя рассматривать как взаимозаменяемые: каждая из них решает свой класс ошибок.

VP-VLA использует структурированные визуальные подсказки, включая перекрестия и ограничивающие рамки, чтобы разделить высокоуровневое рассуждение и низкоуровневое исполнение [1]. Планировщик определяет цель, а исполнительный контроллер выполняет действие, ориентируясь на визуальные якоря. Такой подход показывает, что визуальная подсказка может быть не вспомогательной разметкой, а архитектурной границей между рассуждением и моторным управлением.

TraceVLA вводит визуальное целеуказание по следам траектории: состояние и предыдущие действия кодируются непосредственно в изображении [4]. Это важно для задач, где текущий кадр не содержит всей истории. Визуальный след превращает историю движения в наблюдаемую подсказку и снижает зависимость политики от скрытого состояния.

PixelVLA усиливает пиксельную сторону проблемы. В нем используются пиксельно-ориентированный кодировщик и кодировщик визуальных подсказок, что позволяет объединять текстовые и визуальные входы [2]. Для манипуляции это особенно ценно: модель получает возможность учитывать не только категориальную семантику, но и геометрию объекта на уровне изображения.

VCA предлагает практичный вариант устранения неоднозначности: оператор выбирает цель кликом в 2D-изображении, после чего сегментационная модель уточняет объект [5]. В сценах с похожими предметами это снижает неоднозначность радикальнее, чем добавление прилагательных в текстовую команду. Если перед роботом три одинаковых кубика, фраза «левый красный кубик» зависит от системы координат и ракурса, а клик непосредственно указывает экземпляр.

RoboGround рассматривает маски пространственного заземления как промежуточное представление для управления манипуляцией [6]. Маска не только указывает цель, но и передает сведения о форме и размере объекта. Это делает ее более информативной, чем точка или рамка, особенно если требуется захватить предмет нестандартной формы или поместить его в ограниченную область.

Близкая логика разделения семантики и пространства прослеживается в CLIPort: языково-обусловленная манипуляция строится через связку семантического пути «что требуется выполнить» и пространственного пути «где должно быть выполнено действие» [7]. Для визуального целеуказания это подтверждает необходимость отдельного механизма пространственной привязки, а не только распознавания смысла команды.

Таблица 1. Типология визуальных подсказок для VLA-управления

Вид визуальной подсказки	**Основная функция**	**Преимущество**	**Ограничение**
Ограничивающая рамка	Грубая локализация объекта	Простота формирования и интерпретации	Не передает точную форму
Маска сегментации	Геометрическое выделение объекта или зоны	Содержит форму, площадь и контур	Требует устойчивой сегментации
Ключевая точка или перекрестие	Указание центра или опорной точки	Удобна для захвата и позиционирования	Может быть недостаточной для сложной формы
Визуальный след	Передача истории движения	Улучшает пространственно-временную осведомленность	Требует корректной записи траектории
Клик по объекту	Устранение неоднозначности экземпляра	Минимизирует языковую неопределенность	Предполагает участие оператора

Сопоставление, представленное в таблице 1, показывает, что визуальные подсказки различаются не только формой представления, но и ролью в контуре управления. Ограничивающая рамка и ключевая точка преимущественно решают задачу локализации, маска сегментации уточняет геометрию объекта или зоны, визуальный след добавляет информацию о предшествующих

действиях, а клик по объекту снижает неоднозначность пользовательской инструкции. Поэтому выбор типа подсказки должен определяться не удобством визуализации, а характером ошибки, которую требуется компенсировать: ошибкой выбора экземпляра, неточностью позиционирования, недостатком истории движения или слабой геометрической привязкой цели.

Функционально визуальное целеуказание выполняет четыре задачи: уточняет объект действия, задает пространственную цель, снижает нагрузку на языковую часть модели и повышает диагностируемость управления. Если подсказка построена неверно, источник ошибки можно локализовать в модуле восприятия, а не в VLA-политике целиком.

В инженерной реализации визуальное целеуказание должно учитываться не только при выводе модели, но и при подготовке данных. Если на этапе эксплуатации VLA-политика получает изображения с наложенными масками, а обучалась только на чистых кадрах, возникает сдвиг распределения. Поэтому визуальные подсказки целесообразно использовать и при формировании обучающих данных: в виде дополнительных каналов, аугментированных кадров или пар «исходное изображение – изображение с подсказкой».

Общий контур можно представить так:

$$I' = R(I, M_o, M_z, K) \quad (1)$$

где I – исходное RGB-изображение; M_o – маска целевого объекта; M_z – маска целевой зоны; K – набор ключевых точек или дополнительных визуальных якорей; R – функция рендеринга подсказки; I' – аугментированное изображение, поступающее в VLA-политику.

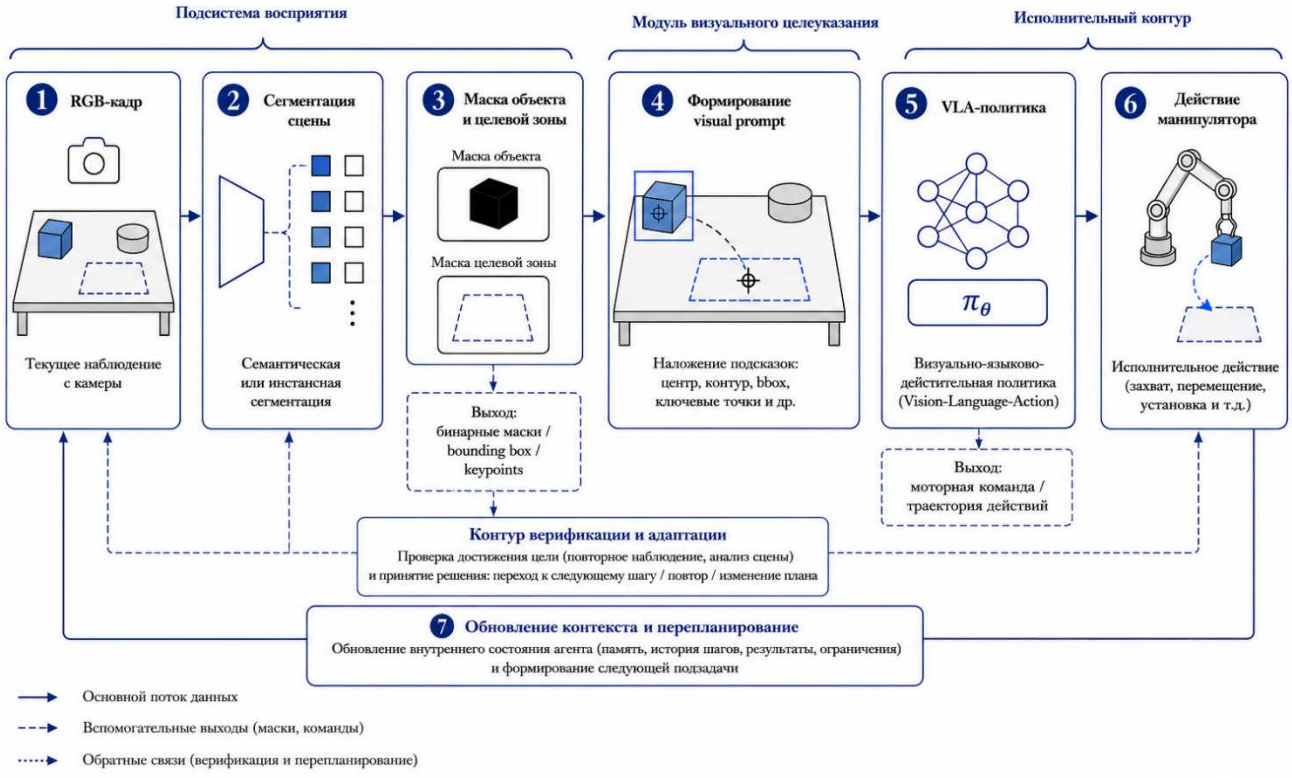

Рисунок 1. Архитектурная схема визуального целеуказания для многошаговой роботизированной манипуляции

Примечание – разработано автором.

Рисунок 1 показывает переход от исходного RGB-наблюдения к действию манипулятора через промежуточные визуальные представления. Подсистема восприятия сегментирует сцену и формирует маски объекта и целевой зоны. Модуль визуального целеуказания накладывает подсказки в виде рамок, контуров, ключевых точек или визуальных следов. VLA-политика получает пространственно уточненное наблюдение и формирует моторную команду, а контур верификации обеспечивает повторное наблюдение, проверку цели и обновление контекста.

Формула (1) отражает перевод семантики в геометрию. Планировщик формулирует подзадачу, сегментационный модуль преобразует объект и целевую зону в маски, а модуль визуального целеуказания накладывает их на изображение. В результате VLA-политика получает не абстрактное описание, а визуально выделенную цель и зону размещения.

Важна и единая визуальная семантика: если объект захвата и зона размещения маркируются устойчивыми типами подсказок, модель получает повторяемую связь между визуальным кодом и функцией области. Такая разметка не должна зависеть от класса объекта: кубик, цилиндр или деталь получают одинаковый функциональный код, если выполняют роль цели.

Визуальное целеуказание не является универсальным решением: его надежность ограничена качеством сегментации, стабильностью отслеживания и согласованностью камер. Поэтому подсказка должна быть связана с контуром проверки: после действия система повторно локализует объект и оценивает достижение цели. Наиболее устойчива схема, где маски используются дважды: до действия – для целеуказания, после действия – для верификации.

В многошаговой задаче каждая подзадача получает собственный визуальный контекст, а не наследует неявное состояние из предыдущего шага. После перемещения первого объекта система заново строит маски и формирует новую визуальную подсказку для следующего объекта. Это снижает риск каскадного накопления ошибок, поскольку каждый шаг начинается с актуализированного восприятия.

Аналитически визуальное целеуказание можно рассматривать как компромисс между полностью символическим планированием и сквозным нейросетевым управлением. Символическая система требует явного описания объектов, зон и правил. Сквозная VLA-архитектура пытается выучить эти связи неявно. Визуальная подсказка занимает промежуточное положение: она не навязывает жесткую символическую модель мира, но делает целевые пространственные отношения наблюдаемыми для нейросетевой политики.

На исполнительном уровне такая схема может сочетаться с различными визуомоторными политиками, включая диффузионные политики, где действие робота формируется как визуально обусловленный процесс генерации управляющей последовательности [8]. Для практического внедрения важны

компактные варианты VLA-моделей, поскольку скорость вывода и объем обучающих данных становятся ограничениями при развертывании на реальном роботе [10].

Результаты

Визуальное целеуказание является одним из перспективных механизмов повышения надежности VLA-моделей в роботизированной манипуляции. Его значение связано не с улучшением визуальной наглядности для человека, а с преобразованием языковой цели в пространственно выраженную структуру, пригодную для исполнительной политики.

Анализ VP-VLA, TraceVLA, PixelVLA, VCA, Segment Anything, RoboGround, CLIPort и Diffusion Policy показывает общий сдвиг к более явному пространственному заземлению VLA-моделей, особенно при точной манипуляции, похожих объектах и длинных сценариях.

Заключение

Для многошаговой системы наиболее обоснована архитектура, в которой визуальное целеуказание связывает агентный планировщик, сегментационную подсистему и локальную VLA-политику. Планировщик задает цель, сегментационная модель выделяет объект и зону, модуль визуального целеуказания формирует аугментированное наблюдение, а VLA-политика выполняет локальное действие. После этого маски снова используются для проверки результата. Такая схема может снижать неоднозначность, повышать диагностируемость и служить основой для дообучения компактных VLA-моделей на локальных наборах данных с явной визуальной семантикой.

Библиографический список

1. Wang Z. et al. VP-VLA: Visual Prompting as an Interface for Vision-Language-Action Models // arXiv preprint. – 2026. – arXiv:2603.22003. – DOI: 10.48550/arXiv.2603.22003.

2. Liang W. et al. PixelVLA: Advancing Pixel-level Understanding in Vision-Language-Action Model // arXiv preprint. – 2025. – arXiv:2511.01571. – DOI: 10.48550/arXiv.2511.01571.
3. Kirillov A. et al. Segment Anything // Proceedings of the IEEE/CVF International Conference on Computer Vision. – 2023. – P. 4015–4026.
4. Zheng R. et al. TraceVLA: Visual Trace Prompting Enhances Spatial-Temporal Awareness for Generalist Robotic Policies // arXiv preprint. – 2024. – arXiv:2412.10345. – DOI: 10.48550/arXiv.2412.10345.
5. Kim D., Jan S., Park H., Lim D. VCA: Vision-Click-Action Framework for Precise Manipulation of Segmented Objects in Target Ambiguous Environments // arXiv preprint. – 2026. – arXiv:2602.23583. – DOI: 10.48550/arXiv.2602.23583.
6. Huang H. et al. RoboGround: Robotic Manipulation with Grounded Vision-Language Priors // arXiv preprint. – 2025. – arXiv:2504.21530. – DOI: 10.48550/arXiv.2504.21530.
7. Shridhar M., Manuelli L., Fox D. CLIPort: What and Where Pathways for Robotic Manipulation // arXiv preprint. – 2021. – arXiv:2109.12098.
8. Chi C. et al. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion // arXiv preprint. – 2023. – arXiv:2303.04137. – DOI: 10.48550/arXiv.2303.04137.
9. Ud Din M., Akram W., Saoud L. S., Rosell J., Hussain I. Vision Language Action Models in Robotic Manipulation: A Systematic Review // arXiv preprint. – 2025. – arXiv:2507.10672. – DOI: 10.48550/arXiv.2507.10672.
10. Wen J. et al. TinyVLA: Towards Fast, Data-Efficient Vision-Language-Action Models for Robotic Manipulation // arXiv preprint. – 2024. – arXiv:2409.12514. – DOI: 10.48550/arXiv.2409.12514.