Введение

Модели класса Vision–Language–Action (VLA) представляют собой перспективное направление в области робототехники и embodied AI, так как они объединяют визуальное восприятие, обработку естественного языка и генерацию действий. Основной предпосылкой их эффективности является использование предварительно обученных Vision–Language моделей (VLM), которые содержат обширные знания об окружающем мире и обеспечивают устойчивое vision–language (VL) соответствие.

Однако при адаптации таких моделей к задачам управления действиями остаётся открытым вопрос: в какой степени исходные VL-представления сохраняются после fine-tuning на action-ориентированных данных. Ряд недавних работ показывает, что наивное дообучение может приводить к деградации визуальных представлений, что отрицательно сказывается на обобщающей способности модели, особенно в условиях выхода за распределение (out-of-distribution, OOD).

Целью данной работы является анализ сохранения VL-представлений в модели SmolVLA при различных стратегиях fine-tuning, а также сравнение их влияния на стабильность обучения и сохранение визуального знания.

Обзор связанных работ

Современные VLA-модели активно используют предварительно обученные VLM в качестве основы для принятия решений в среде. Несмотря на высокое качество визуального и языкового представления, такие модели подвержены эффекту representational drift при адаптации к новым задачам управления.

Для диагностики и смягчения деградации представлений используются методы зондирования скрытых слоёв, анализ attention-карт, а также различные стратегии выравнивания представлений. Особое внимание уделяется тому, как fine-tuning влияет на внутренние VL-структуры и способность модели к OOD-обобщению.

Метод

Архитектура модели

SmolVLA — это компактная Vision–Language–Action модель, которая принимает на вход:

- визуальные наблюдения с камеры,
- текущее состояние системы,
- текстовую инструкцию на естественном языке.

Полученные признаки используются для формирования контекстного представления, которое затем применяется для генерации действий.

Стратегии обучения

В рамках работы были рассмотрены три конфигурации fine-tuning:

1. **Базовая модель (flow-matching loss)**

Стандартная конфигурация обучения SmolVLA, используемая в оригинальной реализации.

2. **L1-регрессия**

Альтернативная функция потерь, применяемая для анализа скорости сходимости и стабильности обучения.

3. **Замороженный vision-энкодер**

Конфигурация, при которой визуальный энкодер не обновляется во время обучения, что позволяет снизить деградацию предварительно обученных визуальных представлений.

Экспериментальная установка

Датасет Для обучения использовался демонстрационный датасет pick-and-place, содержащий эпизоды манипуляций с объектами при различных начальных условиях. Датасет был размещён на платформе Hugging Face Hub для обеспечения воспроизводимости экспериментов.

Параметры обучения Обучение проводилось с использованием фреймворка LeRobot в среде Google Colab. Все модели обучались в течение 200 шагов, что позволило провести сравнительный анализ при ограниченных вычислительных ресурсах. Полученные контрольные точки моделей были загружены на Hugging Face Hub

Результаты

Количественное сравнение

Таблица 1 демонстрирует сравнение различных стратегий fine-tuning.

Таблица 1. Сравнение стратегий fine-tuning SmolVLA

Модель	Функция потерь	Vision-энкодер	Шаг Примечание
SmolVLA (базовая)	Flow-matching	Обучаемый	200 Базовая конфигурация
SmolVLA + L1	L1-регрессия	Обучаемый	200 Более стабильная сходимость
SmolVLA (frozen)	Flow-matching	Заморожен	200 Лучшее сохранение VL

Анализ attention-карт

Анализ attention-карт выполнялся в соответствии с методологией, описанной в разделе 5.1 работы «Don’t Blind Your VLA». Attention-распределения позволяют оценить степень визуального соответствия и фокусировки модели на релевантных областях изображения.

В условиях ограниченных вычислительных ресурсов был проведён качественный анализ. Согласно ранее опубликованным результатам, наивное action fine-tuning приводит к смещению attention-карт, тогда как заморозка

визуального энкодера способствует сохранению устойчивого визуального фокуса.

Анализ пространства представлений (t-SNE)

Для анализа структуры скрытых представлений используется t-SNE-визуализация, позволяющая выявить эффект collapse представлений после fine-tuning. В данной работе обсуждаются ожидаемые тенденции на основе методологии раздела 5.2 указанной статьи.

Предыдущие исследования показывают, что отсутствие ограничений при обучении может приводить к снижению разделимости классов, тогда как сохранение или выравнивание визуальных представлений способствует устойчивой структуре embedding-пространства и лучшему OOD-обобщению.

Обсуждение

Полученные результаты подтверждают наличие компромисса между адаптацией модели к задачам управления и сохранением предварительно обученных VL-представлений. Использование L1-регрессии демонстрирует сопоставимую сходимость на коротких интервалах обучения, тогда как заморозка vision-энкодера наиболее эффективно снижает деградацию визуальных представлений.

Это подтверждает выводы о том, что наивное fine-tuning может негативно влиять на inherited VL-знания, а стратегии ограничения обновлений являются практичным способом их сохранения.

Заключение

В данной работе был проведён анализ сохранения vision–language представлений в модели SmolVLA при различных стратегиях fine-tuning. Показано, что стандартное обучение эффективно решает задачу управления, но сопровождается риском деградации визуальных представлений.

Альтернативные стратегии, такие как использование L1-регрессии и заморозка визуального энкодера, позволяют смягчить этот эффект и обеспечить более устойчивое обобщение. Работа подчёркивает важность контроля и анализа внутренних представлений при адаптации VLM-основанных моделей к action-ориентированным задачам.

Ресурсы и воспроизводимость

Эксперименты выполнялись в Google Colab с использованием библиотеки LeRobot. Все модели и датасеты доступны публично на Hugging Face Hub.

Датасет:
https://huggingface.co/datasets/Sythen/svla_so100_pickplace_copy

Модели:
Baseline: https://huggingface.co/Sythen/my_smolvla_training_fm
L1: https://huggingface.co/Sythen/my_smolvla_training_l1
Frozen: https://huggingface.co/Sythen/my_smolvla_training_frozen

Список литературы

[2] N. Kachaev et al. *Don't Blind Your VLA: Aligning Visual Representations for OOD Generalization*, arXiv, 2025.
[7] B. Liu et al. *LIBERO: Benchmarking Knowledge Transfer for Lifelong Robot Learning*, NeurIPS, 2023.
[8] M. Shukor et al. *SmolVLA: A Vision-Language-Action Model for Affordable and Efficient Robotics*, arXiv, 2025.