

# USING VISION LANGUAGE MODELS FOR OPTIMIZING LIGHT INDUSTRY TASKS

Сидельников Никита, ИТМО

## Лёгкая промышленность сегодня

Прежде чем перейти к описанию проделанной работы, кратко опишем область и существующие в ней сегодня проблемы.

Лёгкая промышленность – это совокупность отраслей промышленности, осуществляющих первичную обработку различных видов сырья и производство из них предметов массового потребления.

Сегодня данная отрасль в нашей стране значительно отстает от многих других. Это происходит из-за достаточно большого списка проблем. Одними из основных являются:

- Кадровый дефицит;
- Технологическая и техническая отсталость;
- Низкий уровень научной и инновационной деятельности;
- Отсутствие сырьевой базы и комплектующих.

Для решения данных проблем можно разрабатывать новые технологии и находить способы применения уже существующих. Например, автоматизация производства может значительно упростить процессы на предприятиях, а благодаря быстрому развитию технологий искусственного интеллекта её можно внедрить на новые этапы производственного цикла, где этого автоматизация не применялась. Моя работа посвящена именно этой теме.

## Постановка задачи

В рамках магистерской диссертации я занимался разработкой MVP системы автоматизированного контроля производственного потока швейного предприятия. Основная задача системы – формирование статистики на разных этапах производственного потока в полностью автоматическом режиме.

Все производство можно разделить на этапы, представленные на картинке:



В рамках MVP работа была сфокусирована на этапах *раскроя* и *ВТО*. Некоторые задачи, которые было необходимо решить:

- Классификация изделий по типу и цвету на этапе ВТО
- Сегментация и подсчет занимаемой лекалами площади

Задача на школу: проверить, возможно ли решить поставленные задачи в zero-shot режиме с помощью Vision Language Models.

## Данные

Для решения задач были собраны несколько наборов данных:

Задача	Кол-во изображений	Кол-во классов
Классификация одежды по типу	~2800	2
Классификация одежды по цвету	~2800	5
Сегментация лекал и полотна ткани	~3200	2

Важно подчеркнуть, что для разметки данных была применена автоматизация. Сервис Roboflow предлагает инструменты для автоматической разметки данных в задачах сегментации с использованием моделей семейства SAM.

Для оценки качества работы будем использовать не все данные, а только 1000 примеров.

## Гипотезы

Сформулируем гипотезы, которые проверим в рамках проекта:

- VLM способна решить задачу классификации изделия в кадре по типу с точностью не меньше 0.9 (accuracy)
- VLM способна решить задачу классификации изделия в кадре по цвету с точностью не меньше 0.9 (accuracy)
- VLM способна решить задачу детектирования изделия в кадре с точностью не меньше 0.85 (IoU)
- VLM способна решить задачу сегментирования лекал в кадре с точностью не меньше 0.8 (IoU)
- VLM способна решить несколько поставленных выше задач за одно обращение к ней

## Список выбранных моделей

Ниже представлен список использованных для проверки гипотез моделей. Статьи, описывающие архитектуры и другие особенности, можно найти по ссылкам.

- [MINICPM-LLAMA3-V-2\\_5](#)
- [VILT-B32-FINETUNED-VQA](#)
- [BLIP-VQA-CAPFILT-LARGE](#)
- [FLORENCE-2-LARGE](#)
- [INTERNVL-2](#)
- [PALIGEMMA](#)

Далее дадим обобщенные результаты качества работы моделей для каждой из задач, а затем представим суммированную таблицу с числовыми метриками.

## Классификация изделий по типу

Данная задача состоит в том, чтобы определить изделие какого типа находится в руках у работника – футболка или толстовка. С решением данной задачи справилась только часть моделей. Качество сильно зависит от входного промпта, важно отметить, что много времени уходит на его подбор. Необходимой точности, в конечном итоге, добиться не удалось.

Примеры ответы нескольких моделей представлены ниже на рисунках.

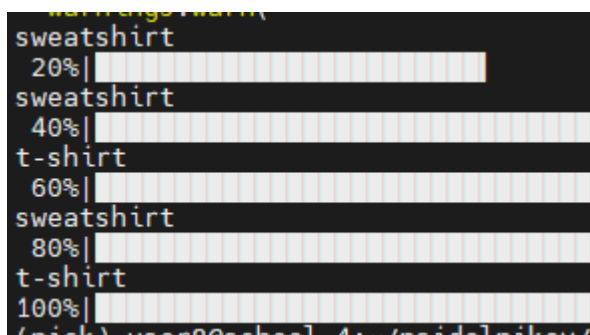


Рисунок 1 – Пример ответа модели minicpm на запрос: 'There is piece of clothing in the hands of person. What is it? tshirt or sweatshirt. Give short answer, type only.'

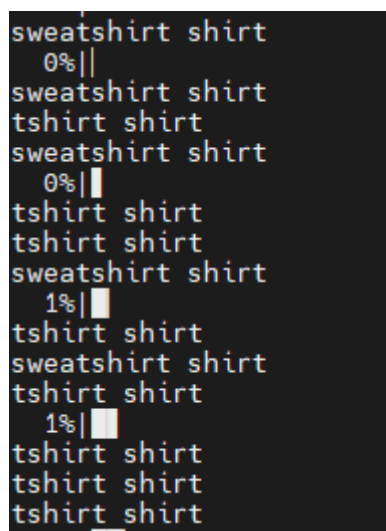


Рисунок 2 – Пример ответа модели vilt на запрос: 'What is the biggest piece of clothing on the picture, tshirt or sweatshirt?'. В левом столбце содержится истинное значение, в правом предсказанное.

## Классификация изделий по цвету

Данная задача состоит в том, чтобы сказать какого цвета изделие находится в руках у сотрудника. Набор данных содержит в себя изделия черного, коричневого, бежевого, белого и розового цветов. С данной задачей справились те же модели, что и с предыдущей. Достичь высокого качества метрики не удалось. Несколько примеров ниже.

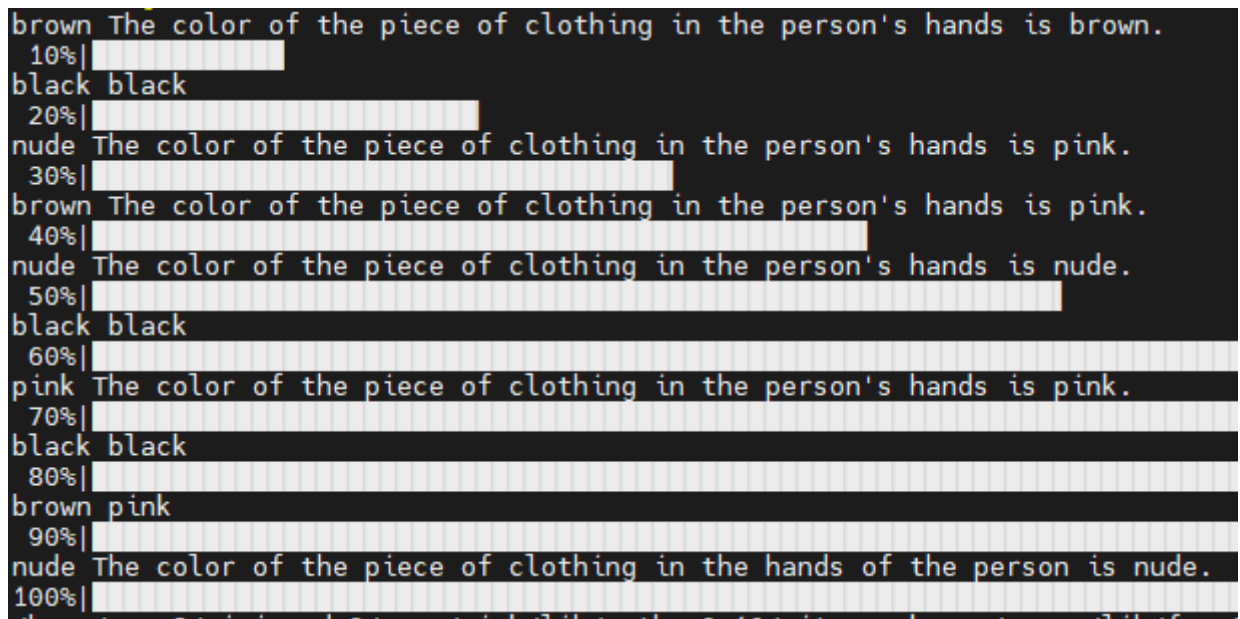


Рисунок 3 – пример ответа модели minicpm на запрос: 'There is piece of clothing in the hands of person. What is its color? Choose one from the list: black, white, pink, brown, nude. Answer shortly, just color name.'

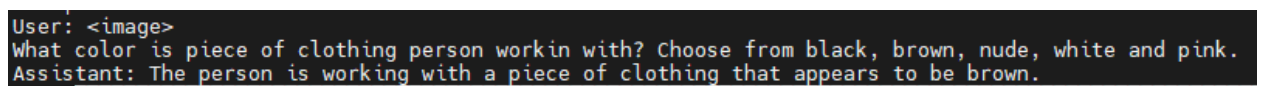


Рисунок 4 – Пример ответа модели internvl2 на запрос: '<image>\nWhat color is piece of clothing person workin with? Choose from black, brown, nude, white and pink.'

## Детектирование

Данная задача состоит в том, чтобы получить ограничивающую рамку изделия, с которым работает сотрудник. С этой задачей справились немного другие модели. Необходимого качества все же достигнуто не было. Примеры запросов и ответов, а также визуализации результатов представлены ниже.

```

[0, 241, 1054, 720]
The biggest piece of clothing in the image is the one being held by the person. It is a purple garment that appears to be mid-sized, likely a shirt or blouse. The bounding box for this item would be approximately from the top left corner of the fabric at (x1, y1) coordinates (0.15, 0.05) to the bottom right corner at (x2, y2) coordinates (0.35, 0.75).
33%|
[0, 261, 930, 720]
The largest piece of clothing in the image is the black garment on the ironing board, which covers a significant portion of the board's surface. The bounding box for this garment would be approximately x1: 28, y1: 195, x2: 628, y2: 1000. These coordinates define the leftmost point (x1, y1) and the rightmost point (x2, y2) of the garment on the ironing board.
67%|
[179, 302, 750, 567.5]
The biggest piece of clothing appears to be the stack of beige garments on the right side of the image. To provide a bounding box, we would typically identify the coordinates where the stack begins and ends in both the x and y directions. However, due to the nature of the image and the angle at which the stack is presented, it's challenging to determine the exact dimensions without additional context or information. If we were to estimate based on visual inspection, the bounding box for the stack could be approximately x1: 650, y1: 250, x2: 1000, y2: 1000, assuming the stack starts slightly above the y=250 mark and extends to the bottom of the image at y=1000. This estimation accounts for the visible portion of the stack and its proximity to the edge of the table.
100%|
(nick) user8@school-4:~/nsidelnikov/src$

```

Рисунок 5 – Пример ответа модели minicpm на запрос: Detect the biggest piece of clothing, give bounding box in format x1, y1, x2, y2

```

warnings.warn(
[0, 241, 1054, 720]
<box>0 494 688 988</box>
33%|
[0, 261, 930, 720]
<box>138 440 509 1000</box>.
67%|
[179, 302, 750, 567.5]
<box>205 604 545 1000</box>
100%|
(nick) user8@school-4:~/nsidelnikov/src$ python mini

```

Рисунок 6 – Пример ответа модели minicpm на запрос: Detect the biggest piece of clothing on the picture and give its bounding box in format x1, y1, x2, y2. Answer should be short, just coordinates.

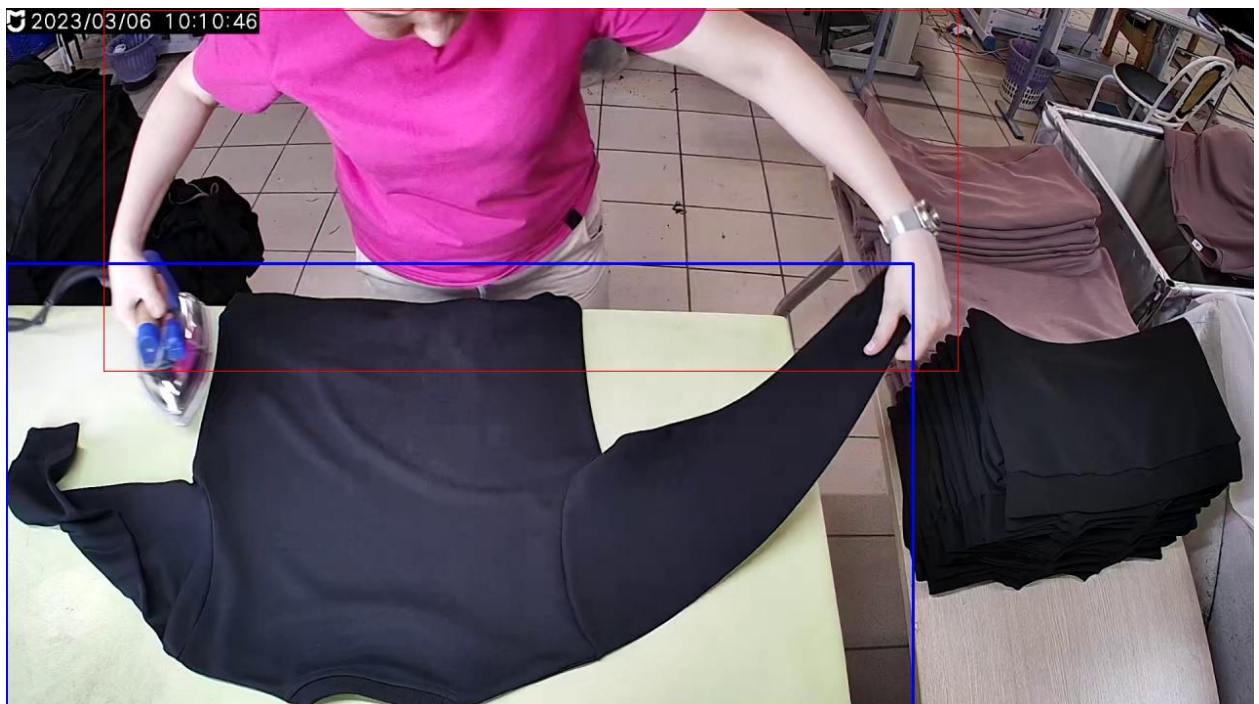


Рисунок 7 – пример плохого ответа florence2





Рисунок 8 – Пример хорошего ответа florence2

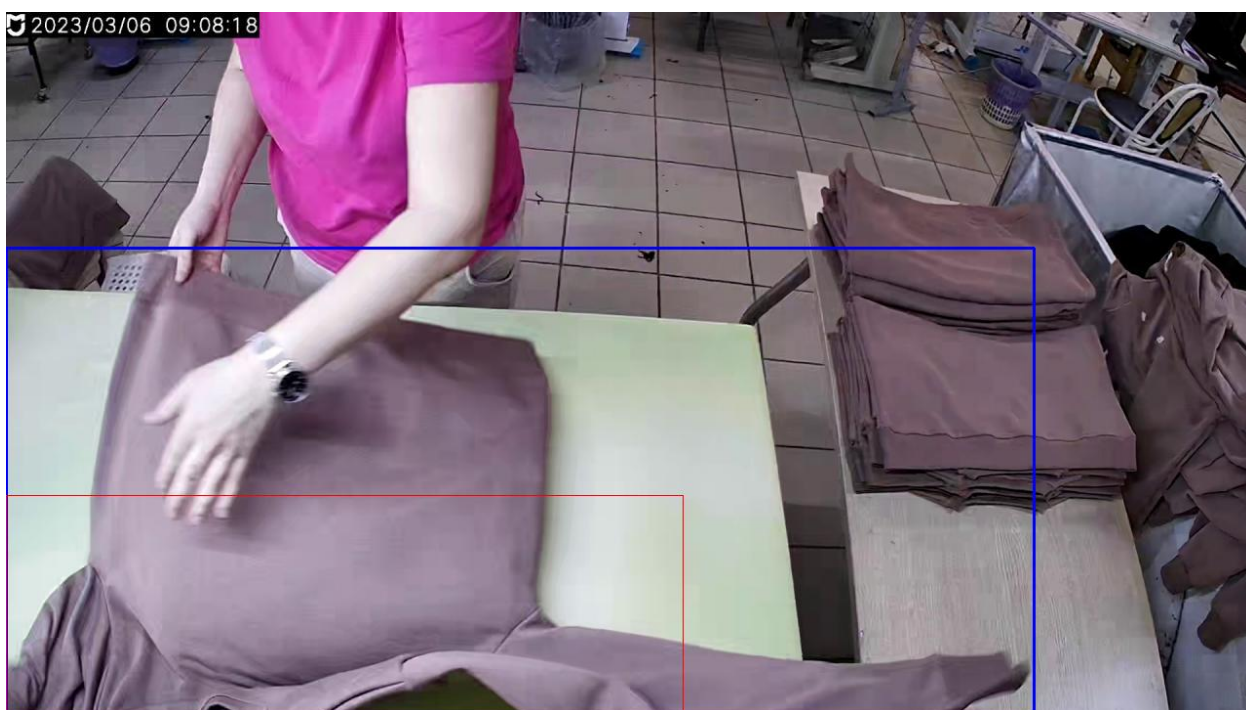


Рисунок 9 – Пример плохого ответа мінісрт

## Сегментация

В рамках данной задачи было необходимо сегментировать картонные лекала, разложенные на поверхности ткани. Ни одна из выбранных моделей не справилась с задачей сегментации. Большинство выдавало просто описательный текст картинки, инструкции для решения задачи сегментации или отвечало, что не может решить данную задачу. Florence2, имеющая функционал для сегментации, не смогла понять, что такое лекала, и где они расположены. Получить хоть как-то правильных масок не удалось. Примеры ответов моделей представлены ниже.

```
User: <image>
Provide segmentation masks of patterns placed on the fabric.
Assistant: To provide segmentation masks of the patterns placed on the fabric, you would typically use a combination of techniques in image processing and machine learning. Given that I can't directly create these masks here, I can guide you on how to achieve this.

### Step-by-Step Process:

### 1. **Identify the Patterns**
Identify patterns such as stripes, checks, and other specific textile designs.

### 2. **Preprocessing**
- **Color Extraction**: Convert the image to grayscale for simpler recognition.
- **Noise Reduction**: Reduce noise using functions in libraries like OpenCV or ImageProcessingToolkit in MATLAB.

### 3. **Detection**
- **Edge Detection**: Use Canny or Sobel algorithms for pattern detection and outline extraction.
- **Pattern Classification**: Use machine learning models, such as Convolutional Neural Networks (CNNs), to identify and segment patterns.

### 4. **Segmentation**
- **Mask Creation**: Generate segmentation masks for the patterns identified in Step 2. This involves creating binary masks where the boundaries of the patterns are clearly demarcated.

### Practical Implementation (Python using OpenCV and Scikit-image):

```python
import cv2
import numpy as np
from skimage.segmentation import slic
from skimage.color import label2rgb
from skimage.util import img_as_ubyte

def segment_fabric(img):
    # Convert to grayscale for simpler processing
    gray = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)

    # Noise reduction
    denoised_img = cv2.fastNlMeansDenoising(gray, None, 10, 10, 7, 21)

    # Edge detection
    edges = cv2.Canny(denoised_img, 50, 150)

    # Segmenting
    segments = slic(edges, n_segments=300, sigma=1.6, enforce_connectedness=True)
    segmented_img = label2rgb(segments, denoised_img, kind='avg')

    return segmented_img

# Load image
img = cv2.imread('fabric.jpg')

# Segment the fabric
```

Рисунок 10 – Пример ответа модели interv12

```
Provide segmentation masks of patterns placed on the fabric in format x1, y1, ..., xn, yn.
Sorry, as a base VLM I am not trained to answer this question.
0%|
Provide segmentation masks of patterns placed on the fabric in format x1, y1, ..., xn, yn.
Sorry, as a base VLM I am not trained to answer this question.
0%|
Provide segmentation masks of patterns placed on the fabric in format x1, y1, ..., xn, yn.
Sorry, as a base VLM I am not trained to answer this question.
```

Рисунок 11 – Пример ответа модели paligemma



## Ответ на несколько вопросов сразу

В рамках данной гипотезы было интересно проверить сможет ли какая-то модель решить сразу несколько описанных выше задач, и одна такая модель нашлась! Ей оказалась – minicpm-llava3-v2.5. Однако, стоит отметить, что при таком формате взаимодействия модель не соблюдает просьбу о форматировании ответа, что усложняет дальнейшую пост-обработку результатов.

```
The clothing item in the photo is a t-shirt, not a sweatshirt. The color of the t-shirt is a shade of brown. The bounding box for the t-shirt is as follows: x1: 0.05, y1: 0.05, x2: 0.45, y2: 0.10.
The clothing in the photo is a long-sleeve garment, which appears to be a sweatshirt due to its length and style. It is black in color. The bounding box for this item is as follows: x1=0, y1=0.20, x2=1.00, y2=1.00.
The clothing item in the photo is a t-shirt. Its color appears to be a shade of nude or light brown. The bounding box for this t-shirt is: t-shirt, nude/brown, box: 0, 344, 622, 944.
The clothing item in the photo is a sweatshirt. The color of the sweatshirt is a shade of purple or lavender. The bounding box for the sweatshirt can be described as follows: the person is wearing a sweatshirt, which occupies the central area of the image, and the coordinates would span from approximately (x1,y1) at the top left corner of the sweatshirt to (x2,y2) at the bottom right corner, where the person's hands are working on it.
The clothing item in the photo is a t-shirt. Its color appears to be nude or a light beige, which is distinct from the pink shirt the person is wearing. The bounding box for the t-shirt is as follows: t-shirt; color, nude; box: x1=0, y1=90, x2=514, y2=1000.
The clothing in the photo is a t-shirt. Its color is black. The bounding box for the t-shirt is as follows: x1: 390, y1: 310, x2: 610, y2: 714.
The clothing in the photo is a t-shirt. It is pink in color. The bounding box of the t-shirt, which is being folded or laid out by the person, can be described as follows: type: t-shirt, color: pink, box: x1: 0, y1: 0, x2: 340, y2: 1000.
The clothing item in the photo is a t-shirt, not a sweatshirt. It is black in color. The bounding box for the t-shirt is: type: t-shirt, color: black, box: x1: 330, y1: 210, x2: 700, y2: 967.
The clothing in the photo is a sweatshirt. The color of the sweatshirt on the table appears to be a shade of brown or taupe, which is often referred to as nude in fashion terminology due to its ability to blend with skin color. The bounding box for the sweatshirt on the table is: type: sweatshirt, color: brown/nude, box: x1: 28, y1: 256, x2: 706, y2: 1000.
The clothing item in the photo is a t-shirt. It is in a nude color, which is a light beige or flesh tone. The bounding box for this t-shirt is as follows: x1=0, y1=0, x2=366, y2=1000.
```

## Числовые результаты

В таблице ниже представлены метрики

Наименование	Размер	Классификация цвета	Классификация типа	Детектирование	m
<a href="#">VILT-B32-FINETUNED-VQA</a>	117.58M	0.00	0.00	-	-
<a href="#">BLIP-VQA-CAPFILT-LARGE</a>	384.67M	0.00	0.00	-	-
<a href="#">FLORENCE-2-LARGE</a>	822.69M	-	-	<b>0.77</b>	-
<a href="#">PALIGEMMA</a>	2.92B	0.53	0.47	0.25	-
<a href="#">INTERNVL-2</a>	8.07B	<b>0.82</b>	0.49	0.3	-
<a href="#">MINICPM-LLAMA3-V-2_5</a>	8.54B	0.54	<b>0.74</b>	0.19	+

\*m – колонка, отражающая способность модели отвечать на несколько вопросов за одно обращение

## **Вывод**

В результате проведённого исследования подтверждение нашла всего одна гипотеза – VLM способна решить несколько простых задач CV за одно обращение. Однако, хорошие результаты работы некоторых моделей (выделены жирным шрифтом) показывают их потенциал применения в качестве помощников в узком классе задач, например, как помощника в разметке для задачи детектирования. Дальнейшее развитие и исследование данной тематики можно считать актуальным.

Также, стоит отметить, что в текущих реалиях рассматривать какую-либо модель данного класса в качестве решения для real-time потоковой обработки данных, которая необходима в данном проекте – невозможно. Инференс моделей занимает много времени и требователен к ресурсам.