Robust Vision-Language-Action Models for Robotic Grasping in Corner Case Scenarios

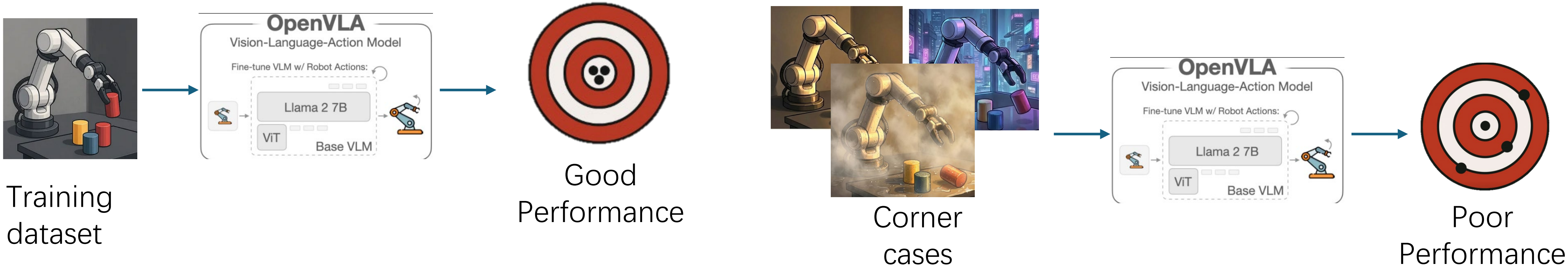
Yuan Li, Jilei Sun, Zijian He, Joseph Min-Chen

Content

- Problem: VLA suffers from corner cases -------------------3
- The proof from the real experiments ---------------------4-5
- Motivation: Use prompt tuning is the most effective way ----6
- Challenge: How to design the neural network -------------7
- Solution: Our structure ------------------------------8
- Result based on our experiments------------------------9
- Future idea for potential publication ---------------------10-11

Problem:VLA suffers from corner cases.

- We train a good VLA model in familiar environments, but it faces a lot of failures in real environment, due to corner cases.
  - including extreme lighting, changed background, mist environment, and so on.



Training dataset

Good Performance

Corner cases

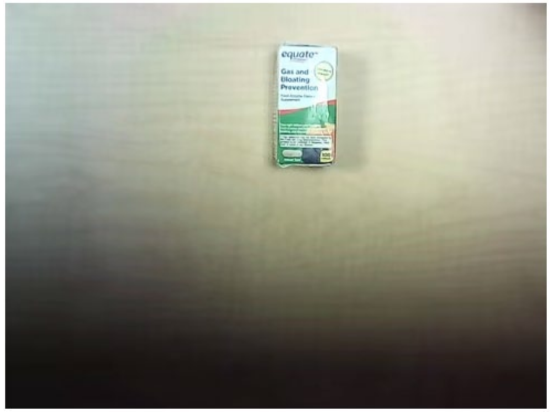
Poor Performance

OpenVLA
Vision-Language-Action Model
Fine-tune VLM w/ Robot Actions:
Llama 2 7B
ViT
Base VLM
OpenVLA
Vision-Language-Action Model
Fine-tune VLM w/ Robot Actions:
Llama 2 7B
ViT
Base VLM

How to avoid failures and improve VLA performance with these corner cases?
Lighting Changed, Background Changed, Mist due to climate.

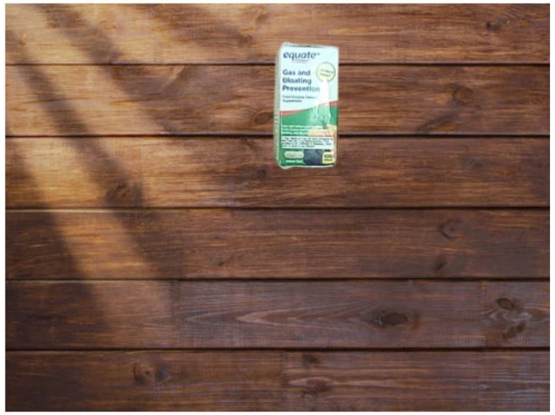
Problem based on real experiments.

- Well trained model (SmoVLA) cannot handle different corner cases.



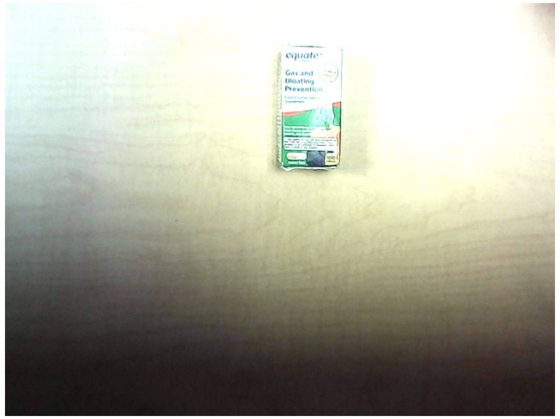
Original Image

Generated by Up camera.



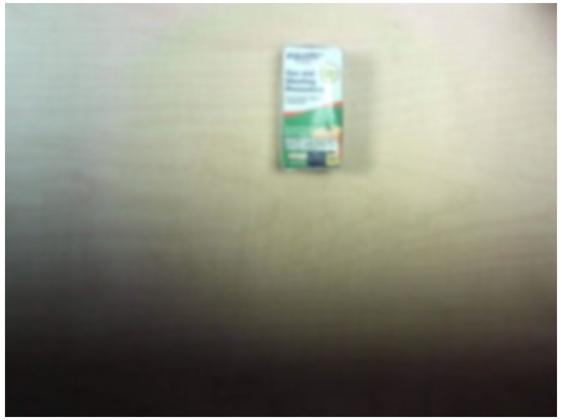
new background

Generated by RemoveBG2.0 and generate BG with stable diffusion 1.5.



Extreme Lighting

Generated by OpenCV to relighting Image



Mist effect

Generated by OpenCV to add mist effect

Original Image

Success

Error < 1

Non-Sensitive

New background

Failure

Error> 90

Sensitive

Extreme Lighting

Failure

Error> 40

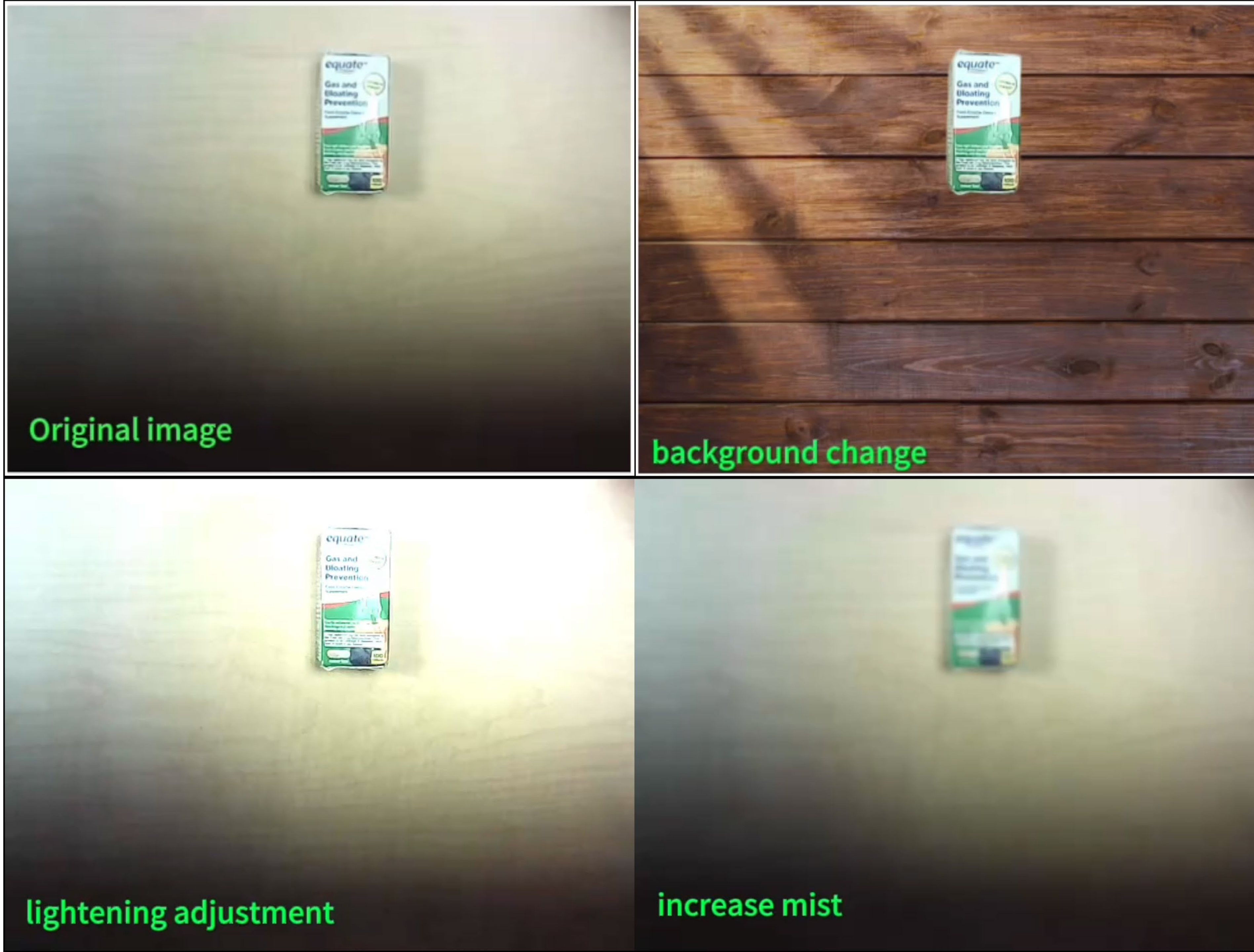
Sensitive

Mist effect

Success

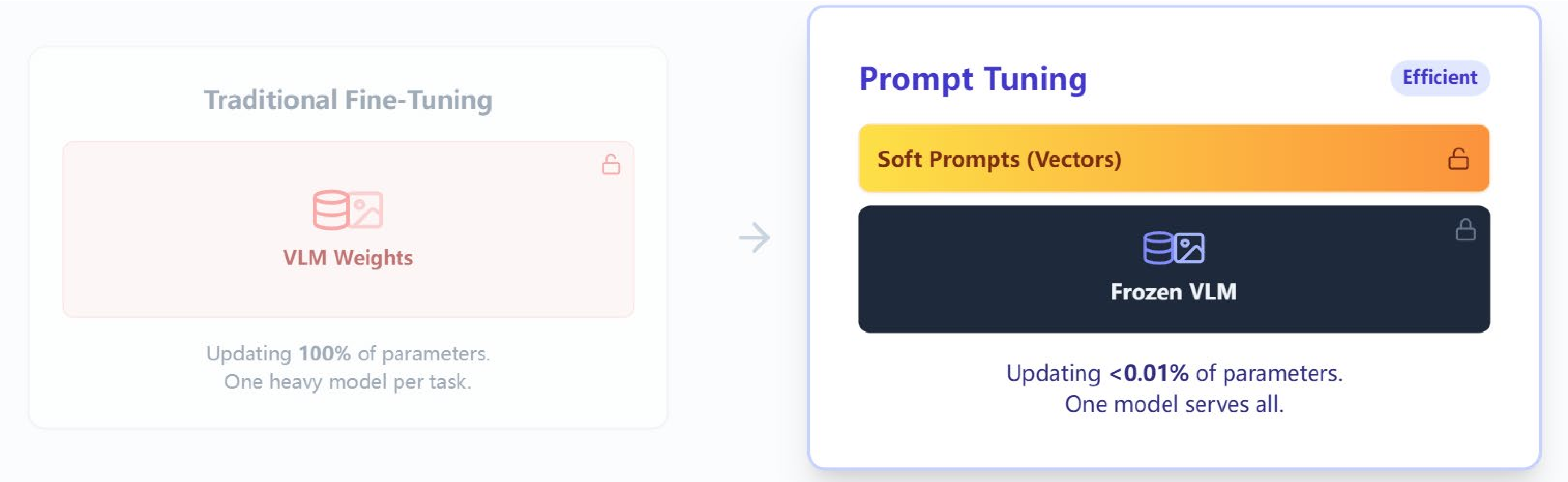
Error < 4

Non-Sensitive



Motivation: An effect way to tackle corner cases

- We want to use prompt tuning to improve the performance.
- A prompt vector for one kind of corner cases.



Traditional Fine-Tuning

VLM Weights

Updating **100%** of parameters.
One heavy model per task.

Prompt Tuning

Efficient

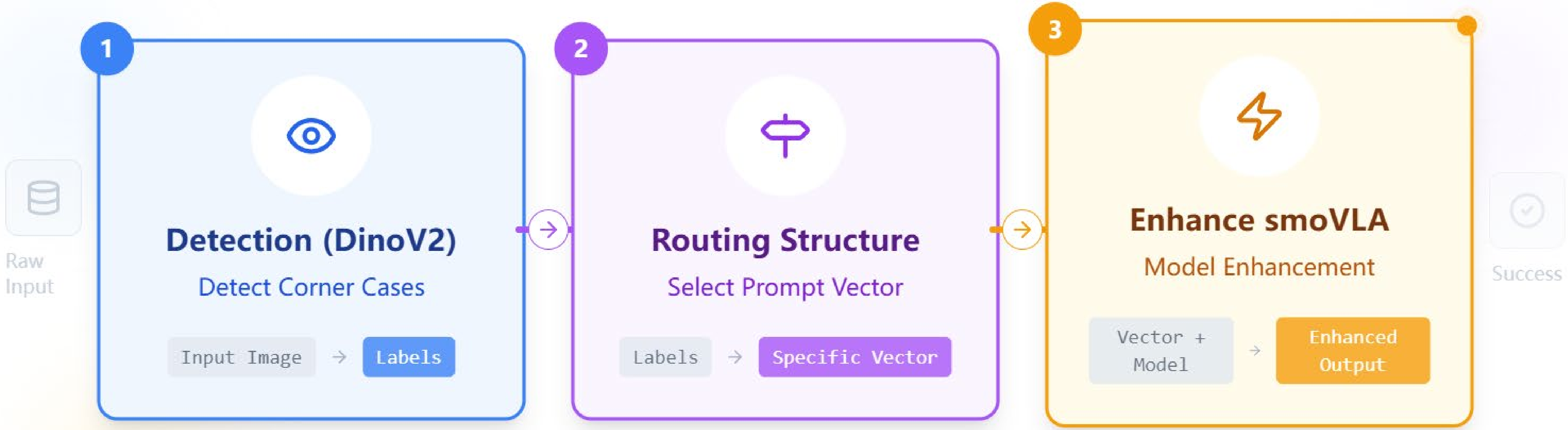
Soft Prompts (Vectors)

Frozen VLM

Updating **<0.01%** of parameters.
One model serves all.

Challenge: a new routing structure for corner cases

- How to use Prompt tuning to handle different kinds of corner cases?
  - First, we use DinoV2 to detect the corner cases' labels
  - Second, We design a routing structure to select prompt vector.
  - Third, we use the prompt vector to improve smoVLA.



Raw Input

1 Detection (DinoV2)

Detect Corner Cases

Input Image → Labels

2 Routing Structure

Select Prompt Vector

Labels → Specific Vector

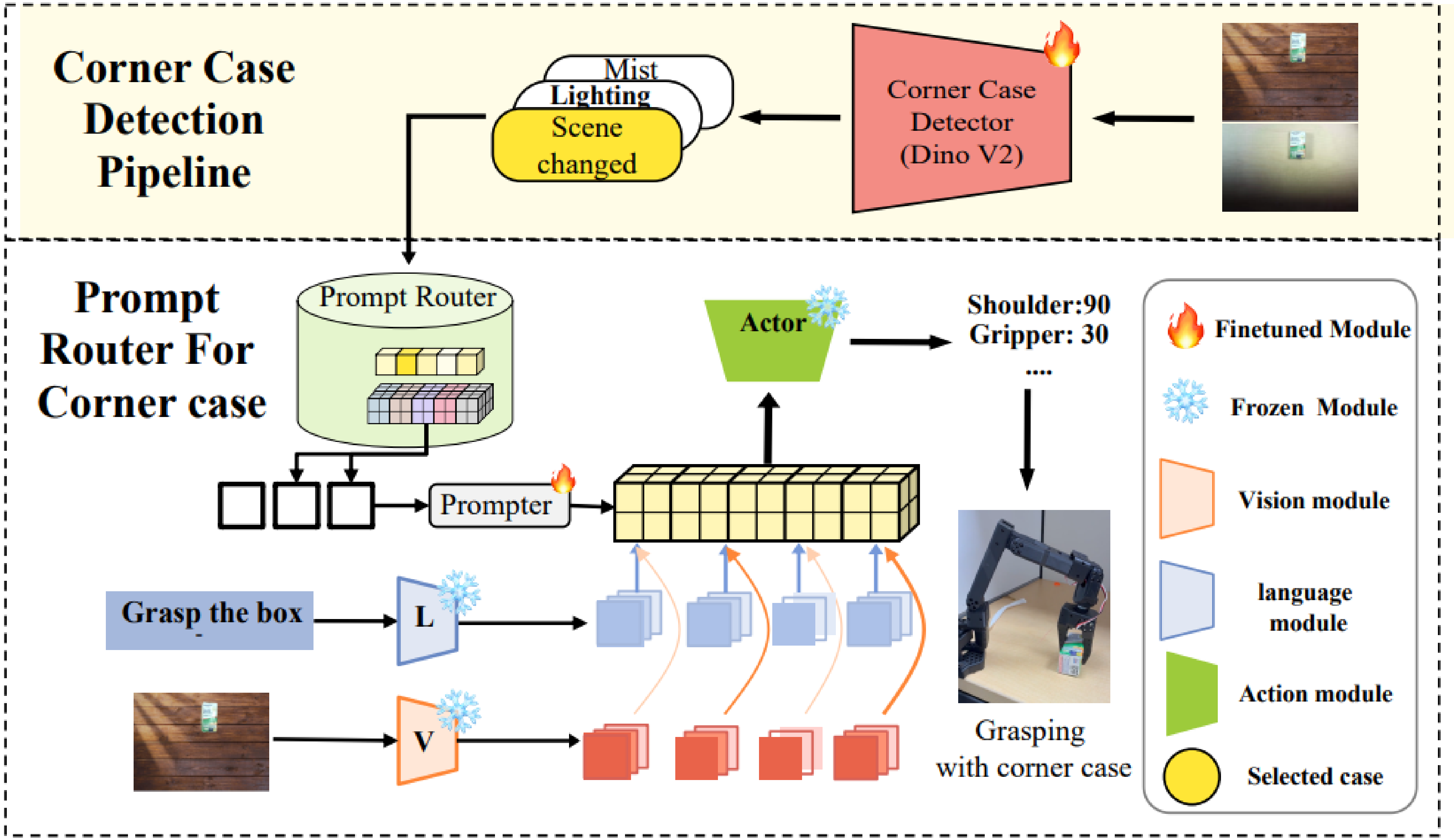
3 Enhance smoVLA

Model Enhancement

Vector + Model → Enhanced Output

Success

Solution: A new routing structure for handling corner cases



Result: We achieve huge success.

Original Image

Success

Success rate
=100%
in ten tests

New background

Success

Success rate
=80%
in ten tests

Extreme Lighting

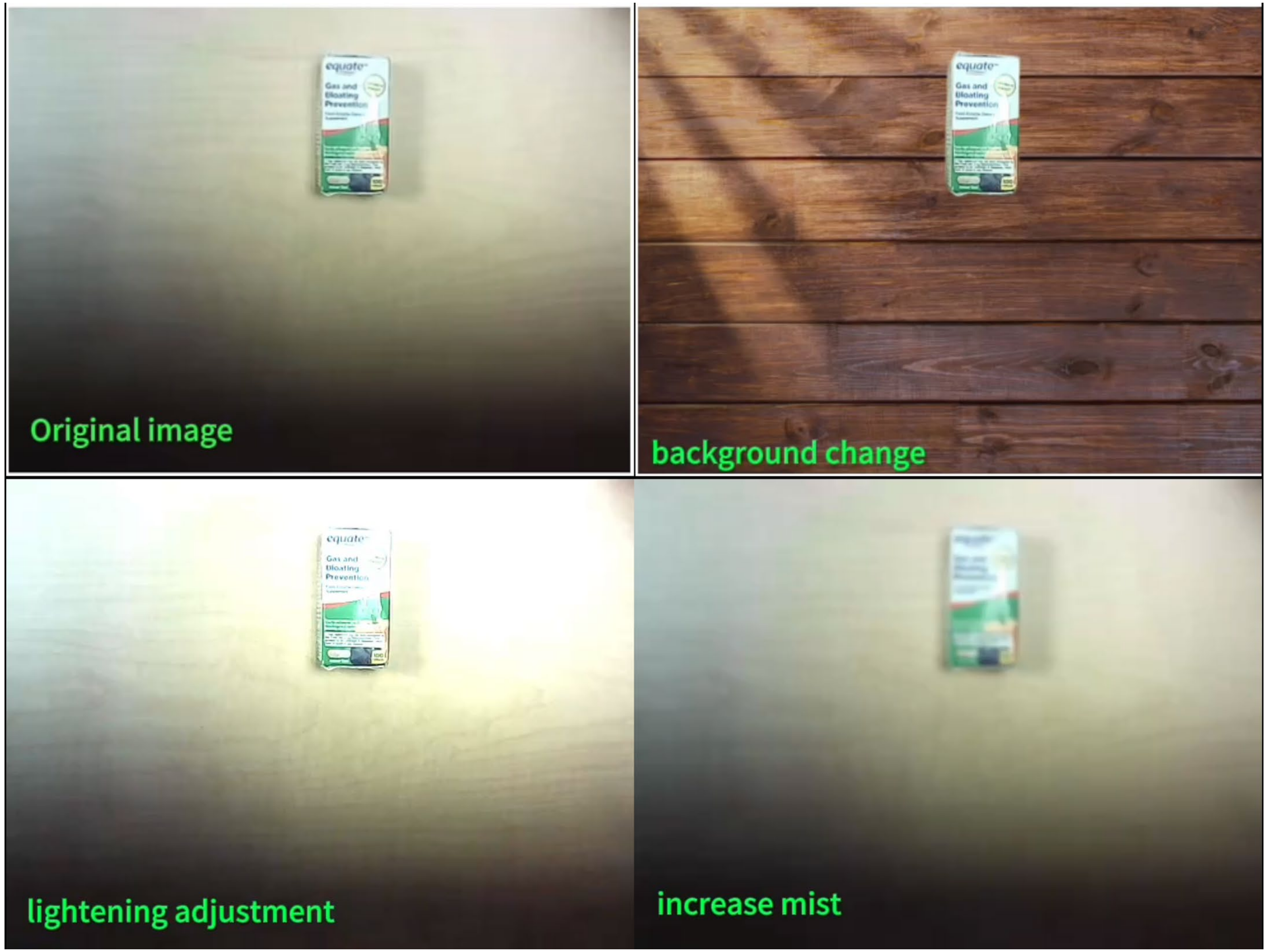
Success

Success rate
=90%
in ten tests

Mist effect

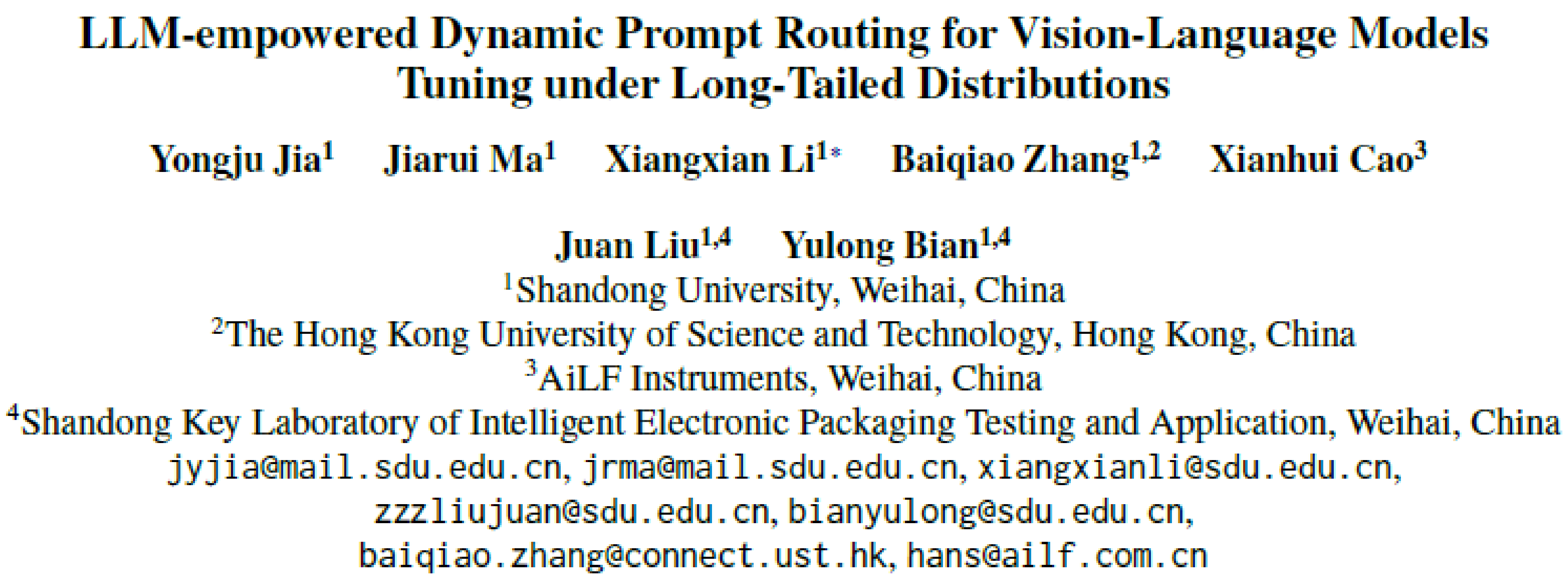
Success

Success rate
=100%
in ten tests



Unfortunately, the idea was published in EMNLP 2025

- In September.



LLM-empowered Dynamic Prompt Routing for Vision-Language Models Tuning under Long-Tailed Distributions

Yongju Jia[1] Jiarui Ma[1] Xiangxian Li[1]* Baiqiao Zhang[1,2] Xianhui Cao[3]

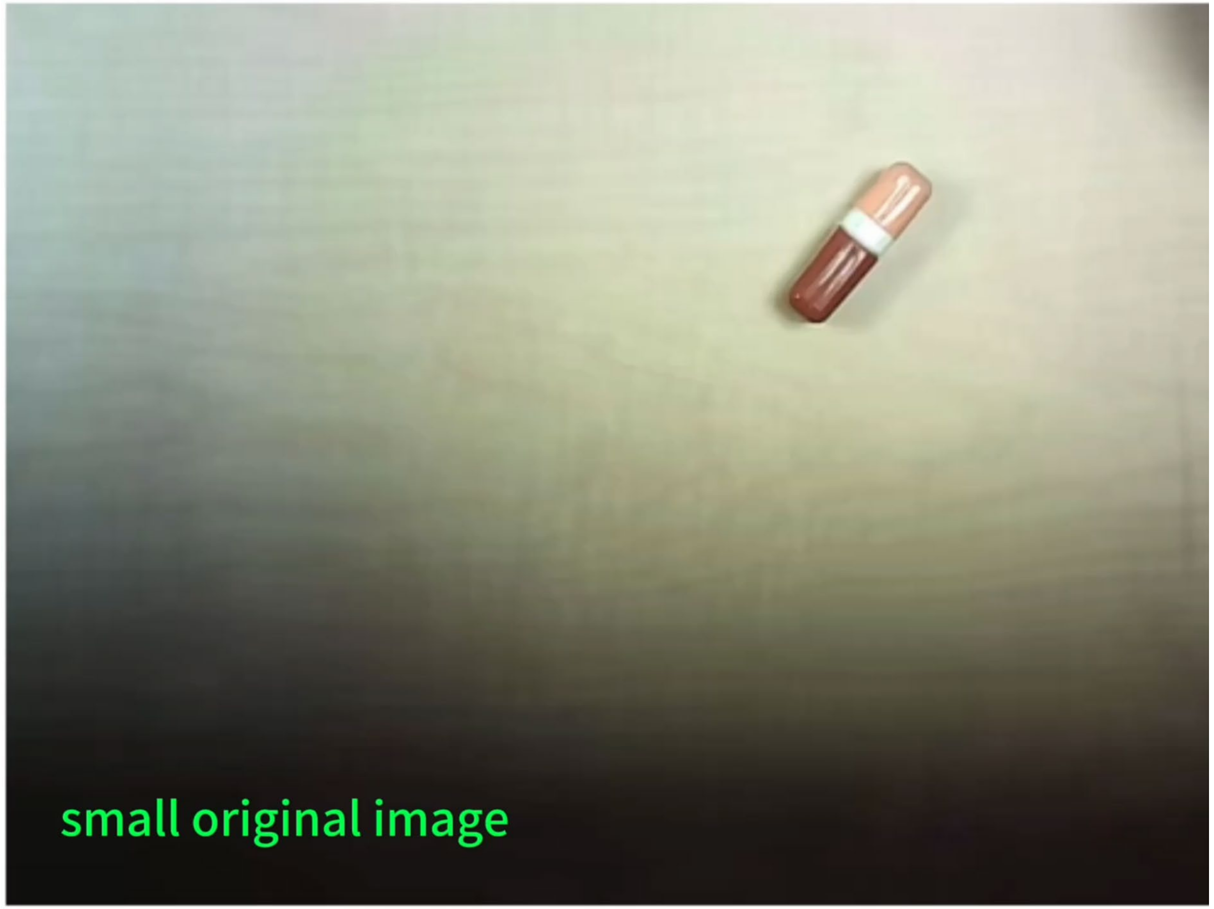
Juan Liu[1,4] Yulong Bian[1,4]

[1]Shandong University, Weihai, China
[2]The Hong Kong University of Science and Technology, Hong Kong, China
[3]AiLF Instruments, Weihai, China
[4]Shandong Key Laboratory of Intelligent Electronic Packaging Testing and Application, Weihai, China
jyjia@mail.sdu.edu.cn, jrma@mail.sdu.edu.cn, xiangxianli@sdu.edu.cn,
zzzliujuan@sdu.edu.cn, bianyulong@sdu.edu.cn,
baiqiao.zhang@connect.ust.hk, hans@ailf.com.cn

But we already find a new idea:

- Key Insight:
  - We found out: different corner case has different difficulties.
  - The changed background is the most difficult, the mist is the least difficult.
  - But the previous papers ignore this:
    - They use the same K prompt vectors to finetune each kind of corner cases.
    - We believe 1 prompt vector is enough for mist, and 10 prompt vectors might not be enough for images with changed background.
    - So how to use prompt pool + graphcut pooling to tackle the problem.
    - Already talked with Qifan.
    - If you are interested, please contact us.

Some other results (small objects/baseline)



Some other results (small objects/tuned)