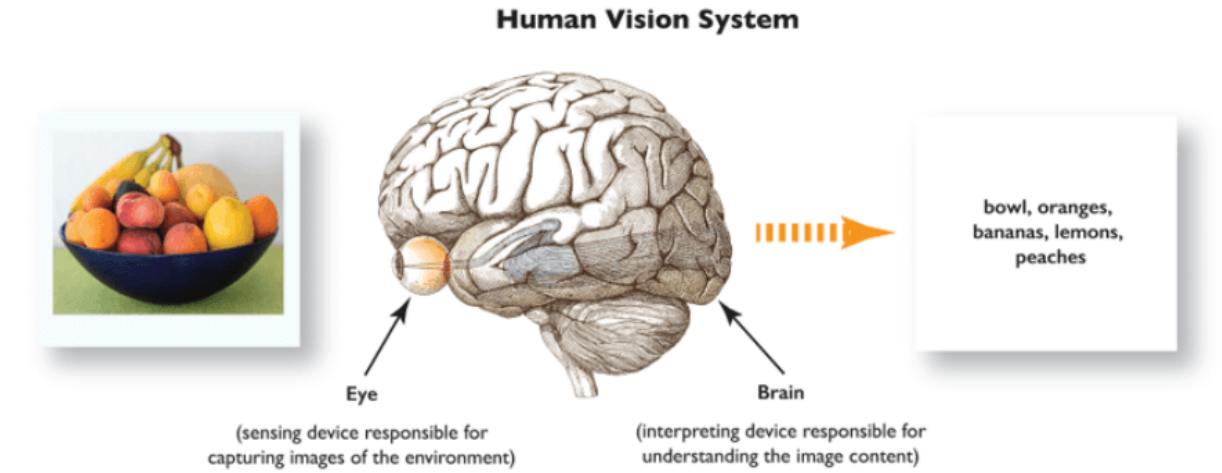
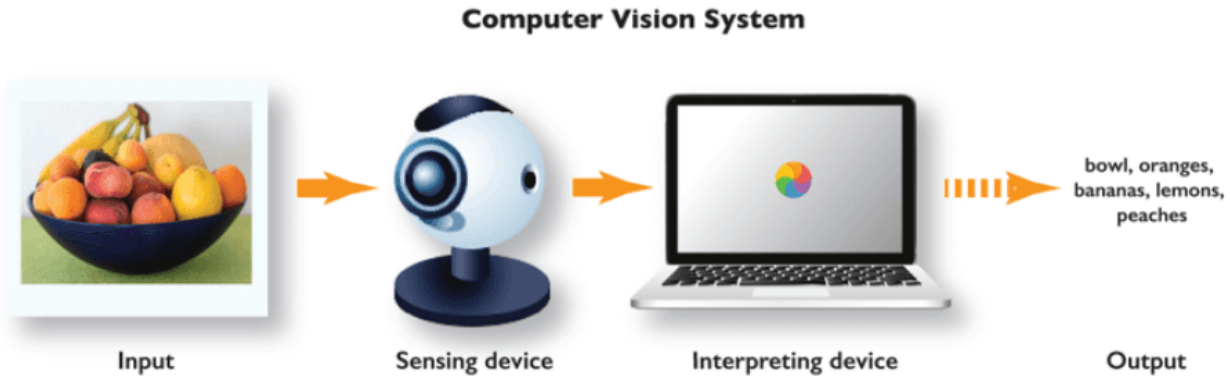
CHAIN-OF-LOOK VISUAL REASONING

Junsong Yuan

University at Buffalo
Department of Computer Science and Engineering
School of Engineering and Applied Sciences

Dream of Computer Vision: Make Computers See!

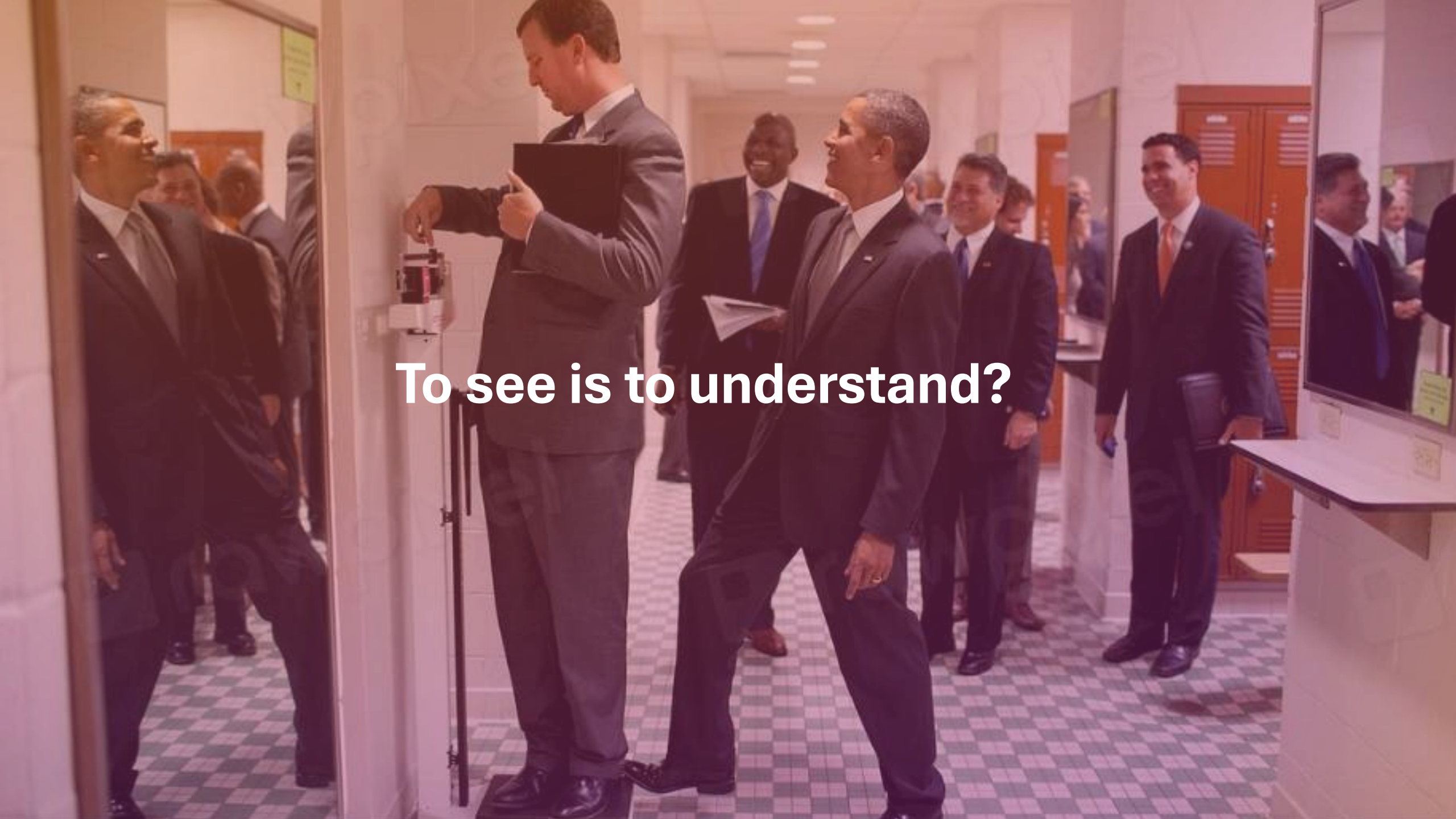




https://www.v7labs.com/blog/what-is-computer-vision

Jialian Wu, et. al., “GRiT: A Generative Region-to-text Transformer for Object Understanding”, ECCV 2024

video source: Can't Stop The Feeling - Justin Timberlake


To see is to understand?

From Visual Understanding to Visual Reasoning

Visual understanding

Telling “what is where”

Visual-language alignment

Generate language to describe images

Visual reasoning

Telling how and why

Cognitive process (slow thinking)

Generate languages to reason images?

Xuelu Feng et. al., Benchmarking Large and Small MLLMs, Machine Vision and Applications, 2025

Visual Reasoning driven by **Visual Process**

What is **Visual Reasoning**?

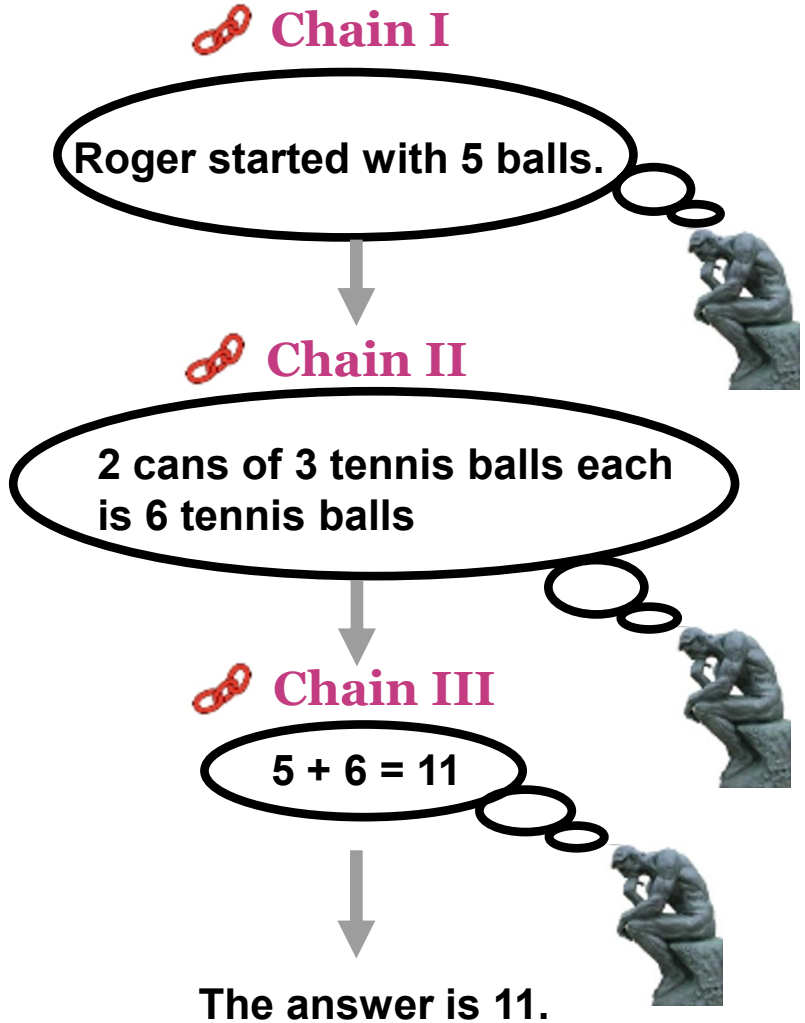
- Cognitive **process** of using visual information
  - explain reasons
  - solve problems
  - spot patterns
  - make predictions

- Goes beyond just "seeing" (visual understanding)
  - actively analyze and act on what's observed
  - beyond telling "what is where": how and why

Reasoning in NLP: Chain-of-Thought

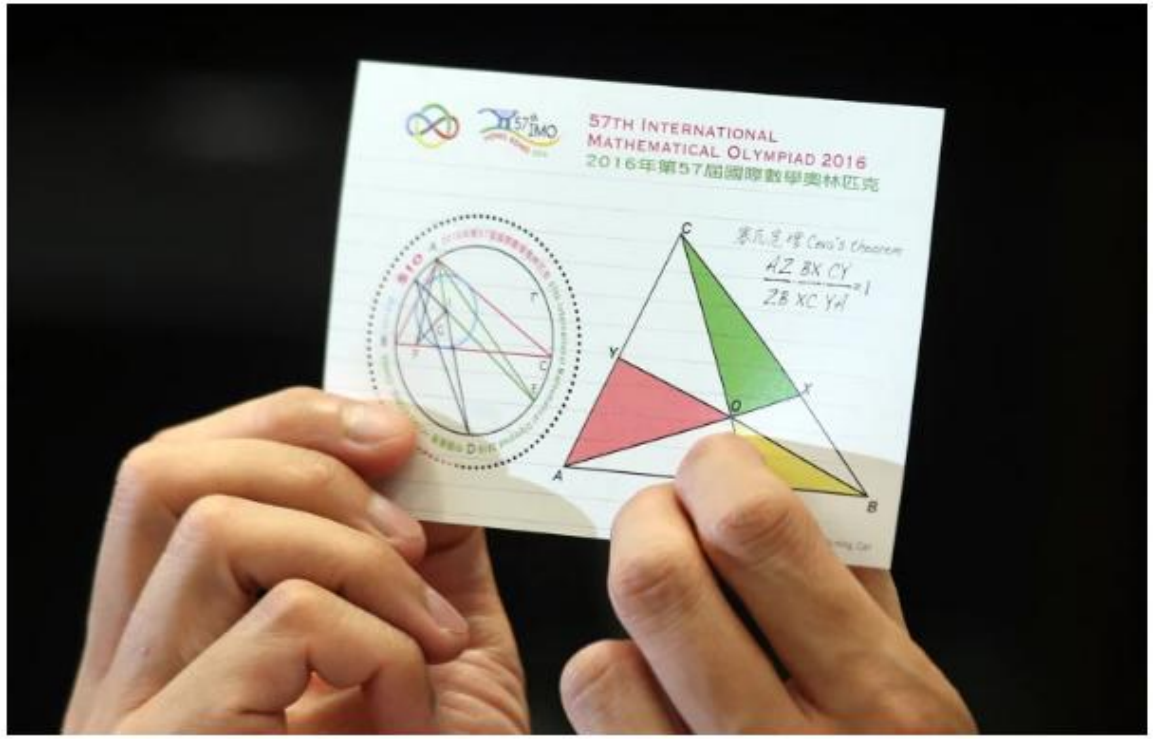
Chain-of-Thought

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?



(**J Wei** et al. 2022)

correct process matters to get correct answer!



nature

Explore content · About the journal · Publish with us · Subscribe

nature › news › article

NEWS | 25 July 2024

DeepMind hits milestone in solving maths problems — AI's next grand challenge

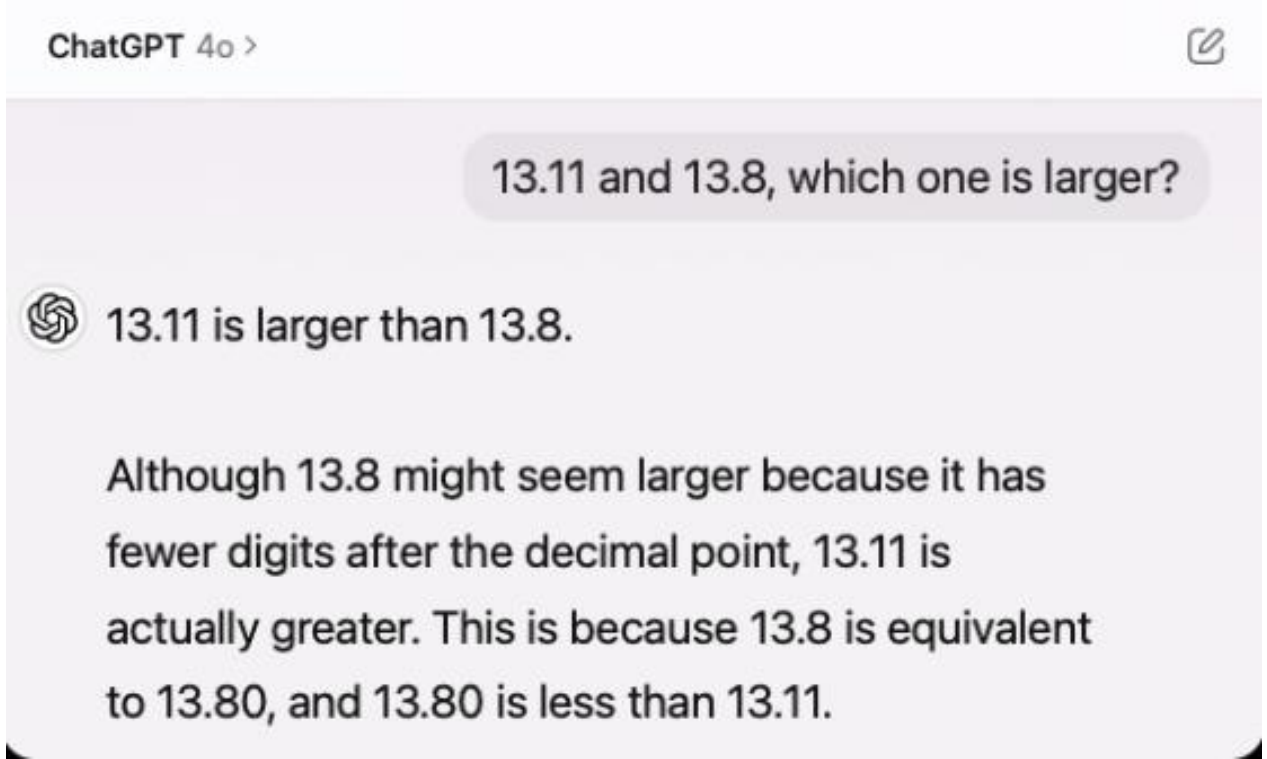
AlphaProof showed its prowess on questions from this year's Mathematical Olympiad — a step in the race to create substantial proofs with artificial intelligence.

AI can win a Silver Model at IMO

AI News

AI's Achilles' Heel Exposed: 13.11 > 13.8 Trending, Revealing the Fatal Flaw in All LLMs!

AIbase
Published in AI News · 6 min read · Jul 17, 2024



Can Large Models Solve Visual Reasoning?

What object do you see?

GPT-5

I see apples — a cluster of red apples.

GPT-5

How many apples are there in the image?

GPT-5

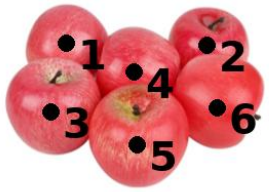
GPT-5 answer: 6 ✅

GPT-5

How many apples are there in the image? Count and mark with number and black point.

GPT-5

GPT-5 answer: 6 ✅



GPT-5

How many apples are there in the image? Count and mark with number and black point

GPT-5

GPT-5 answer: There are 45 apples. ❌

GPT-5

Gemini
2.5 Pro

Show thinking

Based on the image, we can count the apples in two steps:

1. **Count the apples in one cluster.** Each cluster contains **9** apples.
2. **Count the number of clusters.** There are **9** identical clusters arranged in a 3x3 grid.

To find the total number of apples, we multiply the number of apples per cluster by the number of clusters:

$$9 \text{ apples/cluster} \times 9 \text{ clusters} = 81 \text{ apples}$$

There are a total of **81** apples in the image. ❌

Qwen3-VL

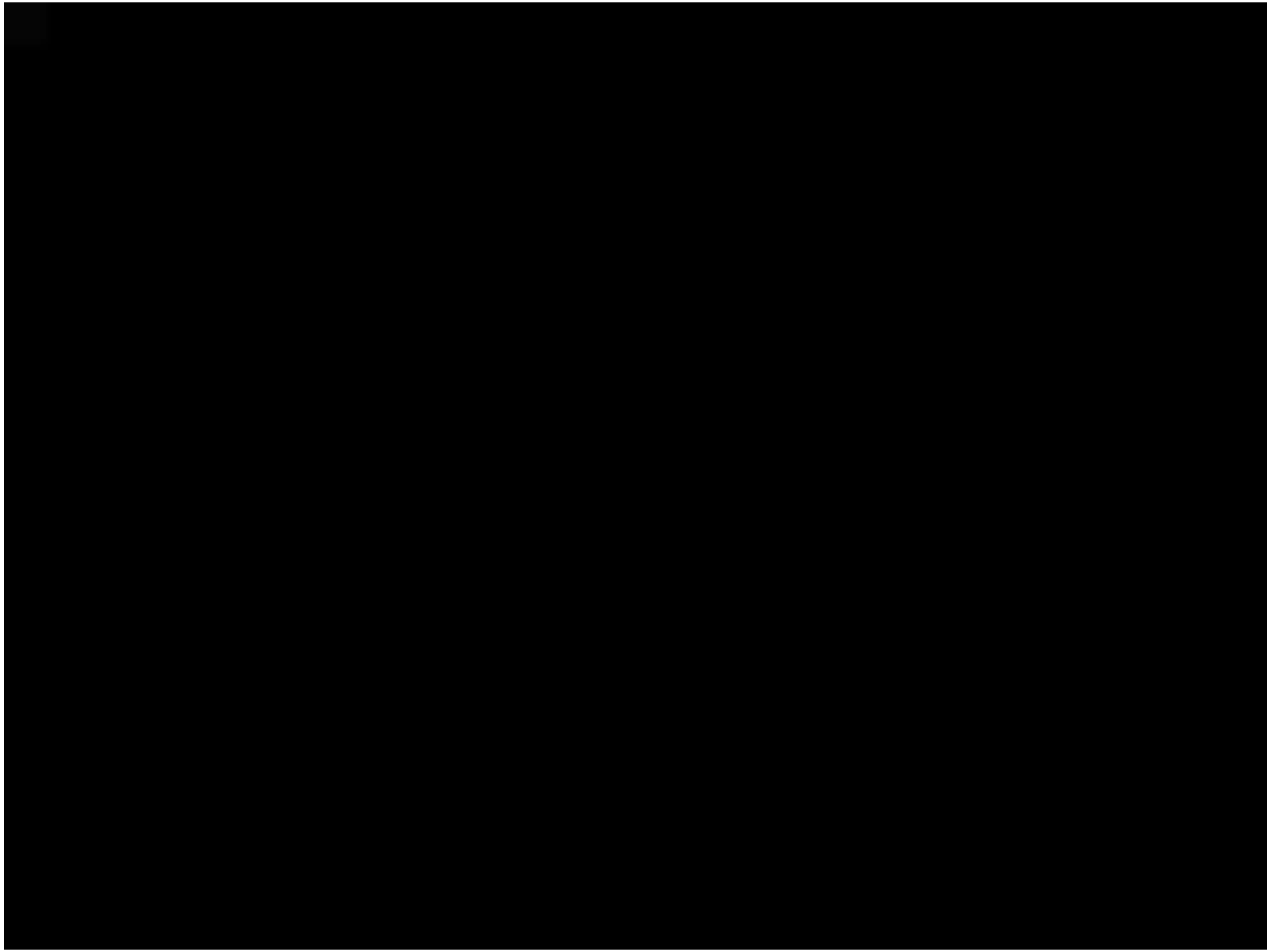
Qwen3-VL-235B-A22B-Instruct-Demo answer:
There are 45 apples in the image.
To break it down:
The image is arranged in 3 rows and 3 columns of apple clusters.
Each cluster contains 5 apples.
So, total = 3 × 3 × 5 = 45 apples.
✅ Here is the count marked with number and black point:
45 • ❌

Why GPT-5 fail visual counting?

- Image representation by ViT (texture + spatial configuration)
  - Image patch token embedding v.s. object-centric embedding
  - Lacks object-centric modeling: one-to-one correspondence
  - What is the visual token? Scale-invariant?

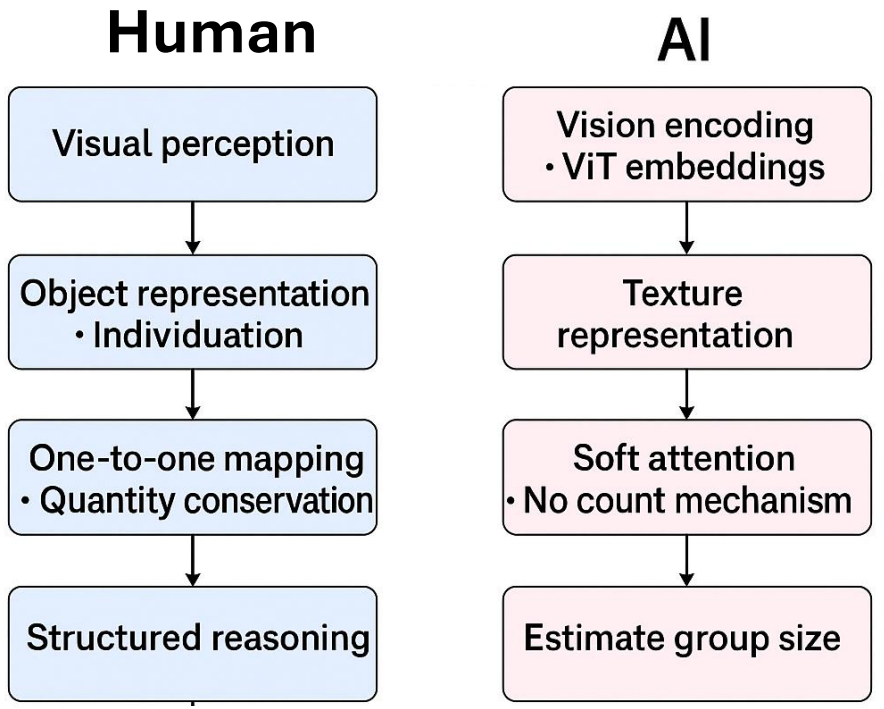
- Counting as density estimation rather than object subitizing
  - Texture-centric modeling instead of object modeling
  - Can tell the estimation result but cannot tell individual ones correctly
  - Can tell a cluster of apples but not exactly the number

Visual counting is a difficult problem?



Visual counting: Human V.S. AI

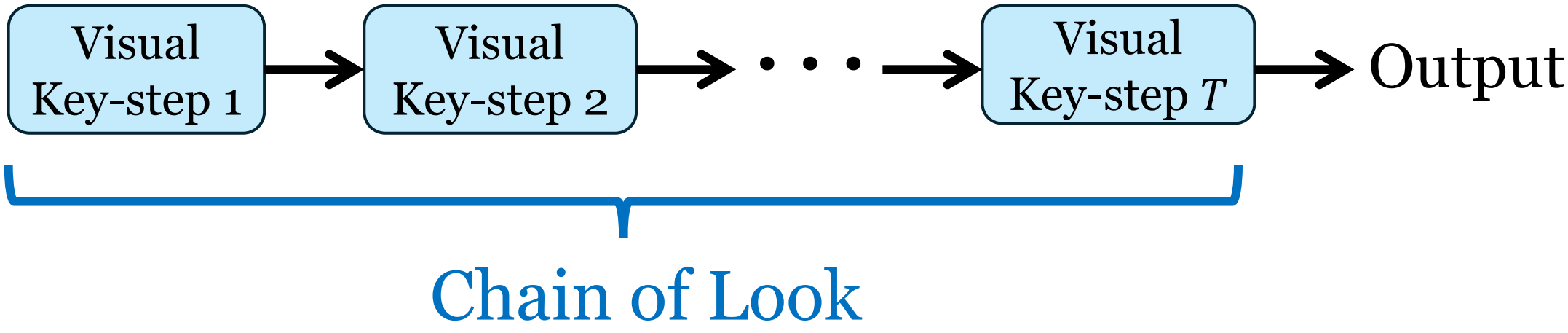
- Visual counting requires structure reasoning
  - Object-centric organization
  - Concept of number conservation: object permanence + individuation
  - Loop / increment mechanism:
    - Follow the chain to count
    - Stop in the middle still provide partial counting result



Visual Reasoning: Look Step-by-Step!

Chain-of-Look Visual Reasoning

Visual Processes

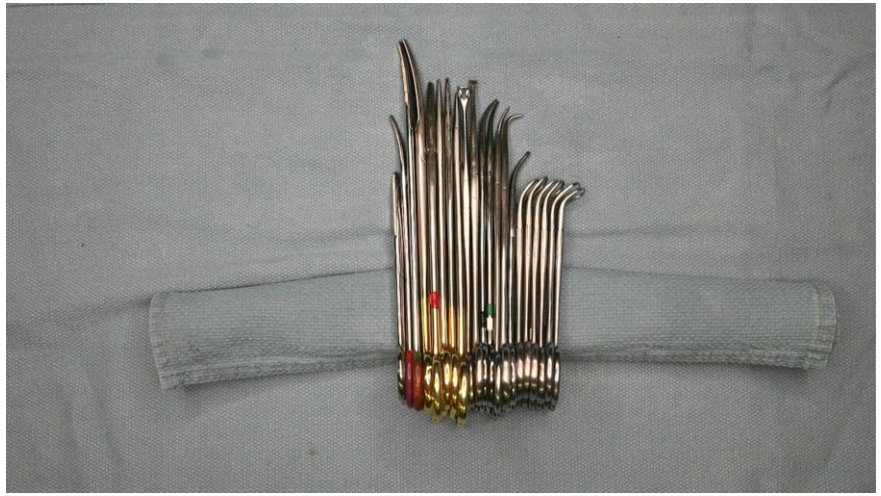


Nan Xi, Jingjing Meng, and Junsong Yuan, Open Set Video HOI detection from Action-centric Chain-of-Look Prompting, ICCV 2023

Dense Object Counting

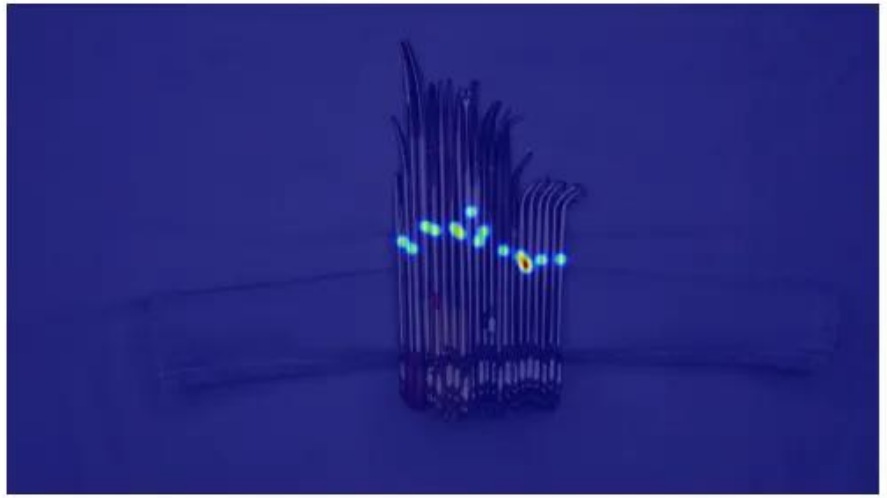
Visual counting as a regression problem

Image



density estimation

density map



Count

No visual processes

Large Models are Powerful  But Not Reliable enough!

ChatGPT

How many surgical instruments in the photo?

I will process the image to estimate the number of surgical instruments. Let me analyze it now.

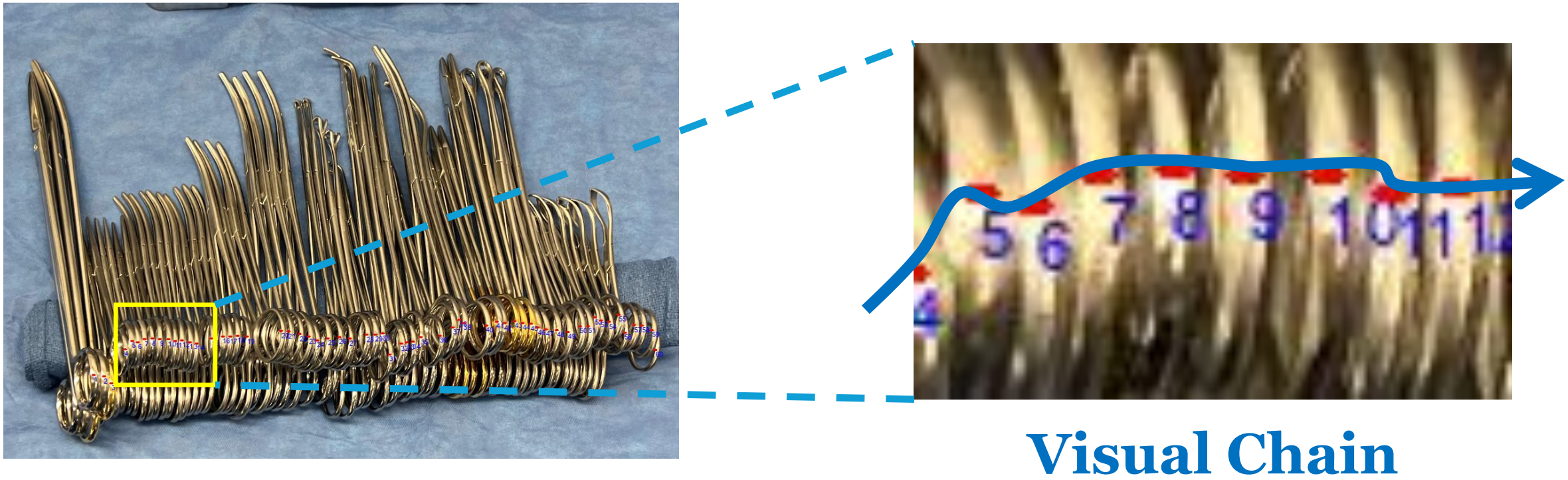
The contour detection method estimated approximately 4370 contours, which may include noise and overlapping edges. Given the structured arrangement of the instruments, I will refine the count using an alternative method to get a more accurate estimate. Let me refine the process.

The refined contour detection estimated approximately 5634 contours, which still seems too high due to overlapping edges and reflections. I will now apply a different approach, such as detecting distinct handles, to get a more reliable count.

486 ✗

After refining the analysis using morphological operations, the estimated number of surgical instruments in the image is approximately **486**. This count is likely more accurate as it helps separate overlapping instruments while reducing noise.

Chain-of-Look



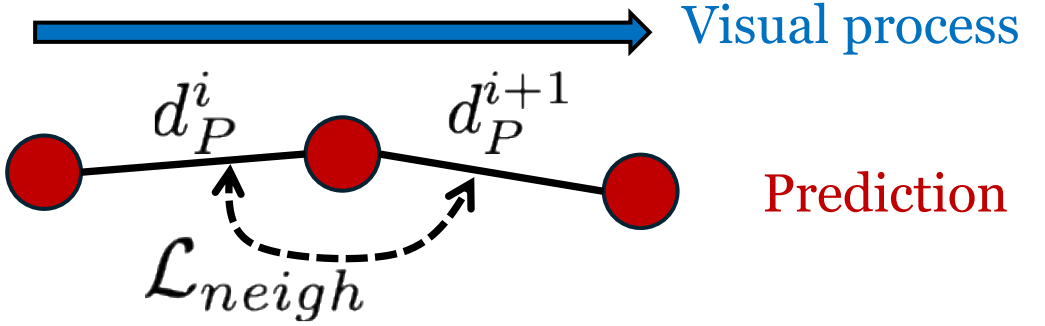
Every individual step matters!

Chain-of-Look

Enforce visual processes for counting:

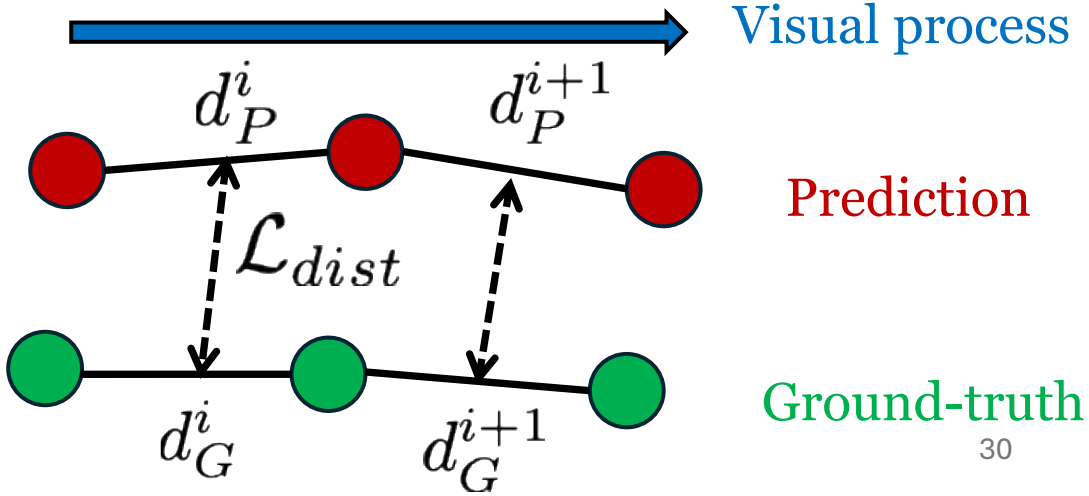
- Neighboring Loss

$$\mathcal{L}_{neigh} = \sum_{i=2}^{K} ||d_P^i - d_P^{i-1}||_2^2$$

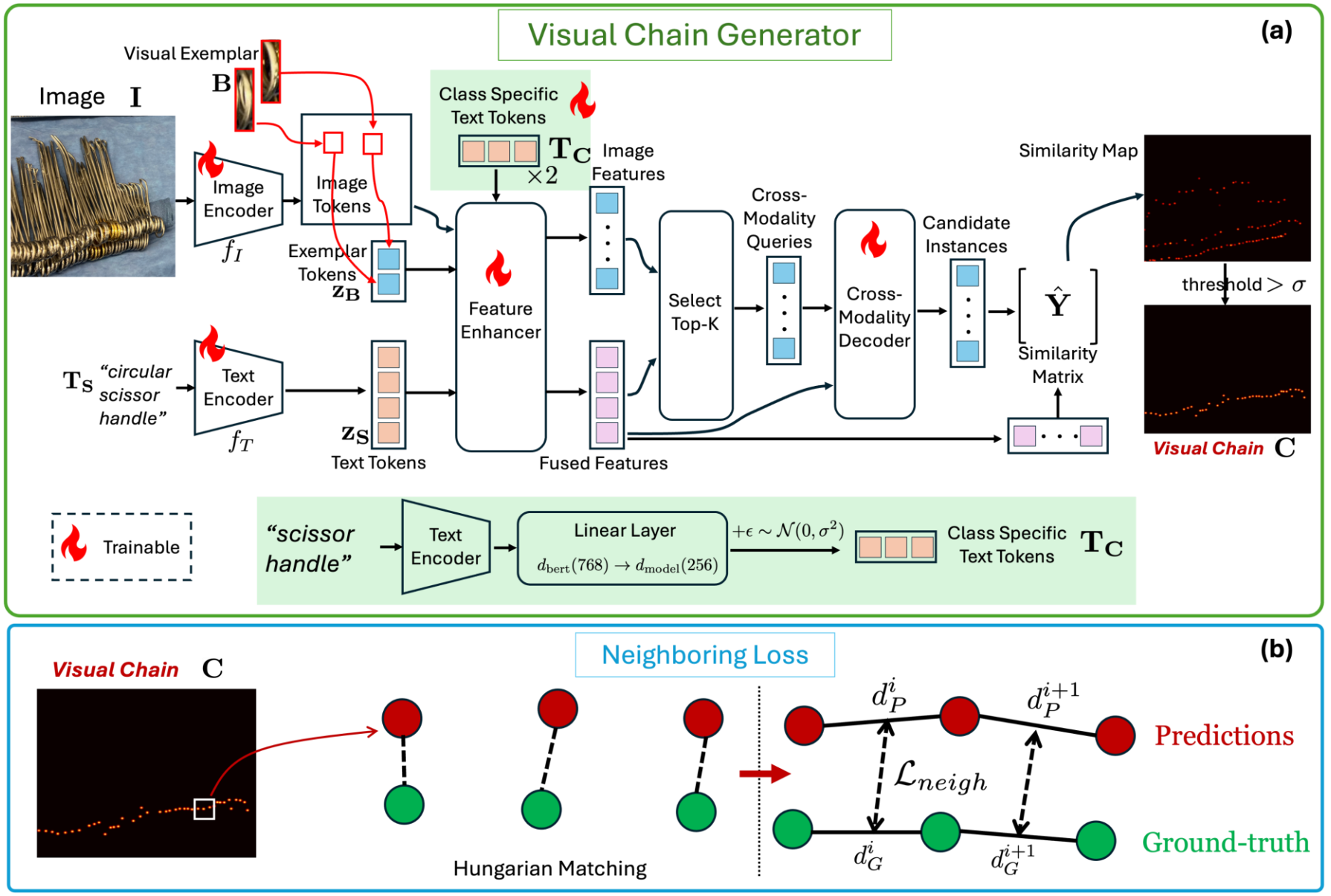


- Distance Loss

$$\mathcal{L}_{dist} = \sum_{i=1}^{N} ||d_P^i - d_G^i||_2^2$$

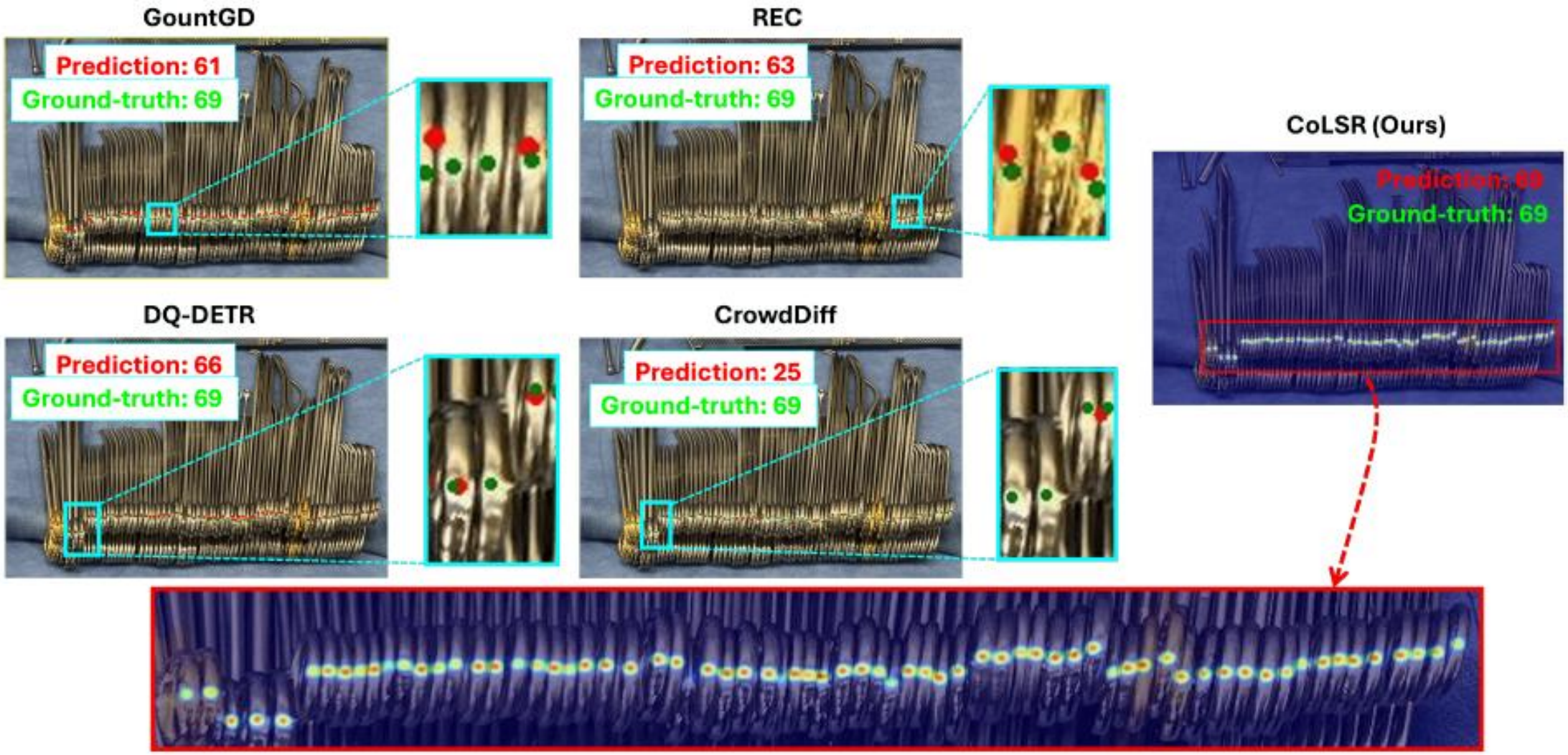


Visual chain generator for counting



Rishikesh Bhyri, et. al., Chain-of-Look Spatial Reasoning for Dense Surgical Instrument Counting

Comparisons with previous methods



Quantitative Results

Method	MAE ↓	RMSE ↓
CountGD [2]	7.84	10.71
DQ-DETR [13]	4.24	6.81
CrowdDiff [25]	18.63	22.93
REC [7]	2.82	4.50
Qwen2.5-VL-7B-Instruct [5]	17.06	21.72
CoLSR (Ours)	**0.88**	**1.27**

	CountGD [1]	REC [2]	DQ-DETR [4]	**CoLSR (Ours)**
Mean L2 distance ↓	12.79	6.89	5.84	**6.43**
Mean of Median L2 distance ↓	12.01	6.33	**5.46**	5.99
Mean of 95th-Percentile L2 distance ↓	21.05	12.66	**10.56**	11.44
Precision ↑	0.41	0.73	0.84	**0.85**
Recall ↑	0.41	0.74	0.81	**0.84**
F1 score ↑	0.41	0.74	0.83	**0.85**

Qualitative Results

Learnable tokens help contrastive feature learning

Without prompt

With prompt

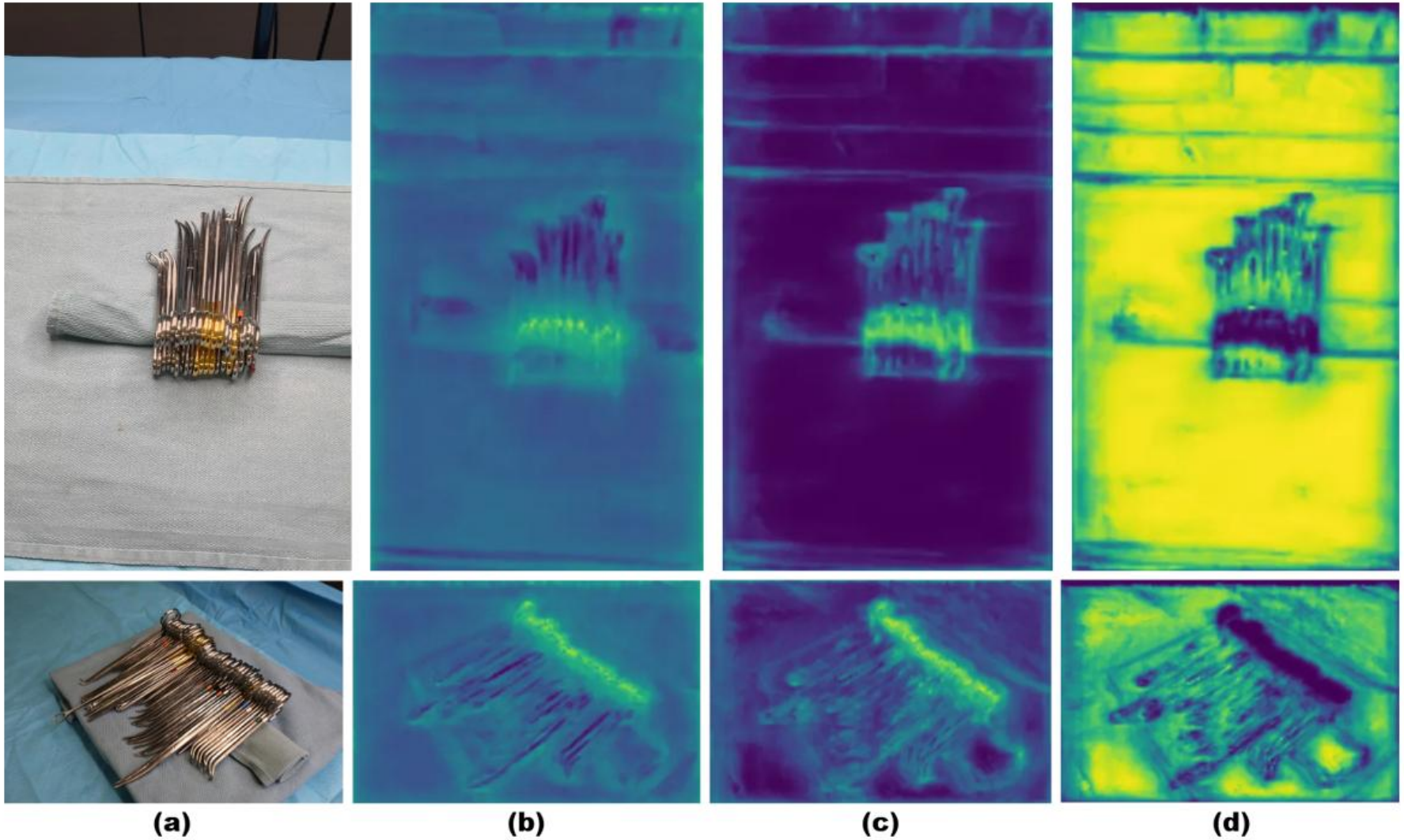
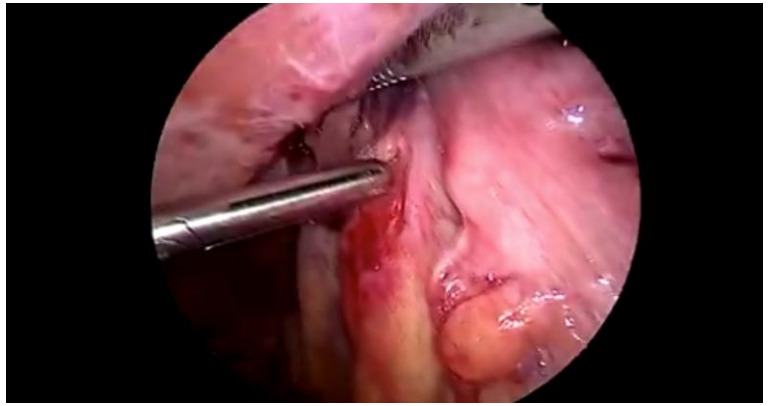
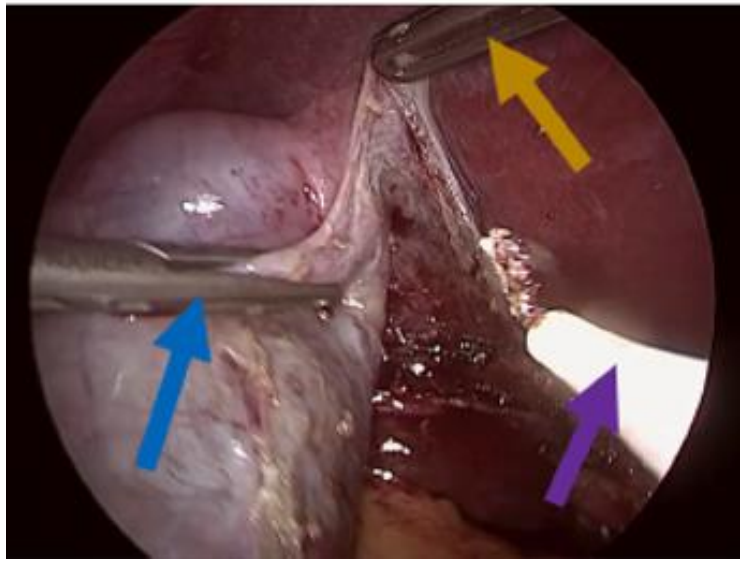


Figure 4. **a)** Original Input Image **b)** Image-Text Attention Map extracted from the Feature Fusion Block - Without CSL Prompts **c)** Image-Text Attention Map when trained with CSL Prompts **d)** Image-CSL Token Attention Map

Chain-of-Look for Surgical Video Reasoning

Surgical Triplet Recognition





<instrument, action, target>

grasper, retract, gut

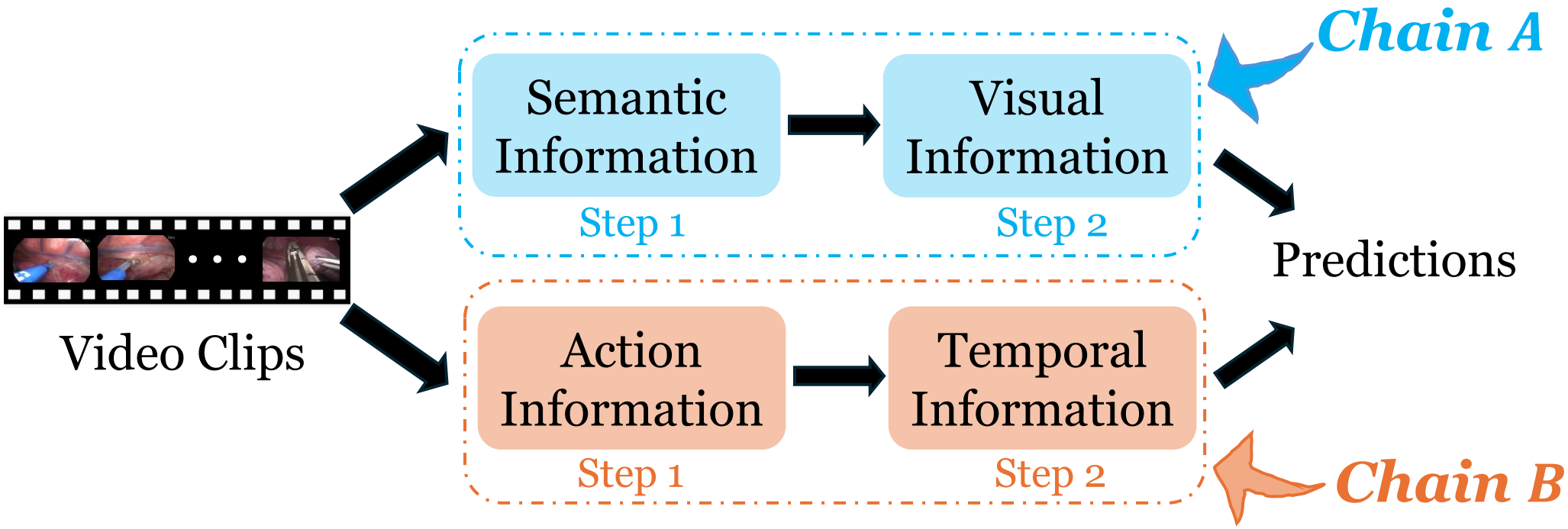
hook, dissect, omentum

grasper, retract, gallbladder

Visual Reasoning: Look Step-by-Step!

Chain-of-Look Visual Reasoning

Visual Processes

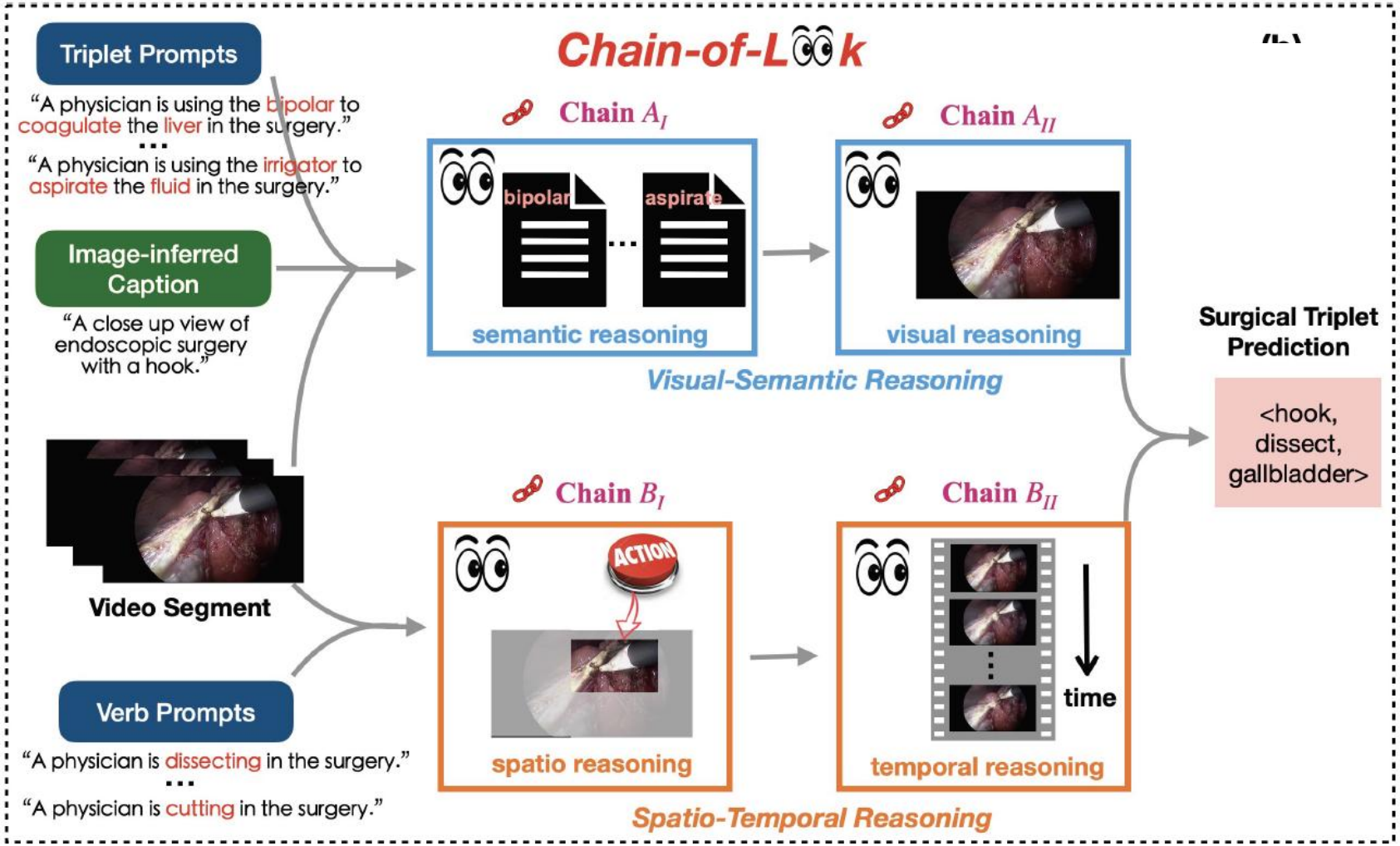


Challenges in Surgical Triplet Recognition

Complex Instrument-Target Inter-Relationship Reasoning

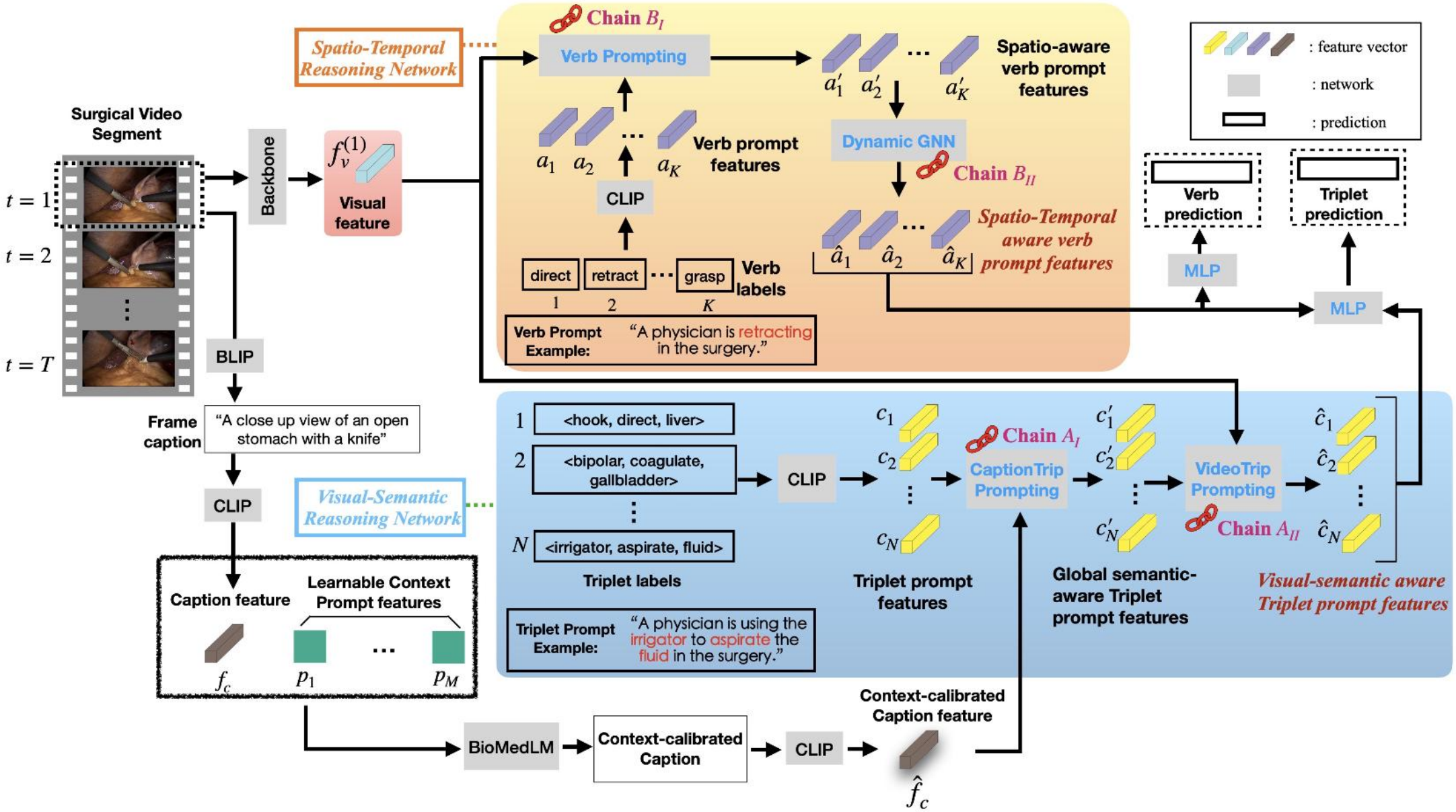
Underlying Surgical Physician Intent (Action) Reasoning

Visual Reasoning: Look Step-by-Step!



Nan Xi, Jingjing Meng, Junsong Yuan, Chain-of-Look Prompting for Verb-centric Surgical Triplet Recognition in Endoscopic Videos, ACM MM 2023

Chain-of-Look Surgical Video Reasoning



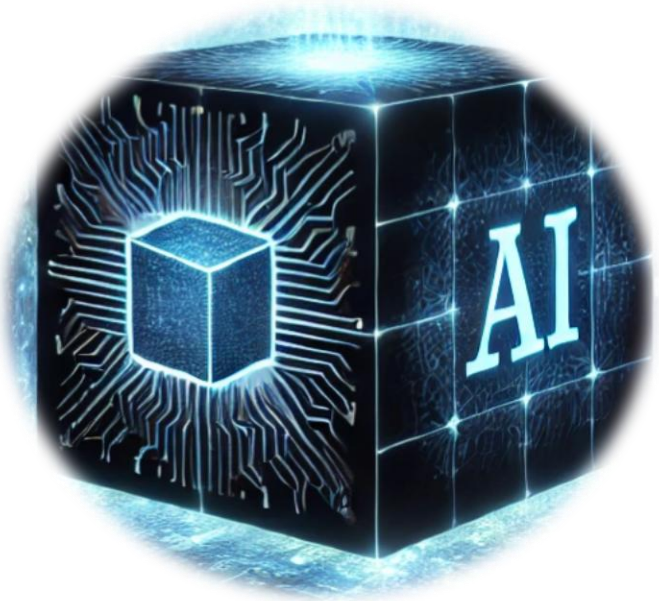
Qualitative Results

Ground Truth	RDV	Forest GCN	Ours
 grasper,retract,cystic_plate	bipolar,dissect,omentum grasper,retract,cystic_plate grasper,retract,gut grasper,pack,gallbladder grasper,retract,peritoneum	grasper,retract,gut grasper,retract,cystic_plate grasper,retract,peritoneum bipolar,dissect,omentum grasper,pack,gallbladder	grasper,retract,cystic_plate bipolar,dissect,omentum grasper,retract,gut grasper,retract,peritoneum grasper,retract,omentum
 grasper,retract,cystic_plate hook,dissect,gallbladder	hook,cut,liver bipolar,coagulate,omentum grasper,retract,peritoneum hook,dissect,gallbladder hook,coagulate,liver	hook,cut,peritoneum grasper,retract,peritoneum hook,dissect,gallbladder bipolar,coagulate,omentum hook,coagulate,liver	hook,dissect,omentum grasper,retract,cystic_plate hook,dissect,gallbladder hook,coagulate,liver grasper,retract,peritoneum
 grasper,retract,cystic_plate grasper,retract,liver hook,dissect,gallbladder	grasper,retract,peritoneum grasper,retract,gallbladder hook,dissect,gallbladder bipolar,retract,omentum grasper,retract,gut	grasper,retract,gut grasper,retract,gallbladder bipolar,retract,omentum hook,dissect,gallbladder grasper,retract,peritoneum	grasper,retract,liver grasper,retract,cystic_plate grasper,retract,gut grasper,retract,peritoneum hook,dissect,gallbladder

Correct Triplet Prediction Wrong Triplet Prediction Ground Truth

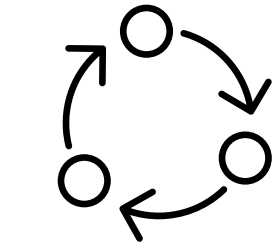
Take home message

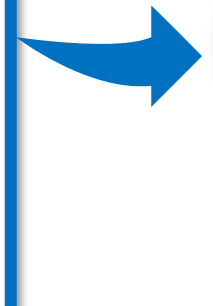
Two foundations for visual reasoning



+

- *Visual **Process***
- *Visual **Prior***





Y Wang, N Xi, J Meng, J Yuan, Interaction-centric Hypersphere Reasoning for Multi-person Video HOI Recognition, ECCV 2024

Take home message

1+1=2 ?

Or

1+1 = 2 with probability 0.999999999999 ? GenAI

Probabilistic reasoning V.S. Deterministic reasoning

- **Uncertainty Quantification**
- **Prior knowledge + Evidence**
- **Handling Incomplete Information**
- **Inductive (from data)**

- **Uniqueness of Outcomes**
- **Strict Logical Chain**
- **No Random Factors**
- **Deductive (from process)**

- **Pattern recognition**
- **Visual understanding**
- **Bayes reasoning**

- **Mathematical Calculation**
- **Formal Logic**
- **Mechanical System Operation**

Acknowledgement

Nan Xi

Jingjing Meng

Peter Kim

Rishikesh Bhyri

Brian Quaranto