NLP Reading Group
Dec 3, 2025

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Chain-of-Visual-Thought:
Teaching VLMs to See and Think Better with Continuous Visual Tokens

Authors: Yiming Qin[1], Bomin Wei[2], Jiaxin Ge[1], Konstantinos Kallidromitis[3], Stephanie Fu[1], Trevor Darrell[1], Xudong Wang[1]

[1]UC Berkeley, [2]UCLA, [3]Panasonic AI Research

Presented by Sriram Sai Ganesh

Sections

1. **Introduction**
2. **Background**
3. **CoVT**
4. **Results**
5. **Ablations**

Introduction

How many clouds within the image?

Intuition: *reasoning over dense spatial features is bottlenecked by discrete tokens.*
Task: ***Visual Question Answering (VQA)*** –
Given an image+text question, provide an answer.

Predominant approaches:

- **Late fusion:** Aggregate features from multiple modality-specific experts (ViT + LM)
- **Early fusion:** Train a VLM with an extended vocabulary (modality-specific discrete tokens)

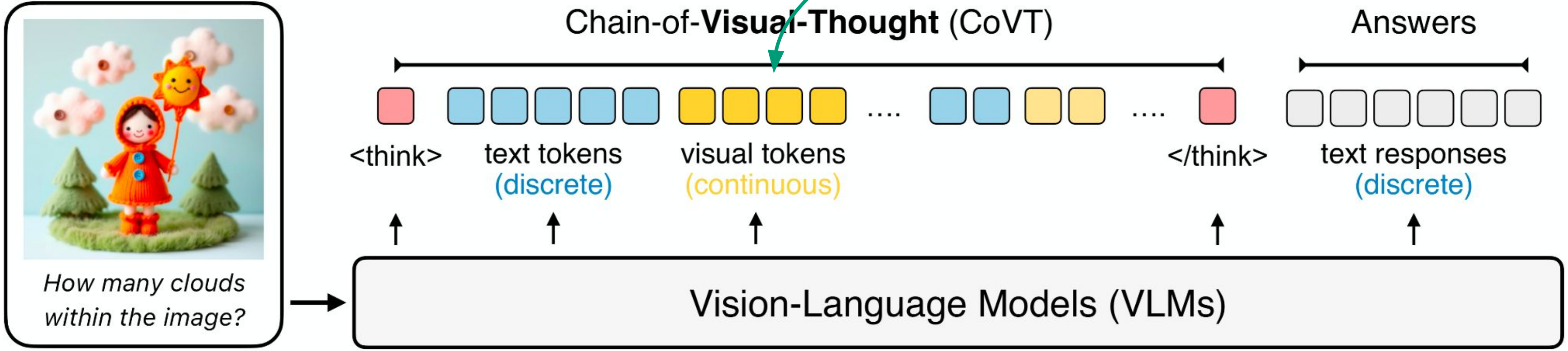
This paper:
Can we better **reason** over **visual** edges/layout/depth using **continuous tokens**?

Limitation: **Rich perceptual cues are poorly represented.**

- Visual information is inherently continuous.
- Aggregation/discretization: dense features like geometry, object boundaries, layout may be lost.

Introduction

Continuous tokens** for visual cues help **ground semantic reasoning.


How many clouds within the image?
Chain-of-**Visual-Thought** (CoVT)
Answers
….
….
<think>
text tokens (discrete)
visual tokens (continuous)
</think>
text responses (discrete)
Vision-Language Models (VLMs)

Limitation: **Rich perceptual cues are poorly represented.**

- Visual information is inherently continuous.
- Discretizing/aggregation: geometry, object boundaries, layout may be lost.

This paper:
Can we better **reason** over **visual** edges/layout/depth using **continuous tokens**?

Figure 1

Overview

Chain of Visual Thought (CoVT)

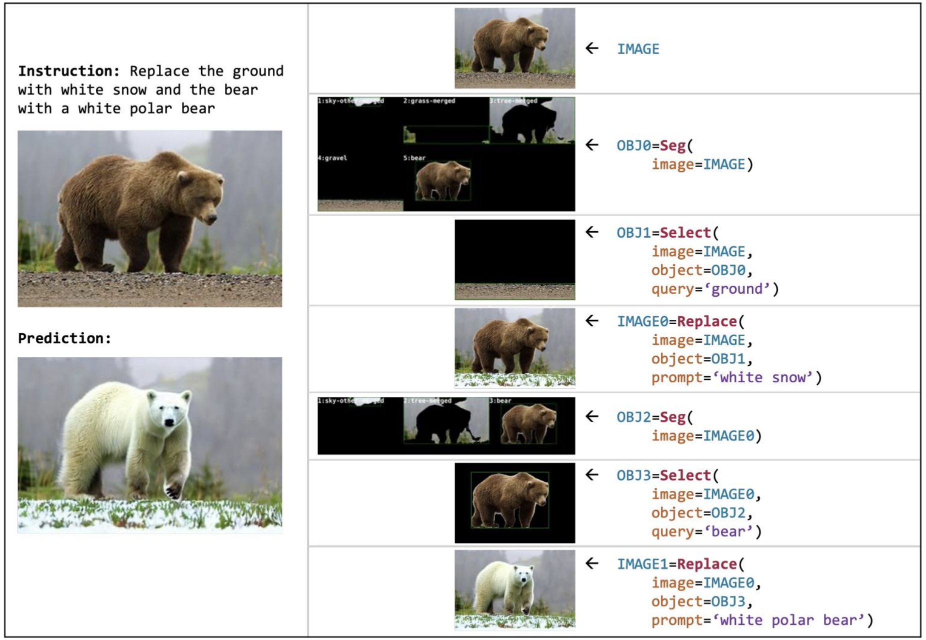
- Defines groups of continuous visual tokens.
  - Contained within <think> steps.
- Each group corresponds to a **perceptual cue**.
  - Segmentation, depth, edge detection, image representation.
- The VLM is trained to compress rich representations into these tokens.
  - Trained on reconstruction loss, target = result from an expert.
  - Visual features are aligned with its token representation

Result: reasoning in tokens without explicit maps or tool calls.

Background work

Scala**b**le Gener**a**tive Co**g**nitive Mod**el**
"**BAGEL**"

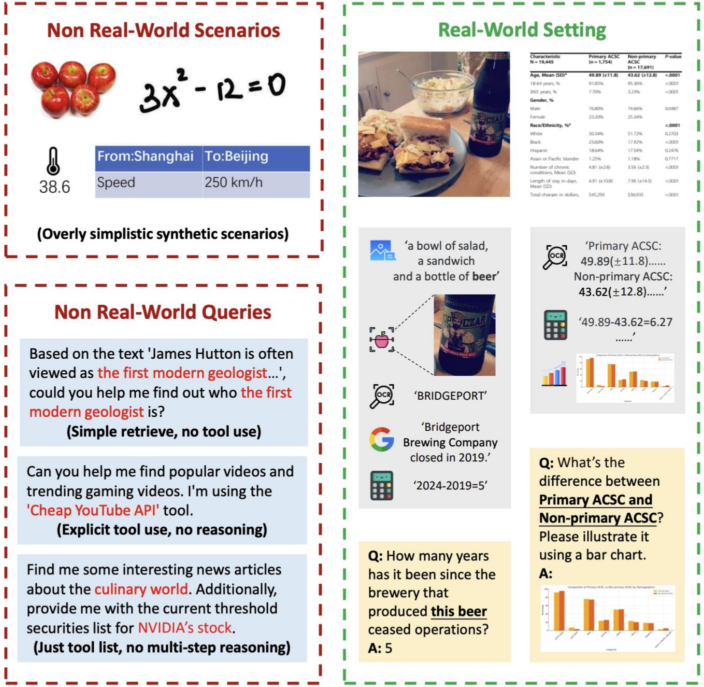
Tool-augmented reasoning



Instruction: Replace the ground with white snow and the bear with a white polar bear

Prediction:

```
← IMAGE
← OBJ0=Seg(
      image=IMAGE)
← OBJ1=Select(
      image=IMAGE,
      object=OBJ0,
      query='ground')
← IMAGE0=Replace(
      image=IMAGE,
      object=OBJ1,
      prompt='white snow')
← OBJ2=Seg(
      image=IMAGE0)
← OBJ3=Select(
      image=IMAGE0,
      object=OBJ2,
      query='bear')
← IMAGE1=Replace(
      image=IMAGE0,
      object=OBJ3,
      prompt='white polar bear')
```



Non Real-World Scenarios

From:Shanghai	To:Beijing
Speed	250 km/h

(Overly simplistic synthetic scenarios)

Non Real-World Queries

Based on the text 'James Hutton is often viewed as the first modern geologist...', could you help me find out who the first modern geologist is?
(Simple retrieve, no tool use)

Can you help me find popular videos and trending gaming videos. I'm using the 'Cheap YouTube API' tool.
(Explicit tool use, no reasoning)

Find me some interesting news articles about the culinary world. Additionally, provide me with the current threshold securities list for NVIDIA's stock.
(Just tool list, no multi-step reasoning)

Real-World Setting

'a bowl of salad, a sandwich and a bottle of **beer**'

'BRIDGEPORT'

'Bridgeport Brewing Company closed in 2019.'

'2024-2019=5'

Q: How many years has it been since the brewery that produced **this beer** ceased operations?
A: 5

'Primary ACSC: 49.89(±11.8)...... Non-primary ACSC: 43.62(±12.8)......'

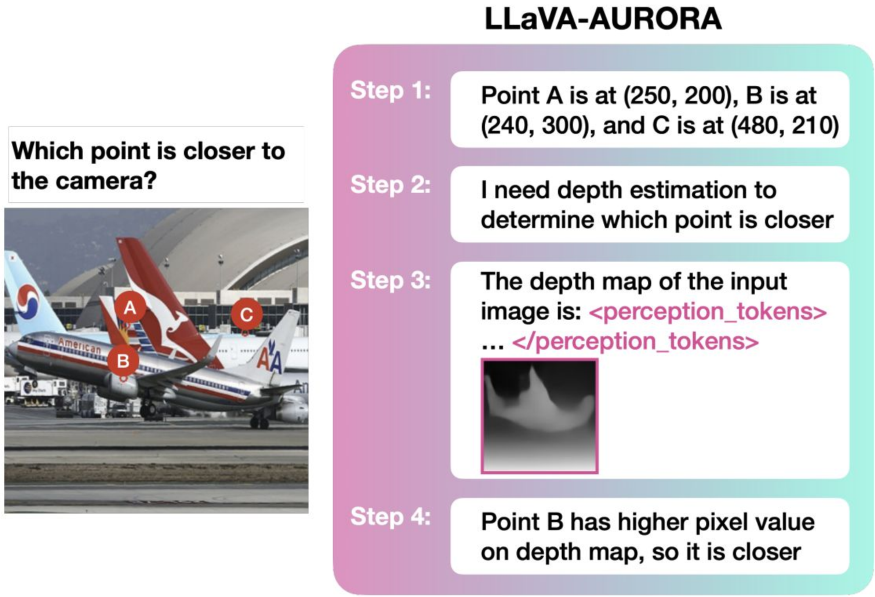
'49.89-43.62=6.27 ......'

Q: What's the difference between **Primary ACSC and Non-primary ACSC**? Please illustrate it using a bar chart.
A:

Visual Programming: Compositional visual reasoning without training
ToolVQA: A Dataset for Multi-step Reasoning VQA with External Tools

Background work

Text Space Reasoning



LLaVA-AURORA

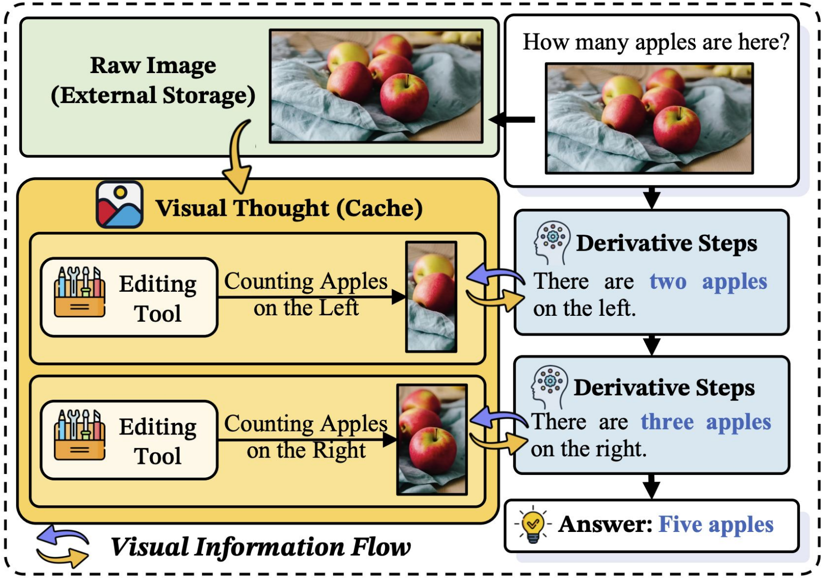
Which point is closer to the camera?

Step 1: Point A is at (250, 200), B is at (240, 300), and C is at (480, 210)

Step 2: I need depth estimation to determine which point is closer

Step 3: The depth map of the input image is: <perception_tokens> … </perception_tokens>

Step 4: Point B has higher pixel value on depth map, so it is closer



How many apples are here?

Raw Image (External Storage)

Visual Thought (Cache)

Editing Tool → Counting Apples on the Left

Editing Tool → Counting Apples on the Right

Derivative Steps
There are **two apples** on the left.

Derivative Steps
There are **three apples** on the right.

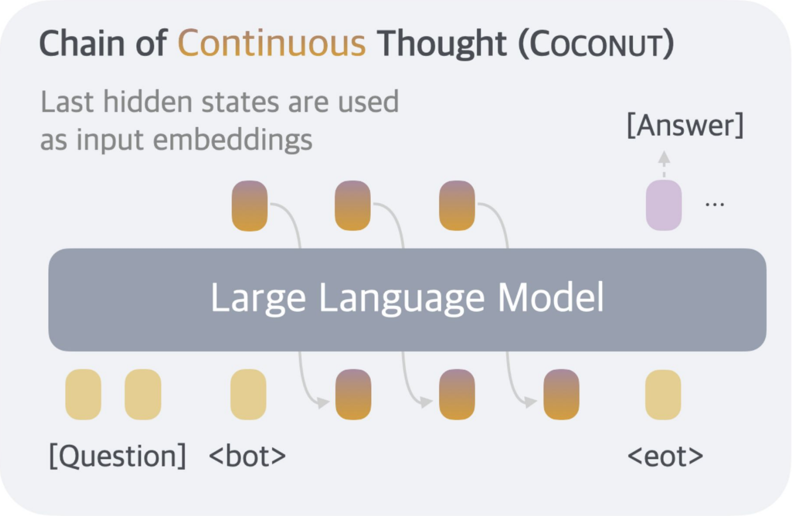
Answer: Five apples

Visual Information Flow

Perception Tokens Enhance Visual Reasoning in Multimodal Language Models (AURORA)
Visual Thoughts: A Unified Perspective of Understanding Multimodal Chain-of-Thought (MCoT)

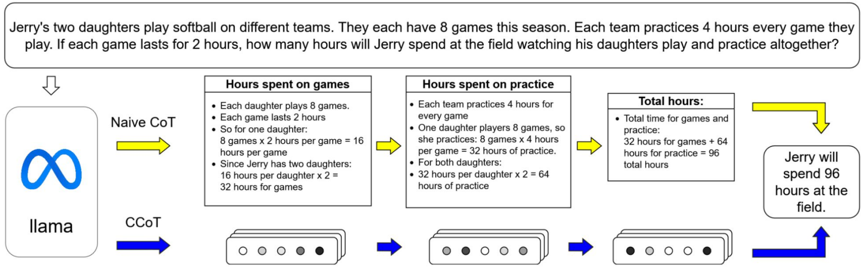
Background work

Latent Space Reasoning



Chain of Continuous Thought (COCONUT)

Last hidden states are used as input embeddings



Jerry's two daughters play softball on different teams. They each have 8 games this season. Each team practices 4 hours every game they play. If each game lasts for 2 hours, how many hours will Jerry spend at the field watching his daughters play and practice altogether?

Hours spent on games

- Each daughter plays 8 games.
- Each game lasts 2 hours
- So for one daughter:
  8 games x 2 hours per game = 16 hours per game
- Since Jerry has two daughters:
  16 hours per daughter x 2 = 32 hours for games

Hours spent on practice

- Each team practices 4 hours for every game
- One daughter players 8 games, so she practices: 8 games x 4 hours per game = 32 hours of practice.
- For both daughters:
- 32 hours per daughter x 2 = 64 hours of practice

Total hours:

- Total time for games and practice:
  32 hours for games + 64 hours for practice = 96 total hours

Jerry will spend 96 hours at the field.

Training Large Language Models to Reason in a Continuous Latent Space (COCONUT)
Compressed Chain of Thought: Efficient Reasoning through Dense Representations (CCoT)

Chain of Visual Thought (CoVT)

Intuition:

- Text-only CoT **accumulates errors.**
- Supervision is **dominated by text responses.**

CoVT:

- A framework that equips VLMs with the ability to reason through **continuous visual tokens.**
- Tailored alignment strategies and a training pipeline to enable VLMs to learn, interpret, and reason effectively within this continuous visual space.

Chain of Visual Thought (CoVT)

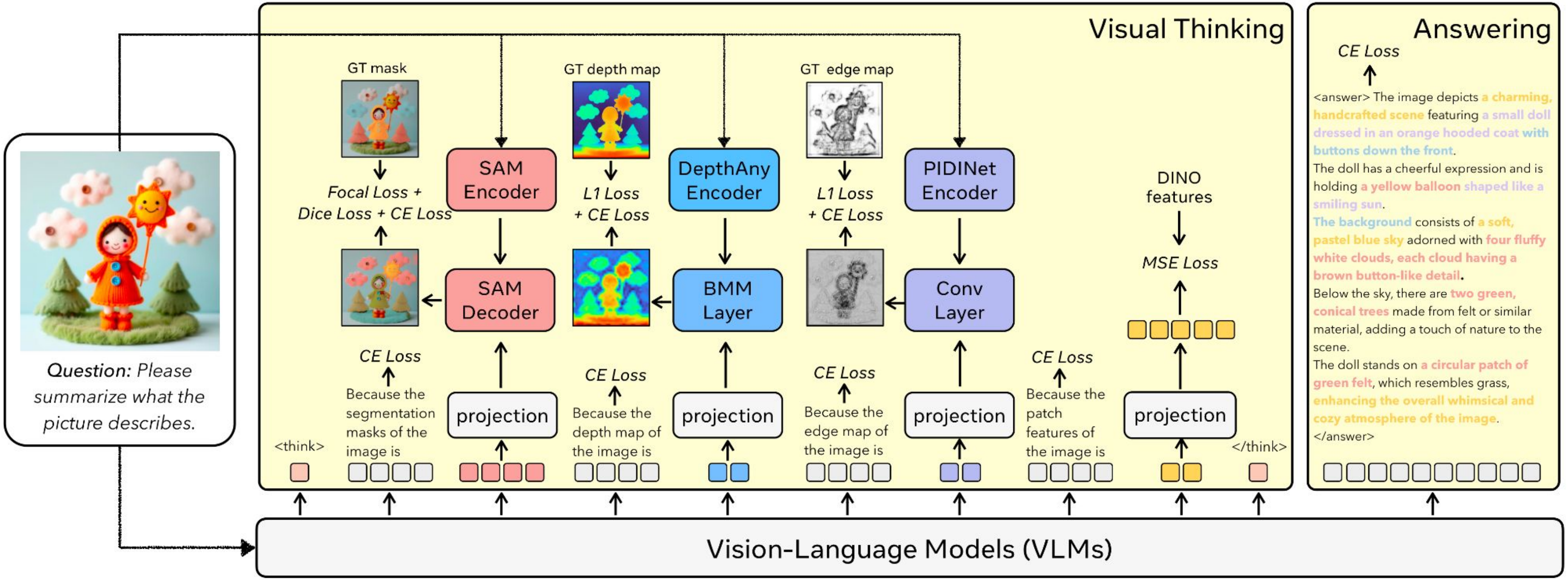
Pipeline

- Equips VLMs with chains of visual thought.
- VLM next token prediction:

$$P(Y \mid \mathcal{V}, \mathcal{T}; \theta) = \prod_{i=1}^{n} P\left(y_i \mid y_{<i}, \mathcal{V}, \mathcal{T}\right)$$

- This work extends this formulation by introducing ***CoVT tokens*** *:*
  - each *y_i* represents either a visual or text token.
- The VLM is trained to function as a dense visual encoder
  - Reconstruction supervision against task-specific decoders.

Chain of Visual Thought (CoVT)



Question: *Please summarize what the picture describes.*

<answer> The image depicts a charming, handcrafted scene featuring a small doll dressed in an orange hooded coat with buttons down the front.
The doll has a cheerful expression and is holding a yellow balloon shaped like a smiling sun.
The background consists of a soft, pastel blue sky adorned with four fluffy white clouds, each cloud having a brown button-like detail.
Below the sky, there are two green, conical trees made from felt or similar material, adding a touch of nature to the scene.
The doll stands on a circular patch of green felt, which resembles grass, enhancing the overall whimsical and cozy atmosphere of the image.
</answer>

Figure 3

Chain of Visual Thought (CoVT)

CoVT Tokens

Perceptual ability of VLMs can be summarized as:

1. Instance recognition
    - segmentation tokens provide positional & shape information.
2. 2D and 3D spatial relationships.
    - Depth tokens provide pixel-level depth information.
3. Structure detection.
    - Edge tokens provide geometry-level details.
4. Deep mining of semantic information.
    - DINO tokens provide patch representations.

Chain of Visual Thought (CoVT)

CoVT Tokens

Each token type is

- assigned a constant **token count** (within <think>)
- supervised by a domain expert

1. Segmentation *(8 tokens)*
   - Supervised by SAM – Segment Anything Model.
2. Depth *(4 tokens)*
   - Supervised by DepthAnything v2.
3. Edge *(4 tokens)*
   - Aligned with PIDINet.
4. DINO *(4 tokens)*
   - Supervised by DINOv2.

$$\hat{M}_i = \mathrm{Decoder}(T_i^{\mathrm{sam}}, f), \quad \hat{M}_i \in [0,1]^{H \times W}$$

$$\hat{D}_i = \mathrm{softmax}\left(T_i^{\mathrm{depth}} \cdot F_i^{\mathrm{depth}\top}\right)$$

Chain of Visual Thought (CoVT)

CoVT Training

- **Loss:**
    - Joint loss function: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ce}} + \gamma\big(\lambda_{\text{seg}} \cdot \mathcal{L}_{\text{visual}}^{\text{seg}} + \lambda_{\text{depth}} \cdot \mathcal{L}_{\text{visual}}^{\text{depth}} + \lambda_{\text{edge}} \cdot \mathcal{L}_{\text{visual}}^{\text{edge}} + \lambda_{\text{dino}} \cdot \mathcal{L}_{\text{visual}}^{\text{dino}}\big)$
    - During inference, visual thinking tokens are not decoded.

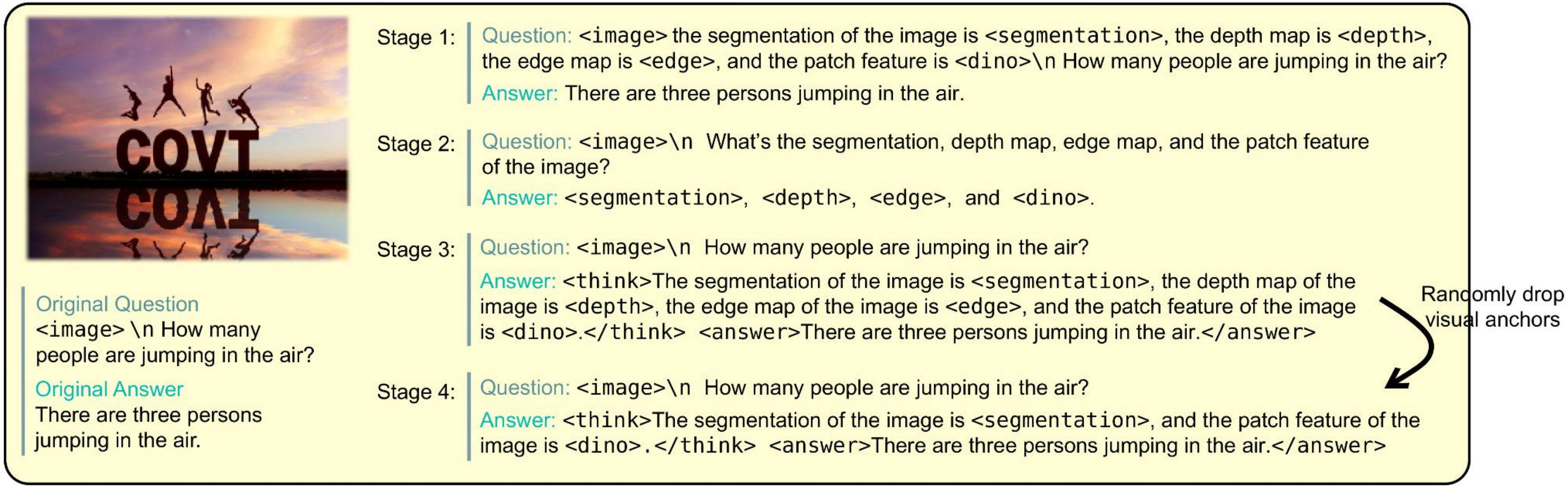
- **Data:**

  4-stage training, progressive leverage of visual tokens:

  1. *Comprehension:* Insert visual tokens after <image> to teach basic semantics.
  2. *Generation:* modify Q/A to guide the VLMs to precisely generate visual tokens.
  3. *Reasoning:* Introduces CoVT format, visual tokens are used in reasoning for final answer.
  4. *Efficient reasoning:* Randomly drop some sets of visual tokens (up to *k*); teach effective feature use.

Chain of Visual Thought (CoVT)

CoVT Training



Original Question
`<image>` \n How many people are jumping in the air?

Original Answer
There are three persons jumping in the air.

Stage 1: Question: `<image>` the segmentation of the image is `<segmentation>`, the depth map is `<depth>`, the edge map is `<edge>`, and the patch feature is `<dino>\n` How many people are jumping in the air?
Answer: There are three persons jumping in the air.

Stage 2: Question: `<image>\n` What's the segmentation, depth map, edge map, and the patch feature of the image?
Answer: `<segmentation>`, `<depth>`, `<edge>`, and `<dino>`.

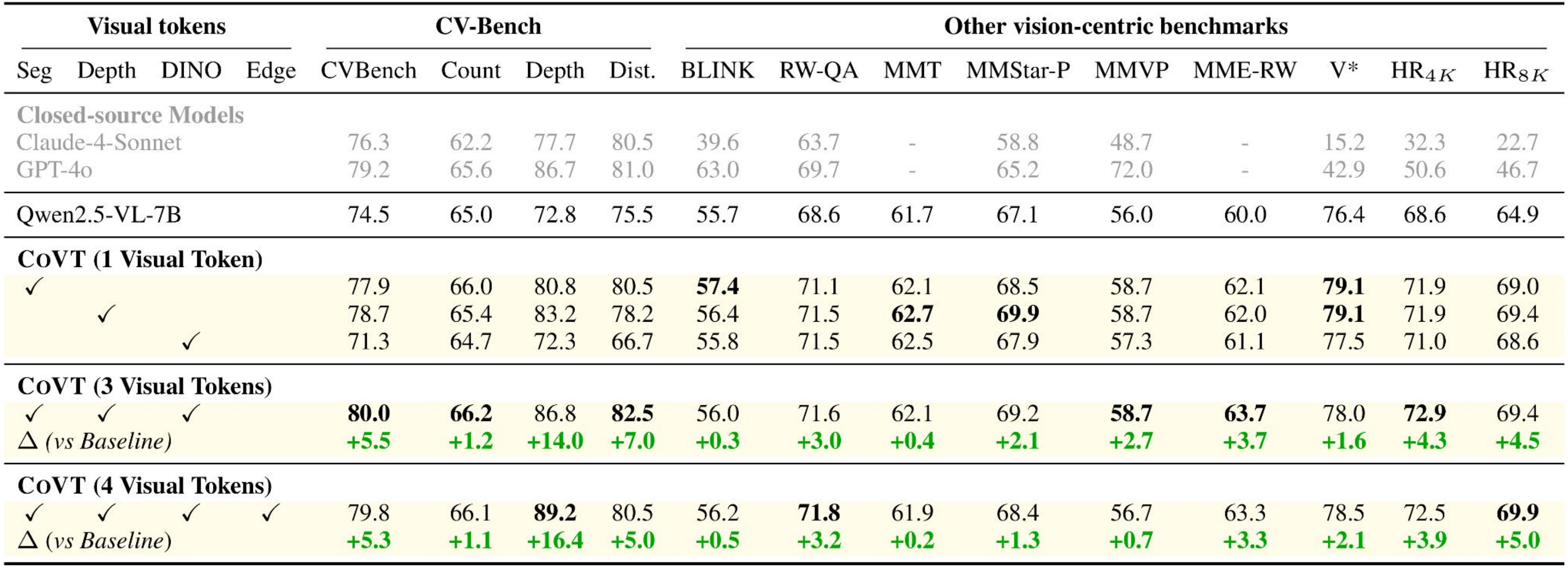
Stage 3: Question: `<image>\n` How many people are jumping in the air?
Answer: `<think>`The segmentation of the image is `<segmentation>`, the depth map of the image is `<depth>`, the edge map of the image is `<edge>`, and the patch feature of the image is `<dino>`.`</think>` `<answer>`There are three persons jumping in the air.`</answer>`

Randomly drop visual anchors

Stage 4: Question: `<image>\n` How many people are jumping in the air?
Answer: `<think>`The segmentation of the image is `<segmentation>`, and the patch feature of the image is `<dino>`.`</think>` `<answer>`There are three persons jumping in the air.`</answer>`

Figure 4.

Results: Quantitative

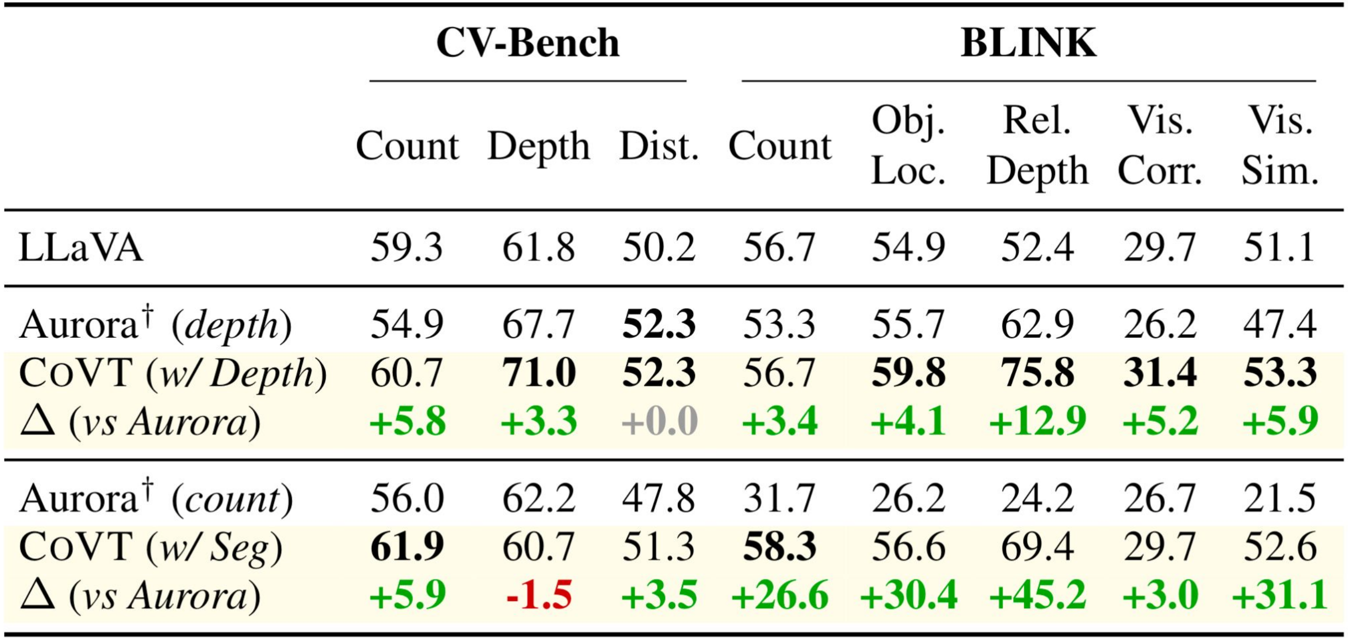


Visual tokens				**CV-Bench**				**Other vision-centric benchmarks**								
Seg	Depth	DINO	Edge	CVBench	Count	Depth	Dist.	BLINK	RW-QA	MMT	MMStar-P	MMVP	MME-RW	V*	HR_{4K}	HR_{8K}
Closed-source Models																
Claude-4-Sonnet				76.3	62.2	77.7	80.5	39.6	63.7	-	58.8	48.7	-	15.2	32.3	22.7
GPT-4o				79.2	65.6	86.7	81.0	63.0	69.7	-	65.2	72.0	-	42.9	50.6	46.7
Qwen2.5-VL-7B				74.5	65.0	72.8	75.5	55.7	68.6	61.7	67.1	56.0	60.0	76.4	68.6	64.9
CoVT (1 Visual Token)																
✓				77.9	66.0	80.8	80.5	**57.4**	71.1	62.1	68.5	58.7	62.1	**79.1**	71.9	69.0
	✓			78.7	65.4	83.2	78.2	56.4	71.5	**62.7**	**69.9**	58.7	62.0	**79.1**	71.9	69.4
		✓		71.3	64.7	72.3	66.7	55.8	71.5	62.5	67.9	57.3	61.1	77.5	71.0	68.6
CoVT (3 Visual Tokens)																
✓	✓	✓		**80.0**	**66.2**	86.8	**82.5**	56.0	71.6	62.1	69.2	**58.7**	**63.7**	78.0	**72.9**	69.4
Δ *(vs Baseline)*				**+5.5**	**+1.2**	**+14.0**	**+7.0**	**+0.3**	**+3.0**	**+0.4**	**+2.1**	**+2.7**	**+3.7**	**+1.6**	**+4.3**	**+4.5**
CoVT (4 Visual Tokens)																
✓	✓	✓	✓	79.8	66.1	**89.2**	80.5	56.2	**71.8**	61.9	68.4	56.7	63.3	78.5	72.5	**69.9**
Δ *(vs Baseline)*				**+5.3**	**+1.1**	**+16.4**	**+5.0**	**+0.5**	**+3.2**	**+0.2**	**+1.3**	**+0.7**	**+3.3**	**+2.1**	**+3.9**	**+5.0**

(CoVT on Qwen-2.5-VL-7B)

Table 2.

Results: Quantitative

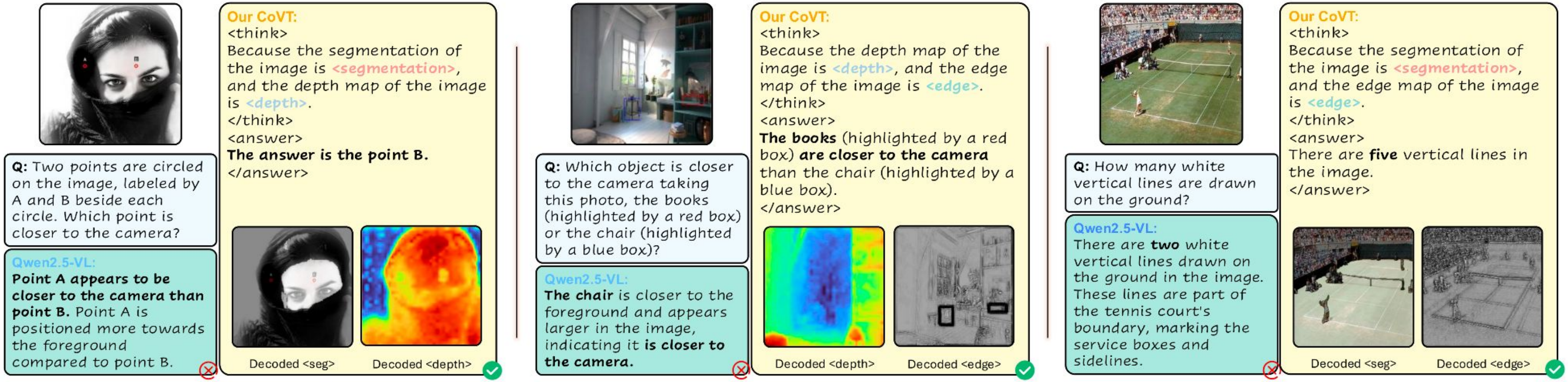


	CV-Bench			**BLINK**				
	Count	Depth	Dist.	Count	Obj. Loc.	Rel. Depth	Vis. Corr.	Vis. Sim.
LLaVA	59.3	61.8	50.2	56.7	54.9	52.4	29.7	51.1
Aurora† (*depth*)	54.9	67.7	**52.3**	53.3	55.7	62.9	26.2	47.4
CoVT (*w/ Depth*)	60.7	**71.0**	**52.3**	56.7	**59.8**	**75.8**	**31.4**	**53.3**
Δ (*vs Aurora*)	+5.8	+3.3	+0.0	+3.4	+4.1	+12.9	+5.2	+5.9
Aurora† (*count*)	56.0	62.2	47.8	31.7	26.2	24.2	26.7	21.5
CoVT (*w/ Seg*)	**61.9**	60.7	51.3	**58.3**	56.6	69.4	29.7	52.6
Δ (*vs Aurora*)	+5.9	-1.5	+3.5	+26.6	+30.4	+45.2	+3.0	+31.1

(CoVT on LLaVA-v1.5-13B, for parity with Aurora)

Table 3.

Results: Qualitative



Q: Two points are circled on the image, labeled by A and B beside each circle. Which point is closer to the camera?

Qwen2.5-VL:
Point A appears to be closer to the camera than point B. Point A is positioned more towards the foreground compared to point B.

Our CoVT:
<think>
Because the segmentation of the image is <segmentation>, and the depth map of the image is <depth>.
</think>
<answer>
The answer is the point B.
</answer>

Q: Which object is closer to the camera taking this photo, the books (highlighted by a red box) or the chair (highlighted by a blue box)?

Qwen2.5-VL:
The chair is closer to the foreground and appears larger in the image, indicating it **is closer to the camera.**

Our CoVT:
<think>
Because the depth map of the image is <depth>, and the edge map of the image is <edge>.
</think>
<answer>
The books (highlighted by a red box) **are closer to the camera** than the chair (highlighted by a blue box).
</answer>

Q: How many white vertical lines are drawn on the ground?

Qwen2.5-VL:
There are **two** white vertical lines drawn on the ground in the image. These lines are part of the tennis court's boundary, marking the service boxes and sidelines.

Our CoVT:
<think>
Because the segmentation of the image is <segmentation>, and the edge map of the image is <edge>.
</think>
<answer>
There are **five** vertical lines in the image.
</answer>

Figure 5.

Ablations

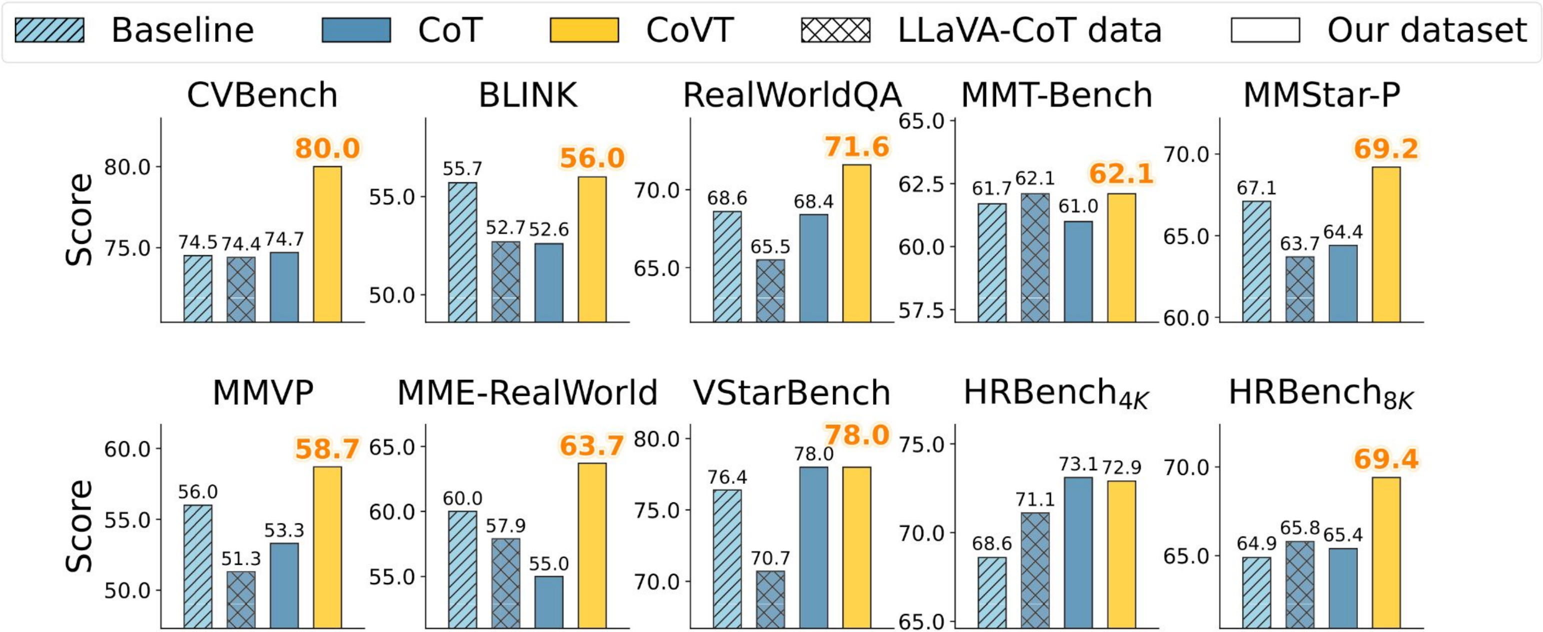


Fig. 6: Text-only CoT vs. CoVT.

Figure 6.

Ablations

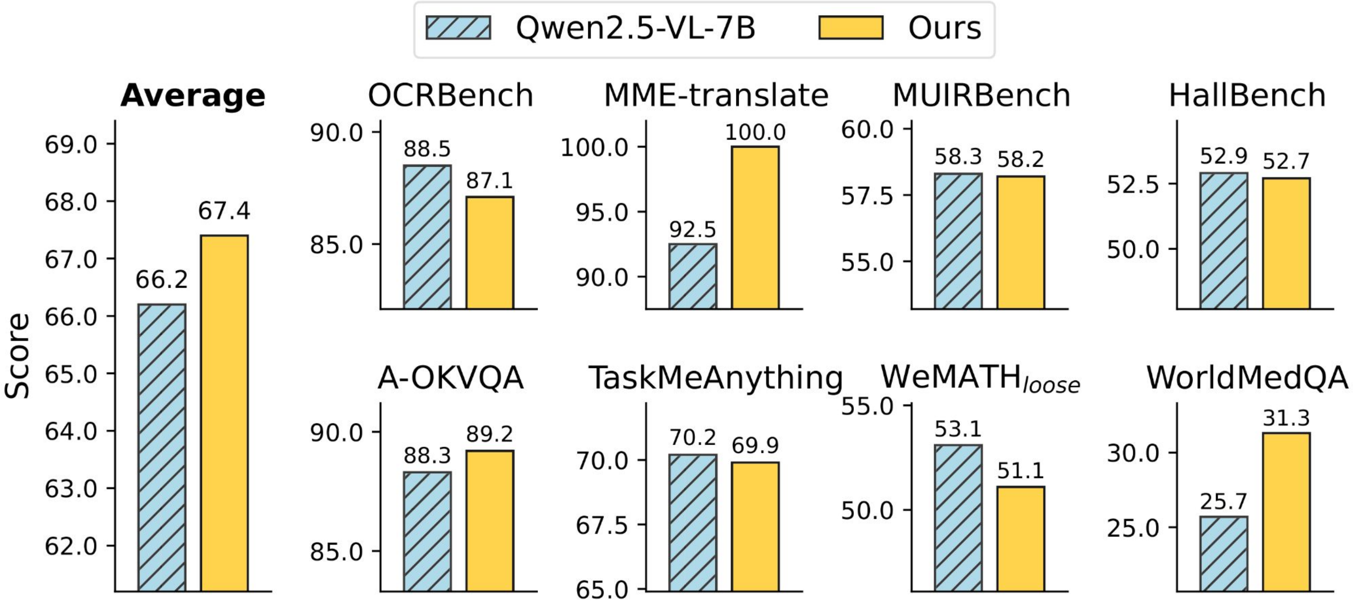


Fig. 7: Non vision-centric tasks.

Figure 7.

Ablations

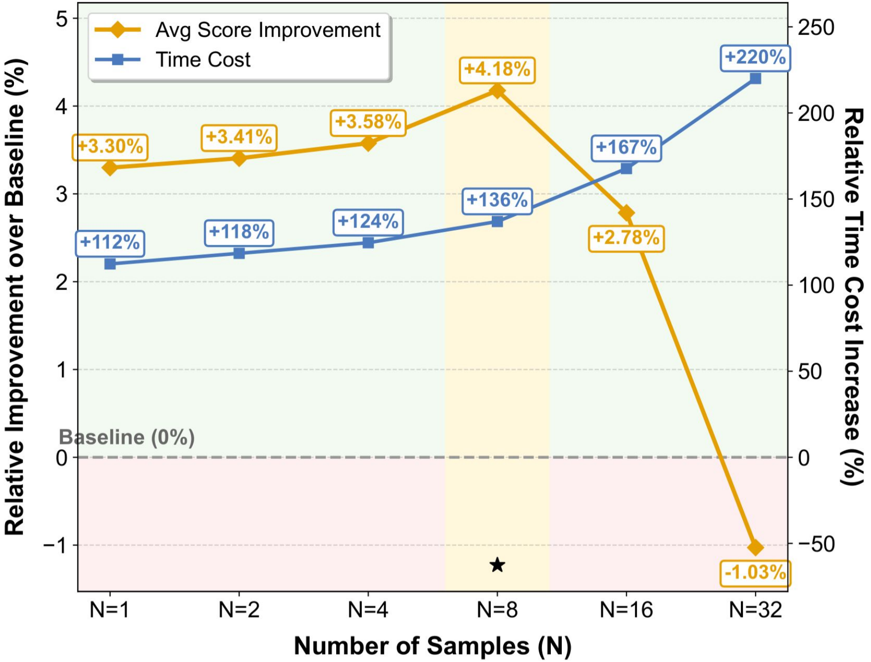
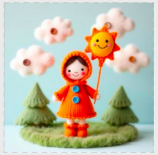


Fig. 12: performance vs. # of segmentation tokens

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Figure 12.

Examples



Question

Describe the scene in the picture in detail, and find out how many clouds are there in the sky.

Qwen2.5-VL

The image depicts a charming, handcrafted scene featuring a small, knitted figure dressed in an orange hooded outfit with matching boots. The figure is holding a cheerful, smiling sun-shaped balloon on a stick. The background includes three fluffy, cloud-like shapes in various shades of pink and white, floating above a small, green, grassy area with two coniferous trees. The sky is a soft, light blue color.

There are four clouds visible in the sky: one large pink cloud on the left, one smaller white cloud near the top right, and two more white clouds further to the right.

CoVT

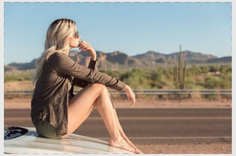
<think> Because the segmentation of the image is <|sam_pad|><|sam_pad|><|sam_pad|><|sam_pad|><|sam_pad|><|sam_pad|><|sam_pad|><|sam_pad|>, the depth map of the image is <|depth_pad|><|depth_pad|><|depth_pad|><|depth_pad|>, and the perception feature of the image is <|dino_pad|><|dino_pad|><|dino_pad|><|dino_pad|>. </think>

<answer> The image depicts a charming, handcrafted scene featuring a small doll-like figure standing on a grassy patch. The figure is dressed in an orange outfit with matching boots, and it has a cheerful expression. The doll is holding a stick with a smiling sun at the top, which appears to be made from felt or similar material.

In the background, there are two green felt trees, each with a simple triangular shape and a brown trunk. Above the trees, there are four fluffy white clouds, each adorned with a brown button-like detail. The sky is painted in a soft blue color, providing a serene backdrop for the scene.

The overall composition is whimsical and playful, likely intended to evoke a sense of joy and innocence. The use of bright colors and simple shapes contributes to the overall aesthetic appeal of the image.

To answer the question about the number of clouds: There are four clouds visible in the sky above the trees. </answer>



Question

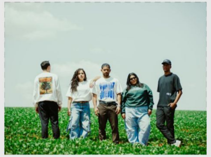
What is the beautiful woman sitting on in the picture?

Qwen2.5-VL

The woman in the picture is sitting on the hood of a car.

CoVT

<think> Because the segmentation of the image is <|sam_pad|><|sam_pad|><|sam_pad|><|sam_pad|><|sam_pad|><|sam_pad|><|sam_pad|><|sam_pad|>, and the depth map of the image is <|depth_pad|><|depth_pad|><|depth_pad|><|depth_pad|>. </think>
<answer> The beautiful woman is sitting on the hood of a white car. </answer>



Question

How many men are facing the camera in the picture?

Qwen2.5-VL

In the picture, there are three men facing the camera.

CoVT

<think> Because the segmentation of the image is <|sam_pad|><|sam_pad|><|sam_pad|><|sam_pad|><|sam_pad|><|sam_pad|><|sam_pad|><|sam_pad|>, and the depth map of the image is <|depth_pad|><|depth_pad|><|depth_pad|><|depth_pad|>. </think>
<answer> 2 </answer>

Figs. 13 & 14.