

Seeing, Speaking, and Reasoning in a Visual World

Cees Snoek

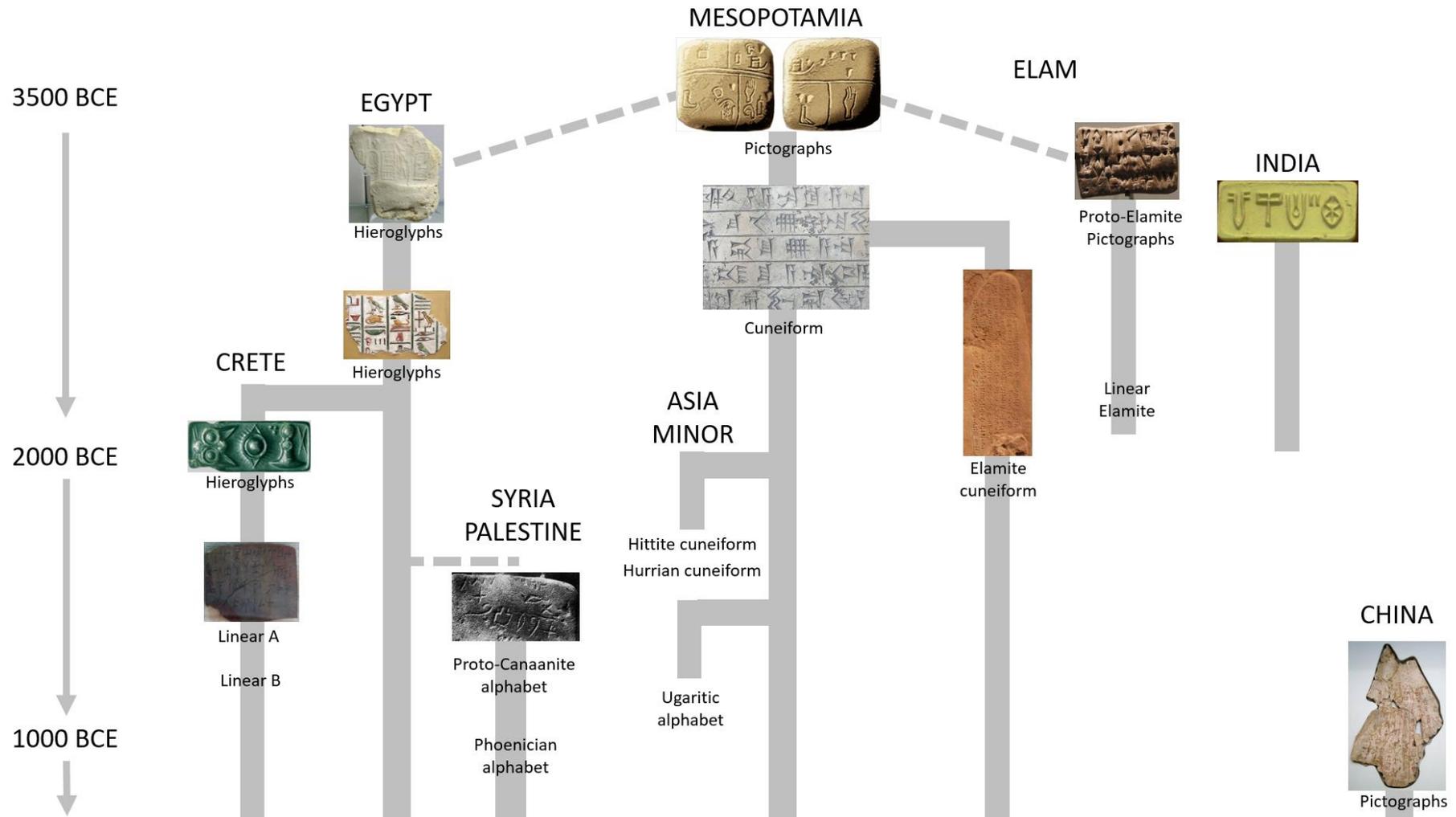


UNIVERSITY OF AMSTERDAM

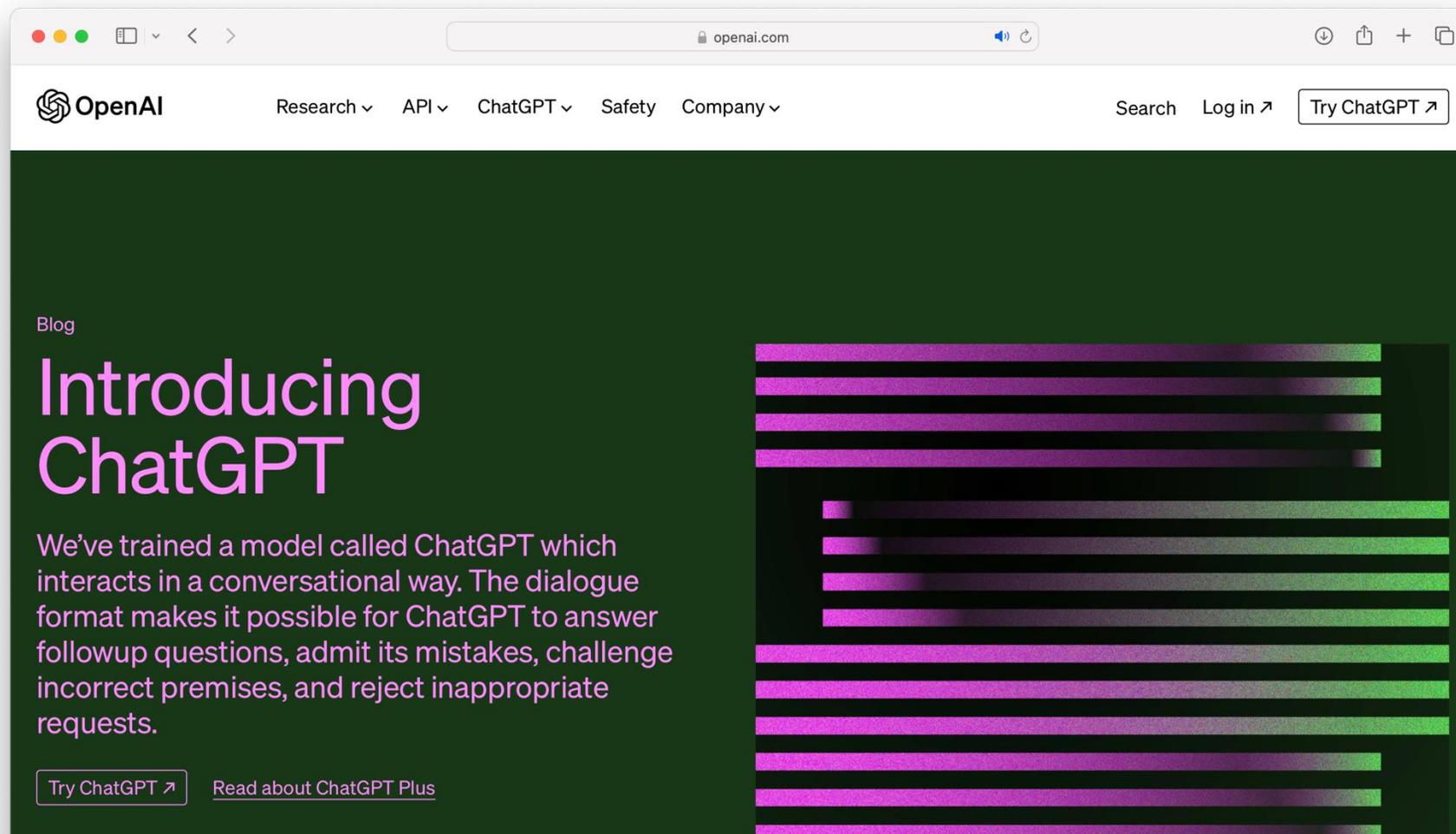




Human invention of written language



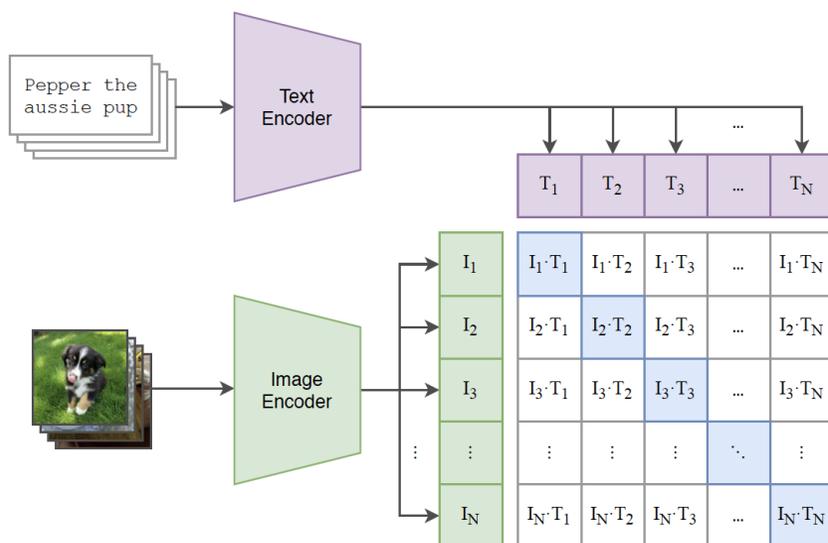
Human invention of ChatGPT



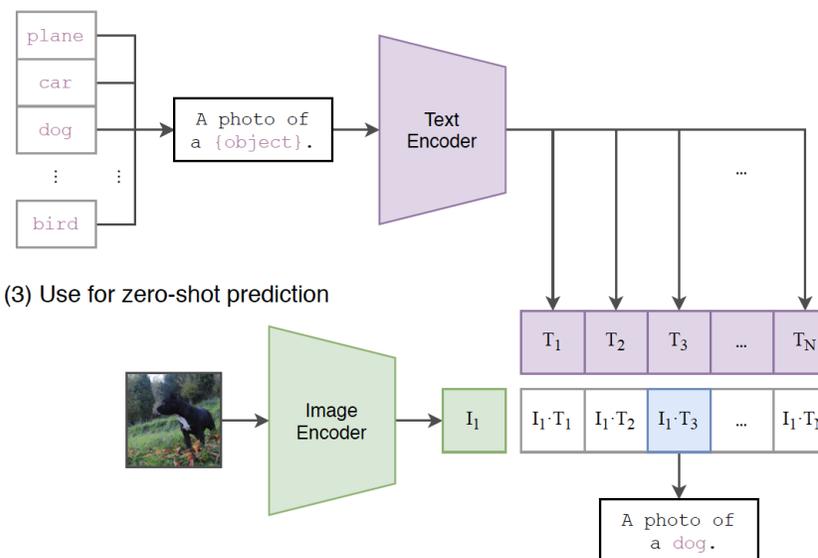
Vision and language even more powerful

1. Collect millions of images and their description from the Internet
2. Learn associations between encoded image and text
3. Amazing predictive abilities

(1) Contrastive pre-training



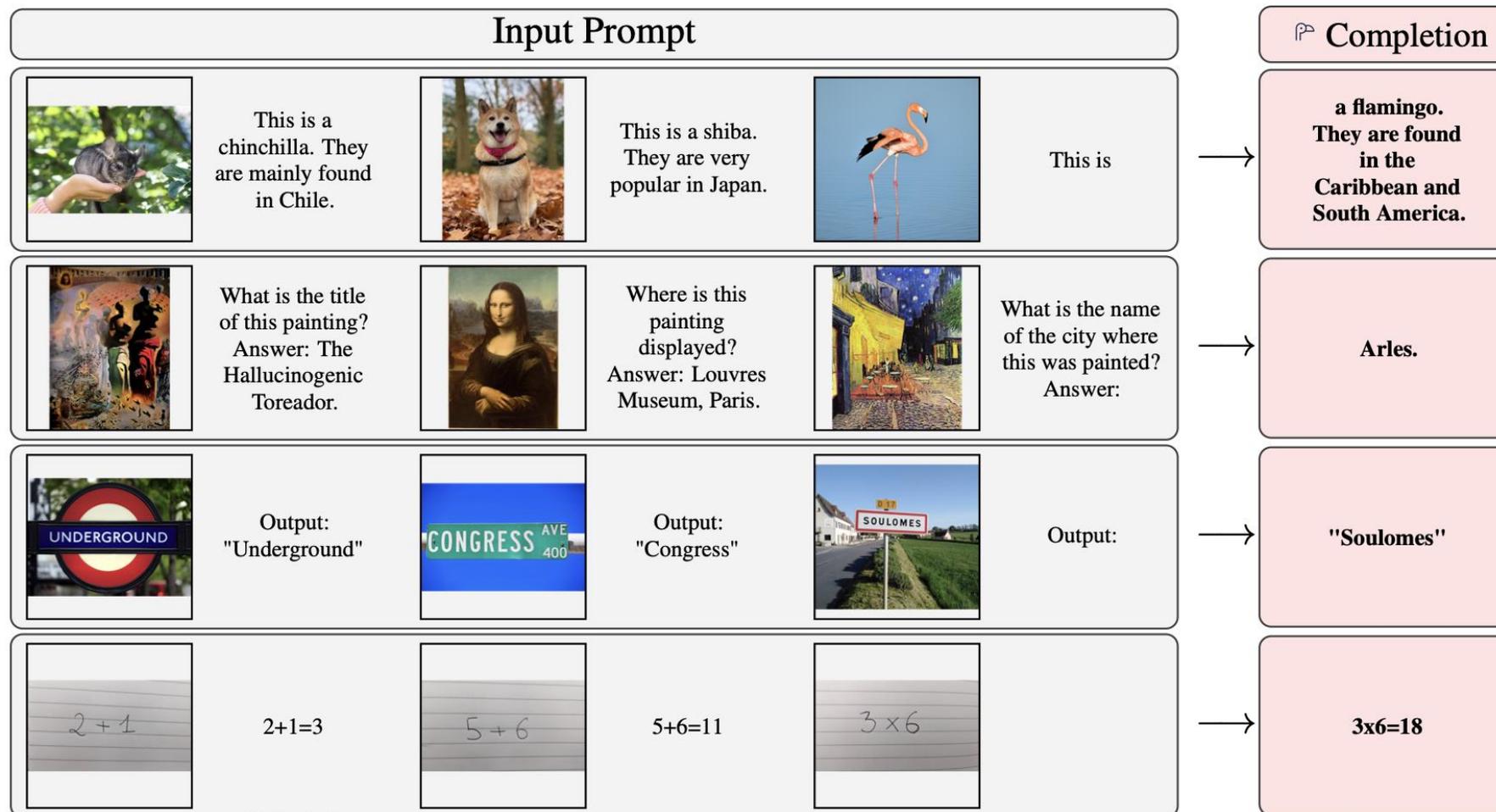
(2) Create dataset classifier from label text



(3) Use for zero-shot prediction



Task-specialization emerges at scale



This talk

A plead to leverage synergy between vision, language and reasoning to better understand the (visual) world

I. Seeing and Speaking

II. Seeing and Reasoning

Part I: Seeing and Speaking



NeoBabel: A Multilingual Open Foundation Model for Visual Generation

Mohammad Mahdi Derakhshani, Dheeraj Varghese, Marzieh Fadaee, and Cees G. M. Snoek

 Cohere Labs



UNIVERSITY OF AMSTERDAM



Funded by
the European Union

Text-to-image generation is amazing

A golden retriever lying peacefully on a wooden porch, with autumn leaves scattered around.



A young woman with freckles wearing a straw hat, standing in a golden wheat field.



A single drop of water clinging to a green leaf, with sunlight creating a faint rainbow prism.



An ancient stone bridge arching over a crystal-clear mountain stream, surrounded by lush greenery.

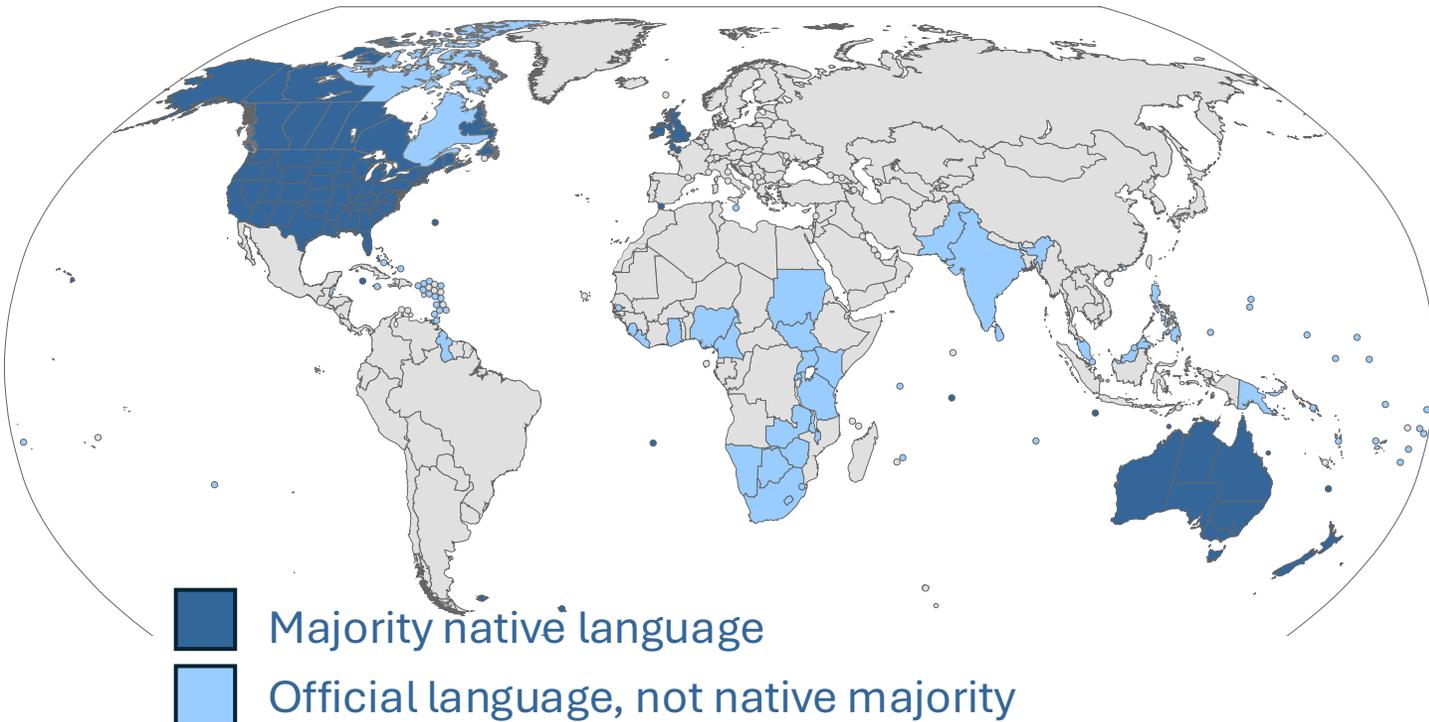


Credit: DeepSeek's Janus Pro 7B

Problem: A monolingual AI world

Text-to-Image generation is English-centric.

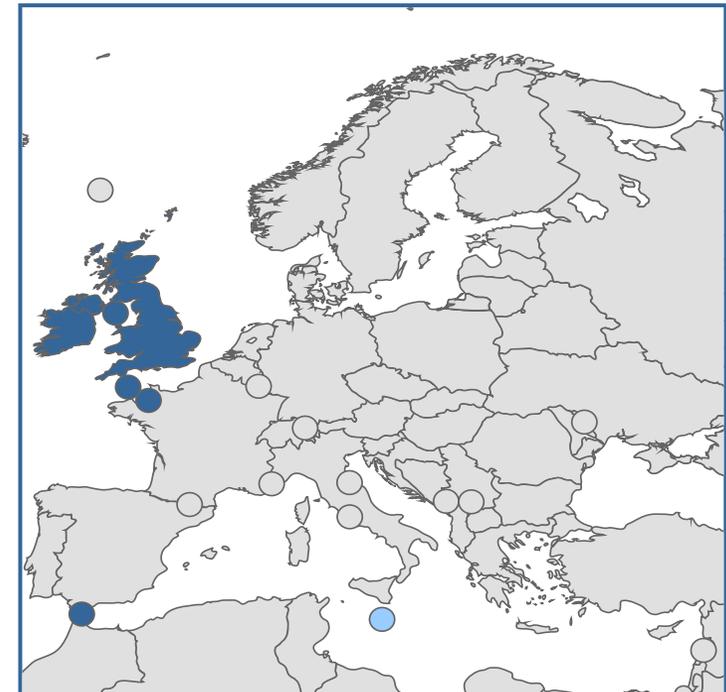
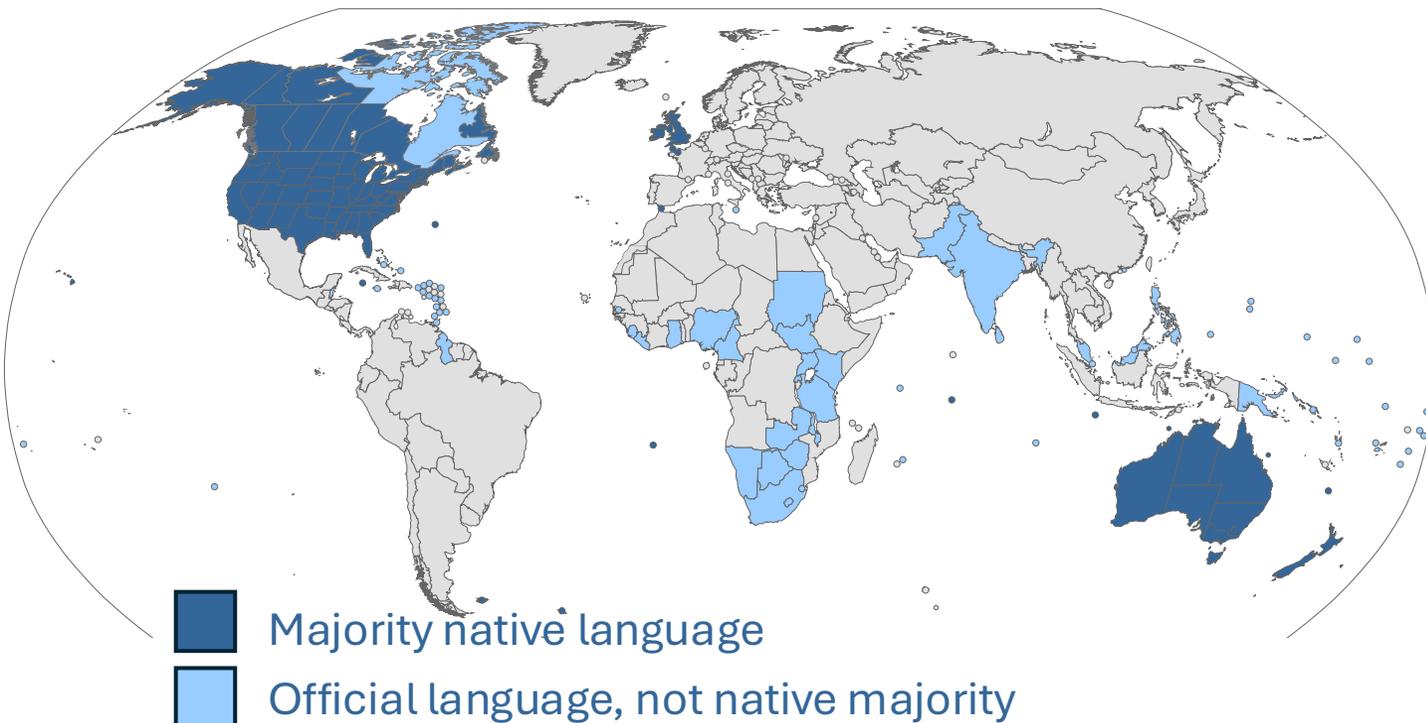
- State-of-the-art models serve $\sim 5\%$ of the world's native speakers.
- This creates a significant **barrier to global access** and perpetuates **digital inequity**.



Problem: A monolingual AI world

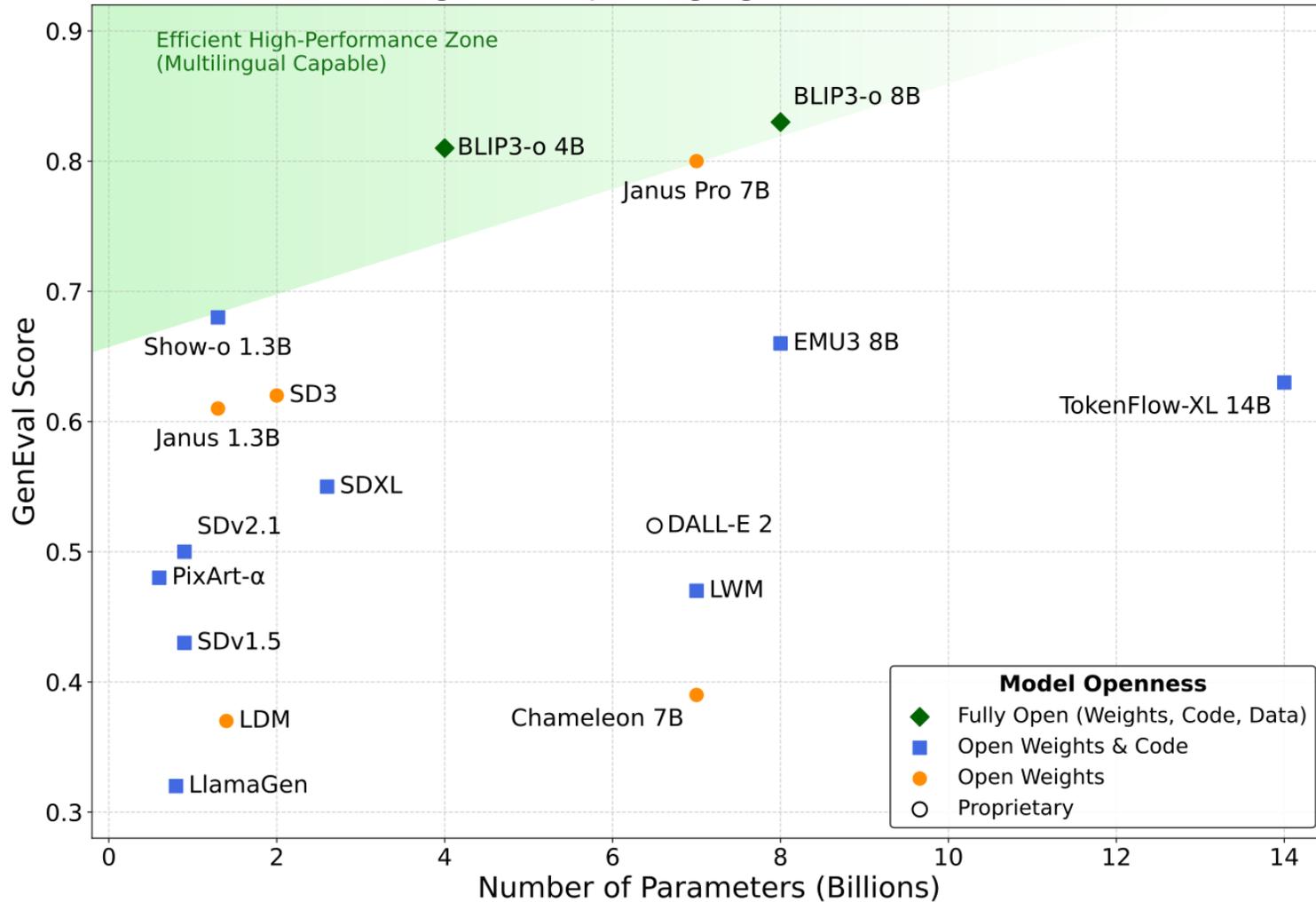
Text-to-Image generation is English-centric.

- State-of-the-art models serve $\sim 5\%$ of the world's native speakers.
- This creates a significant **barrier to global access** and perpetuates **digital inequity**.



Text-to-Image Generation state-of-the-art

English-only image generation



Field moves fast

Almost all models English-only

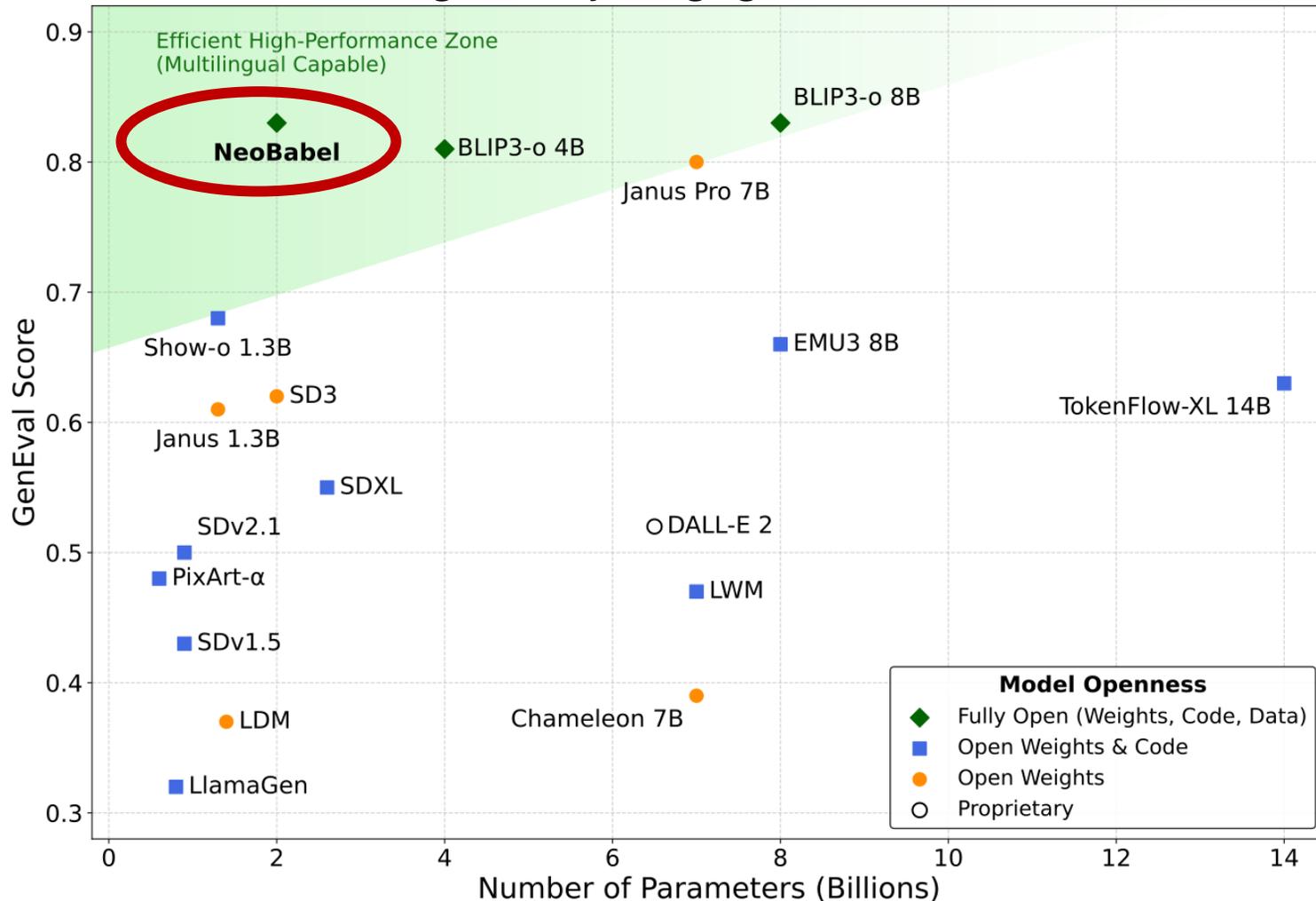
Some rely on multilingual LLM

No native multilinguality

Limited openness

NeoBabel: Natively multilingual, efficient & open

English-only image generation



NeoBabel is SOTA for English

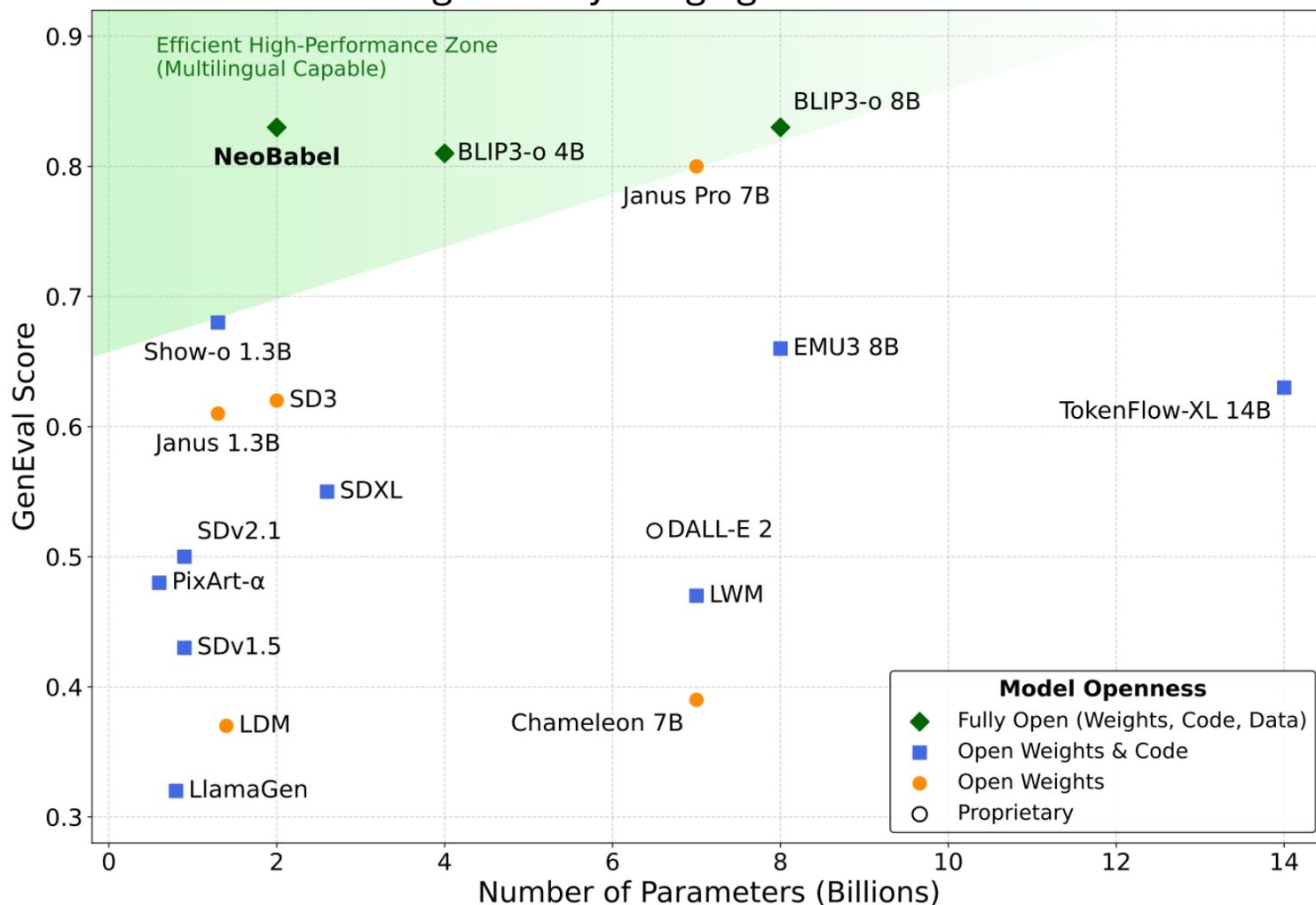
Model size 2-4x smaller

Fully open

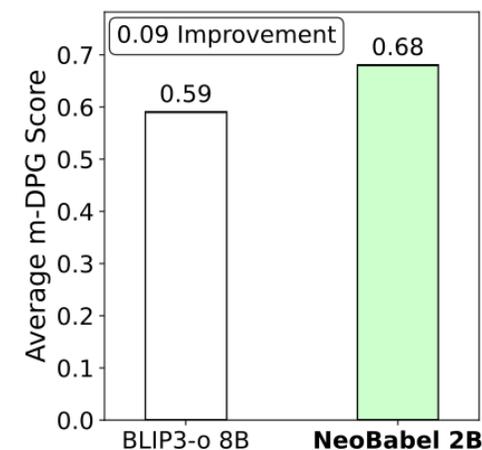
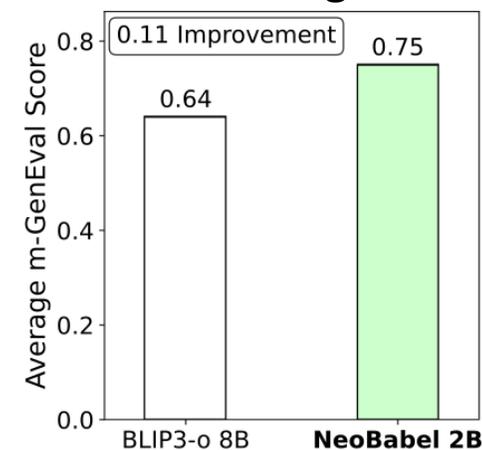
Native support for six languages

NeoBabel: Natively multilingual, efficient & open

English-only image generation



Multilingual image generation



English

Chinese

Dutch

French

Hindi

Persian

NeoBabel contributions

First natively multilingual text to image generation system.

- Large-scale multilingual image data set
- Unified multimodal architecture with staged training

Comprehensive multilingual benchmark and metrics.

Open toolkit for inclusive research.

Data

Challenge: High-quality multilingual data is scarce

We expand existing datasets by model-generated recaptioning and translation

Original English-Only Dataset			
Dataset	Image Source	Caption Source	Size
ImageNet 1K	Web	Class labels	1M
CC12M	Web	Alt-text (noisy)	12M
SA-1B	Photography	LLaVA	10M
LAION-Aesthetic	Web	Alt-text (noisy)	12M
JourneyDB	Synthetic	GPT-3.5	4M
BLIP3-o Instruct	Web + Synthetic	GPT-4o / human	60K
			39M

ImageNet translation example

Translation by Gemini Experimental models.

ImageNet 1k



English caption: streetcar **French caption:** tramway
Chinese caption: 有轨电车 **Hindi caption:** स्ट्रीटकार
Dutch caption: tram **Persian caption:** واگن برقی

ImageNet 1k



English caption: American chameleon
Chinese caption: 美洲变色龙
Dutch caption: Amerikaanse kameleon
French caption: caméléon américain
Hindi caption: अमेरिकी गिरगिट
Persian caption: آفتابپرست آمریکایی

ImageNet 1k



English caption: Bernese mountain dog
Chinese caption: 伯恩山犬
Dutch caption: Berner Sennenhond
French caption: bouvier bernois
Hindi caption: बर्नीज माउंटेन डॉग
Persian caption: سگ کوهستانی برن

CC12M recaptioning example

Recaptioning by InternVL (Zhe Chen et al. CVPR 2024)



Conceptual Caption 12M

Original caption: The Scotch House Vintage Black Velvet Scottish Kilt Style Jacket 3.

Recaptioned: This image displays a black velvet jacket with silver buttons on a white mannequin. The jacket is tailored, with a fitted waist and peaked lapels. The mannequin is wearing a black velvet bowtie. The jacket has three silver buttons on the front, and four silver buttons on each sleeve. The background is a white brick wall. The style of the image is a fashion photograph.

JourneyDB recaptioning and translation

Recaptioning by InternVL (Zhe Chen et al. CVPR 2024)



JourneyDB

Original caption: a beautiful princess with long hair walking with her tiger in a garden, dream, enchantment, wonder, stardust, HD, 8k, Paul Guy Gantner, Mandy Jungens, Camilla D'Errico, --ar 3:2 --q 2 --v 4

Recaptioned: the image is a beautifully detailed and vividly colored painting featuring two main subjects, a woman with long flowing hair adorned with flowers on the left side of her head, wearing an elegant off-the-shoulder dress that has intricate gold embroidery along one sleeve. She stands beside a majestic tiger looking directly at the viewer. The background consists of lush green foliage interspersed with pink blossoms to the left and orange poppies scattered across the ground towards the right. The setting appears to be outdoors near ancient stone ruins, which are partially visible behind them. The lighting suggests it might be late afternoon as there's a warm glow illuminating both figures from above, creating soft shadows around their forms. Overall, the scene conveys a sense of harmony between nature and wildlife through its rich colors and serene atmosphere.

JourneyDB recaptioning and translation

Translation by NLLB model by Meta AI, UC Berkeley, Johns Hopkins University.



JourneyDB

English caption: The image is a beautifully detailed and vividly colored painting featuring two main subjects, a woman with long flowing hair adorned with flowers on the left side of her head, wearing an elegant off-the-shoulder dress that has intricate gold embroidery along one sleeve ...

Chinese caption: 图片是一个精彩细节和生动的彩色绘画,特色的两个主要主题,一个长发的女人,头部左侧装饰着花朵,穿着优雅的肩膀上衣,有一袖上有复杂的金色刺...

Dutch caption: De foto is een prachtig gedetailleerd en levendig gekleurd schilderij met twee hoofdpersonen, een vrouw met lang vloeiend haar, versierd met bloemen aan de linkerkant van haar hoofd, met een elegante off-the-shoulder jurk met ingewikkelde gouden borduurwerk langs een mouw ...

French caption: L'image est une peinture magnifiquement détaillée et vivante avec deux sujets principaux, une femme aux cheveux longs et fluides ornés de fleurs sur le côté gauche de sa tête, portant une élégante robe à épaule qui a des broderie en or compliquée le long d'une manche. Elle se tient à côté d'un majestueux tigre regardant directement le spectateur ...

Hindi caption: यह चित्र एक खूबसूरत विस्तृत और जीवंत रंगीन चित्र है जिसमें दो मुख्य विषय हैं, एक महिला अपने बाएं हाथ के सिर पर फूलों से सजाए गए लंबे बहते बाल के साथ, एक सुरुचिपूर्ण ऑफ-द-कंधे पोशाक पहनी हुई है जिसमें एक आस्तीन के साथ जटिल सोने की कढ़ाई है। ...

Persian caption: تصویر یک نقاشی زیبا و دقیق و رنگارنگ است که دو سوژه اصلی را شامل می شود، یک زن با موهای بلند و پر از گل در سمت چپ سرش، پوشیدن یک لباس ظریف که در کنار یک آستین دارای نقاشی طلا پیچیده است

BLIP3-o translation example

Translation by NLLB and Gemini Experimental models.



BLIP3-o Instruct

English caption: A surreal 70s magazine photo of a hamburger man on stilts walking on the beach with rich detail and soft lighting.

Chinese caption: 一张超现实的 70 年代杂志照片，画面是一个汉堡包人在高跷上行走在海滩上，细节丰富，光线柔和。

Dutch caption: Een surrealistische jaren 70 tijdschriftfoto van een hamburgerman op stelten die op het strand loopt met rijke details en zachte verlichting.

French caption: Une photo surréaliste des années 70 d'un homme hamburger sur des échasses marchant sur la plage avec de riches détails et un éclairage doux.

Hindi caption: एक असली 70 के दशक की पत्रिका की तस्वीर एक हैम्बर्गर आदमी की समुद्र तट पर स्टिल्ट्स पर चलते हुए, समृद्ध विवरण और नरम प्रकाश के साथ।

Persian caption: یک عکس مجله سورئال دهه ۷۰ از یک مرد همبرگری روی داربست که در ساحل با جزئیات غنی و نور ملایم قدم می زند.

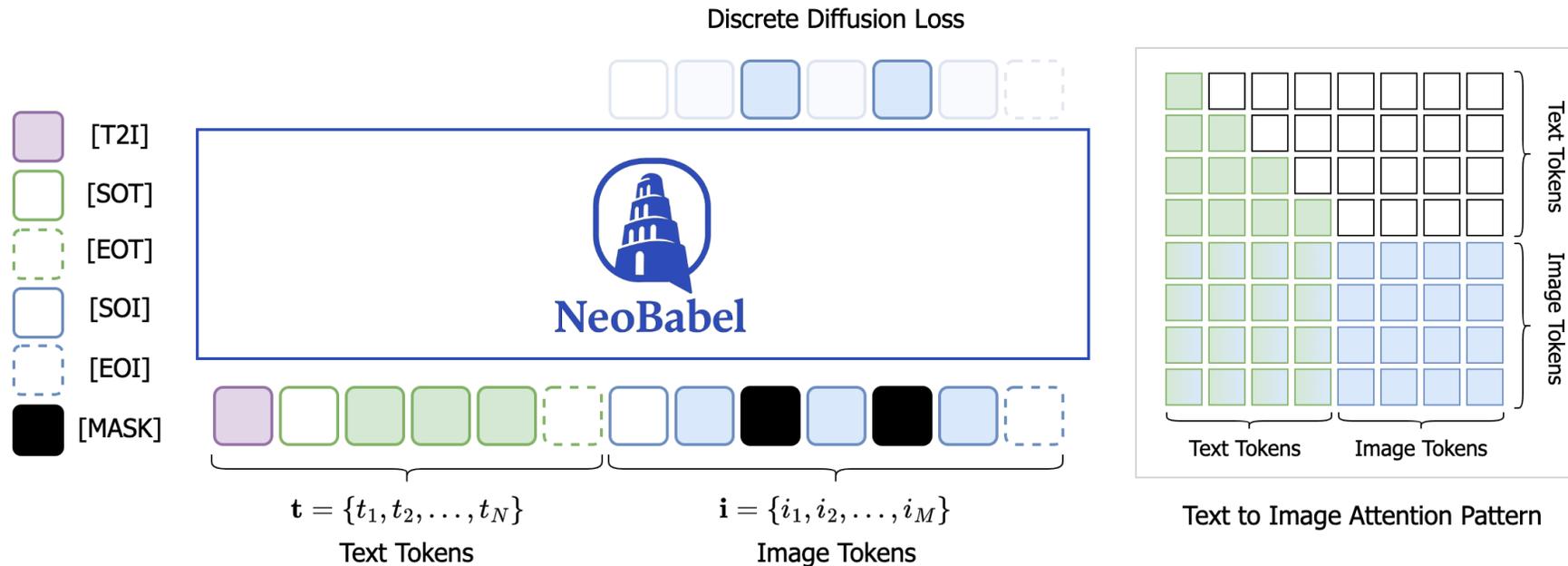
NeoBabel multilingual dataset

We increase the total size from 39M to 124M image–caption/label pairs

Original English-Only Dataset				NEOBABEL Multilingual Expansion		
Dataset	Image Source	Caption Source	Size	Recaptioning	Translation	New Size
ImageNet 1K	Web	Class labels	1M	–	✓	6M
CC12M	Web	Alt-text (noisy)	12M	✓	–	12M
SA-1B	Photography	LLaVA	10M	✓	–	10M
LAION-Aesthetic	Web	Alt-text (noisy)	12M	✓	✓	72M
JourneyDB	Synthetic	GPT-3.5	4M	✓	✓	24M
BLIP3-o Instruct	Web + Synthetic	GPT-4o / human	60K	–	✓	360K
			39M			124M

Training

NeoBabel multimodal architecture



Text Tokens: Generated via Gemma 2 tokenizer.

Image Tokens: Obtained from MAGVIT-v2, following Show-o

Architecture: Gemma 2 LLM decoder-only transformer

Training: Discrete diffusion predicts masked visual tokens.

Inference: Starts from fully masked image; tokens are iteratively unmasked.

Training stages

Progressive Pretraining

Stage 1 – Pixel Dependency Learning

6M datasets: m-ImageNet 1K

Stage 2 – Scaling Alignment with Large-Scale Multilingual Data

94M datasets: m-SA-1B, m-CC12M, m-LAION-Aesthetic

Stage 3 – Refined Multilingual Pretraining

96M datasets: m-LAION-Aesthetic and m-JourneyDB

256 x 256 resolution



Progressive Instruction Tuning

Stage 1 – Initial Multilingual Instruction Alignment

96M datasets: m-LAION-Aesthetic, m-JourneyDB, m-BLIP3o-Instruct

Stage 2 – Instruction Refinement

96M datasets: m-LAION-Aesthetic, m-JourneyDB, m-BLIP3o-Instruct

512 x 512 resolution

Evaluation

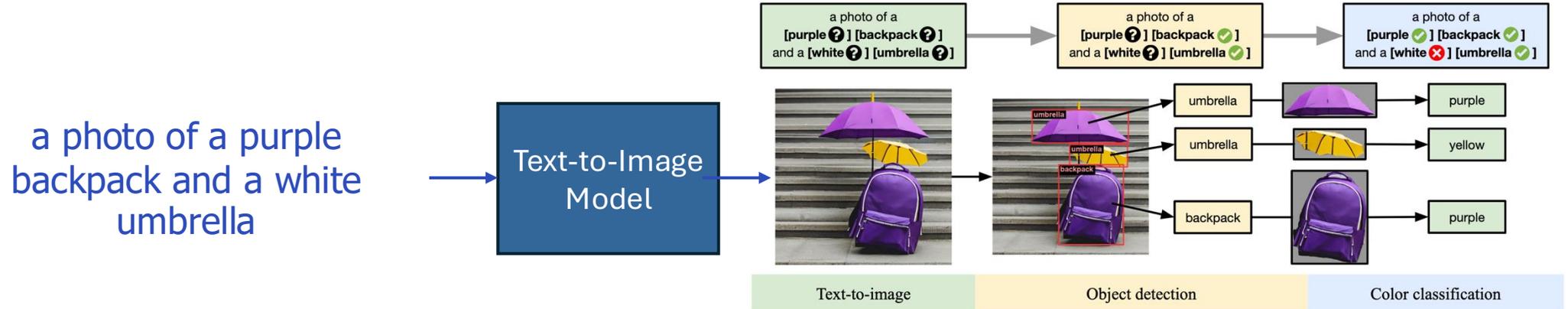
Existing benchmarks are English-only

We introduce a comprehensive multilingual evaluation suite including:

- **Multilingual extensions** of GenEval & DPG to 6 languages.
- **Two new metrics** for evaluation:
 - Cross-Lingual Consistency: Do translated prompts generate similar images?
 - Code-Switching Similarity: Can the model understand mixed-language prompts?

From GenEval

Six object-focused tasks to evaluate compositional image properties

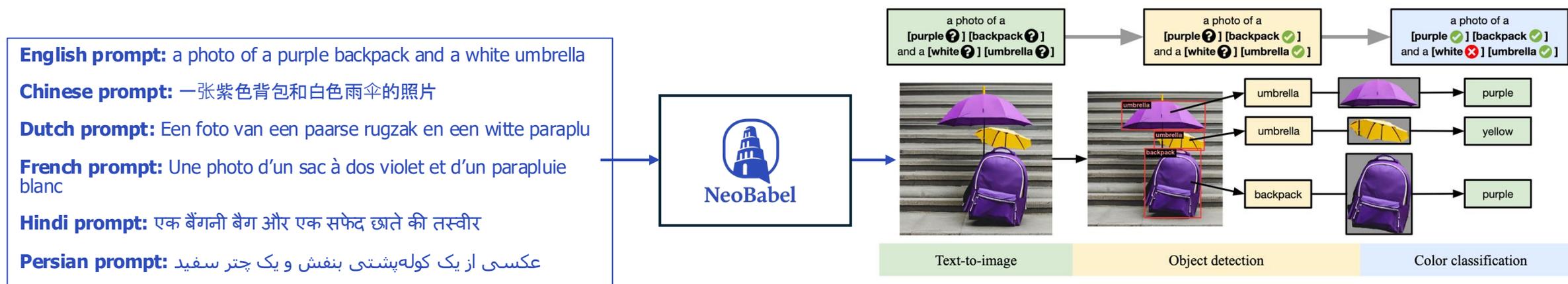


For each short prompt (~ 100 per task), paired with objects and color:

- Modern object detectors **verify** object presence, count, and position.
- Discriminative models **assess** fine-grained properties such as color.
- Final-score = verify whether image contains all items, average per task.

From GenEval to m-GenEval

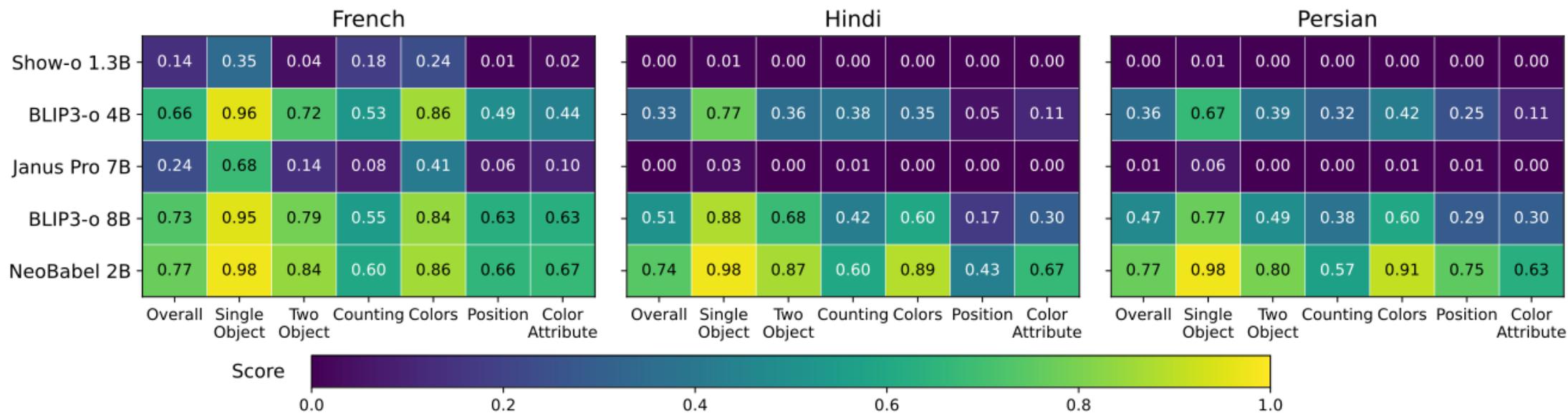
Six object-focused tasks to evaluate compositional image properties



For each short prompt (~100 per task), paired with objects and color:

- Modern object detectors **verify** object presence, count, and position.
- Discriminative models **assess** fine-grained properties such as color.
- Final-score = verify whether image contains all items, average per task and language.

m-GenEval comparison



Janus Pro and BLIP3-o rely on multilingual LLMs but are trained on English-only data, leading to a sharp performance drop in non-English languages.

NeoBabel maintains strong and consistent results across all tested languages.

Qualitative results for Dutch

One object

Een foto van een bankje



two objects

Een foto van een eettafel en een beer



counting

Een foto van vier honden



colors

Een foto van een paarse wortel



position

Een foto van een rugzak rechts van een sandwich



color attribute

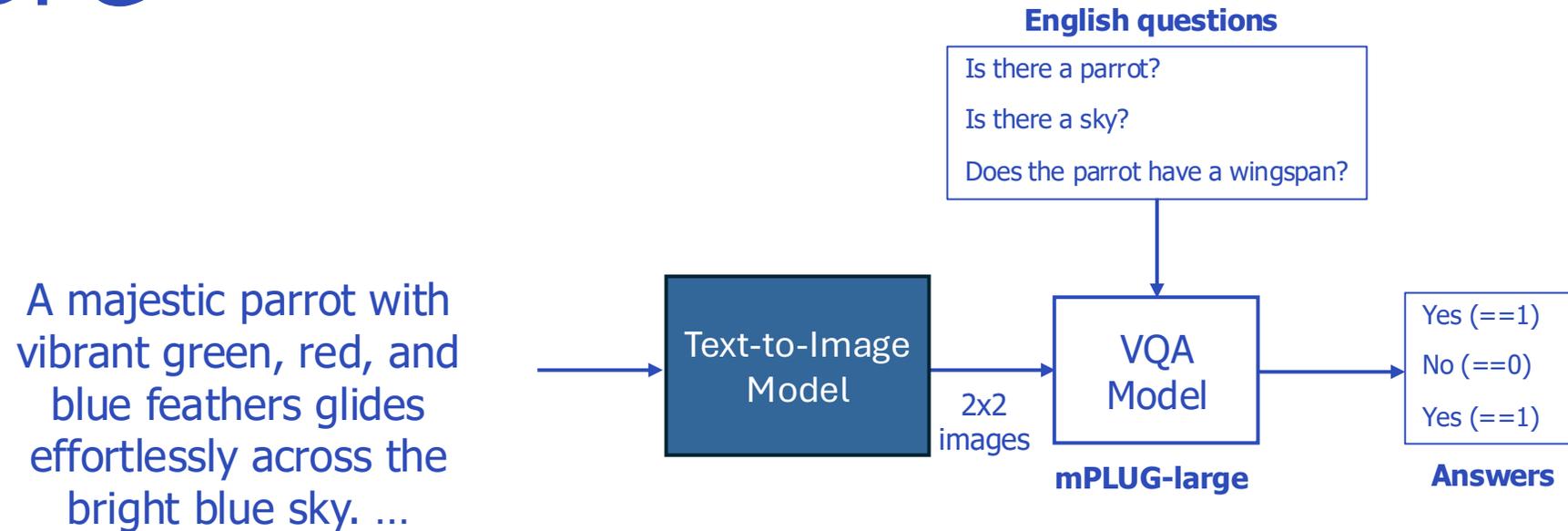
Een foto van een paars wijnglas en een zwarte appel



BLIP3-o



From DPG



For each dense prompt (out of 1K), 4 images are generated in a grid, paired with natural-language questions.

A VQA model answers each question; logical constraints filter valid responses.

Final score = average of valid question scores across all selected images.

From DPG to m-DPG

Multilingual Prompts

English prompt: A majestic parrot with vibrant green, red, and blue feathers glides effortlessly across the bright blue sky. ...

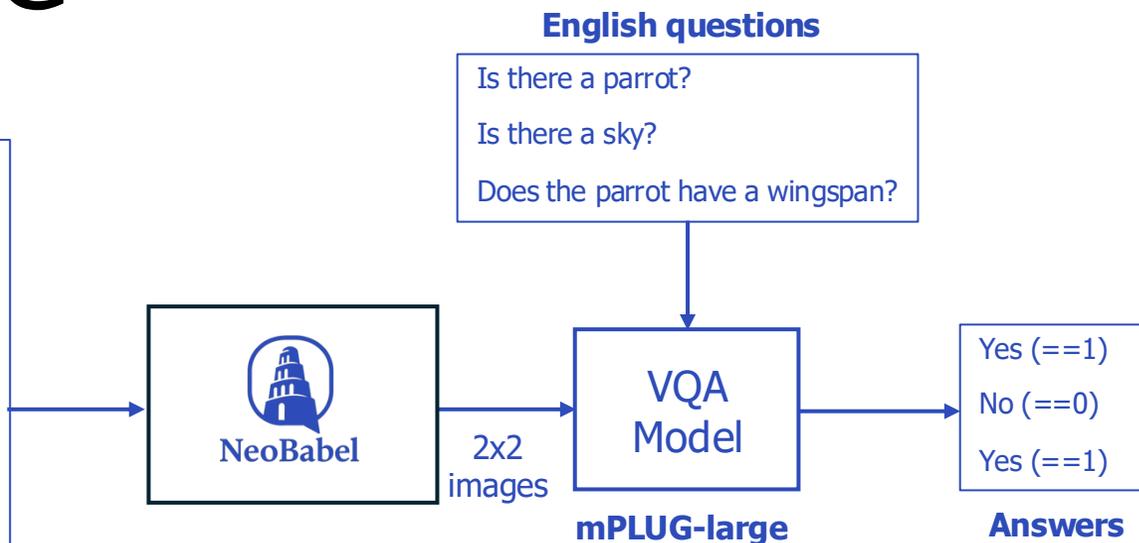
Chinese prompt: 一只雄伟的鹦鹉，长着鲜艳的绿色、红色和蓝色的羽毛，在蔚蓝的天空中轻松滑翔。...

Dutch prompt: Een majestueuze papegaai met levendige groene, rode en blauwe veren glijdt moeiteloos door de helderblauwe lucht. ...

French prompt: Un perroquet majestueux aux plumes vertes, rouges et bleues vibrantes plane sans effort dans le ciel bleu vif. ...

Hindi prompt: एक शानदार तोता, जिसके पंख जीवंत हरे, लाल और नीले रंग के हैं, चमकीले नीले आकाश में अनायास ही उड़ता है। ...

Persian prompt: یک طوطی باشکوه با پرهای سبز، قرمز و آبی پر... بدون زحمت در آسمان آبی روشن سر می خورد.



For each dense prompt (out of 1K), 4 images are generated in a grid, paired with natural-language questions.

A VQA model answers each question; logical constraints filter valid responses.

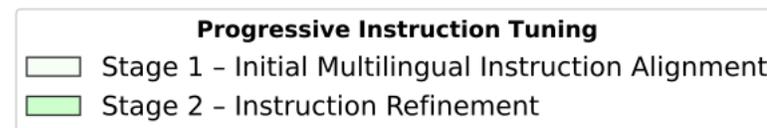
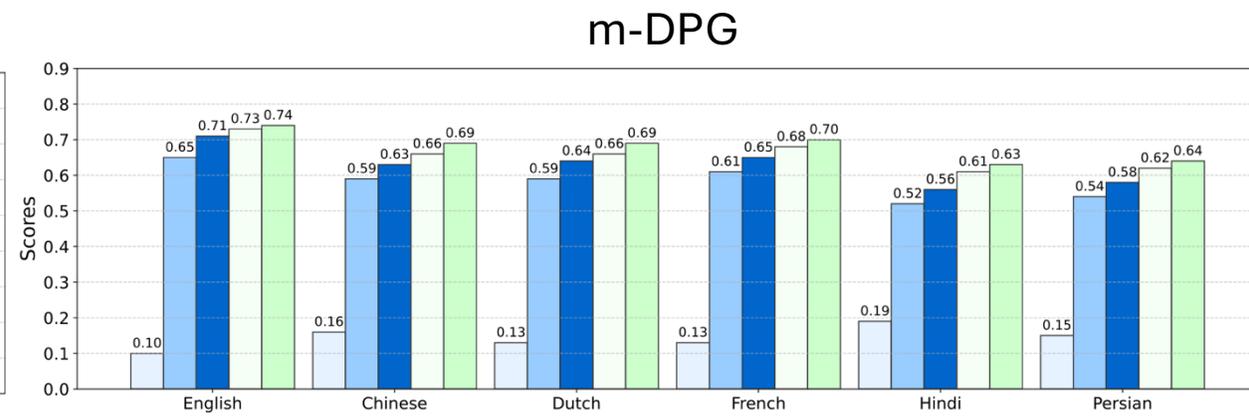
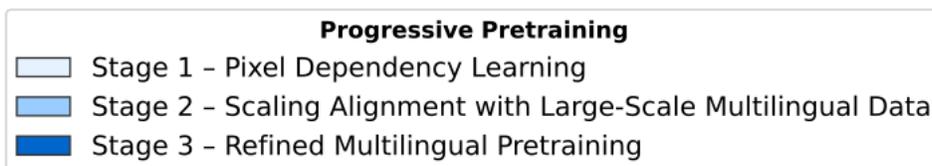
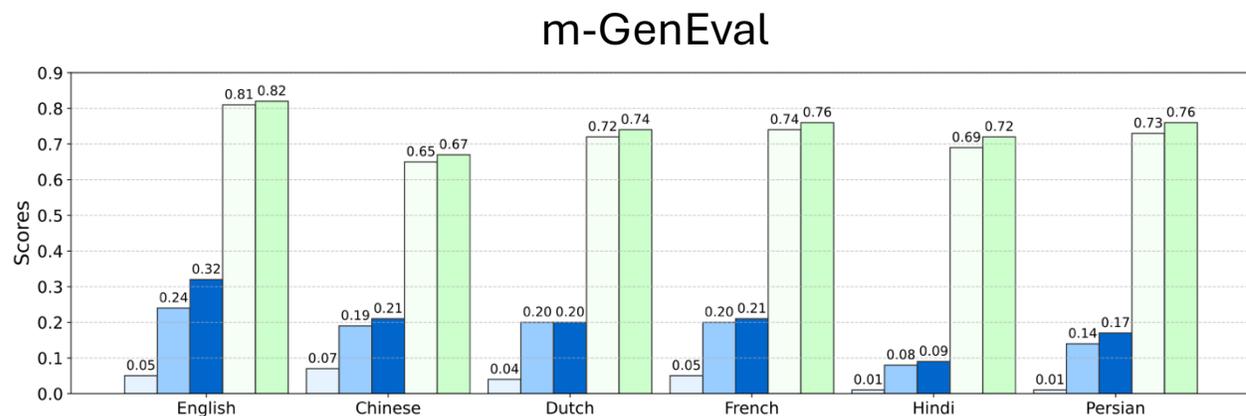
Final score = average of valid question scores across all selected images, per language.

m-DPG comparison

Model	Params.	English	Chinese	Dutch	French	Hindi	Persian	Overall
Show-o	1.3B	0.67	0.10	0.22	0.32	0.04	0.04	0.23
EMU3	8B	0.80	–	–	–	–	–	-
TokenFlow-XL	14B	0.73	–	–	–	–	–	-
Janus	1.3B	0.79	0.56	0.42	0.53	0.17	0.13	0.43
Janus Pro	7B	0.84	0.50	0.61	0.68	0.12	0.12	0.47
BLIP3-o	4B	0.79	0.60	0.58	0.59	0.47	0.49	0.58
BLIP3-o	8B	0.80	0.56	0.59	0.61	0.50	0.53	0.59
NEOBABEL	2B	0.75	0.70	0.69	0.70	0.63	0.65	0.68

Despite small parameter count, NeoBabel achieves competitive results in English and consistently outperforms alternatives across five non-English languages.

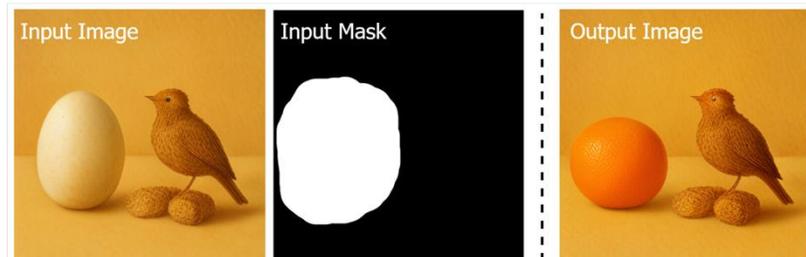
Effect of training stages



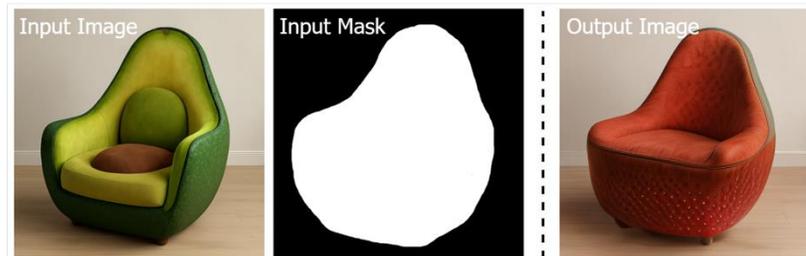
Multilingual image data quality and resolution enable fine-grained improvements, scale drives generalization.

Application

Multilingual Inpainting



Input Prompt in English: An orange and a bird made of wheat bread.



Input Prompt in Chinese: 草莓形状的扶手椅。



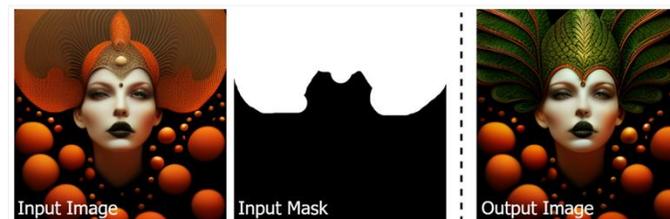
Input Prompt in Dutch: Een glas witte wijn op een spiegelend oppervlak.



Input Prompt in French: Cette œuvre d'art surréaliste et extrêmement détaillée représente le profil d'un homme à la barbe fournie et aux traits ornés d'éléments organiques complexes. Sa peau semble ornée de textures variées, rappelant des feuilles, des vignes et des motifs délicats, qui se marient harmonieusement avec des motifs naturels tels que des fleurs et des papillons. L'œil est éthéré, avec des iris d'un bleu vif cernés de reflets dorés sur de longs cils. La composition utilise des couleurs vibrantes comme l'or, le bleu, le vert, le violet et des tons terreux, créant une profondeur à travers les différentes couches du tableau.



Input Prompt in Hindi: पूरी तरह खिले हुए जीवंत गुलाब का क्लोज-अप शॉट लें



Input Prompt in Persian: یک سربرند مزین با طرح‌های پیچیده شبیه پر یا فلس، عمدتاً با رنگ‌های سبز زیتونی که با تزئینات سفالی ترکیب شده‌اند

Code-Switching

Multilingual Input Prompt: A close-up photograph of a Corgi dog. De hond draagt een zwarte hoed en een ronde, donkere zonnebril. Le Corgi a une expression joyeuse, avec la bouche ouverte et la langue tirée, donnant une impression de bonheur ou d'excitation.

English Translation: A close-up photograph of a Corgi dog. The dog is wearing a black hat and round, dark sunglasses. The Corgi has a joyful expression, with its mouth open and tongue sticking out, giving an impression of happiness or excitement.



Multilingual Input Prompt: एक आत्मसंतुष्ट बिल्ली का क्लोज-अप फोटो. گریه عینک آفتابی. 它的胡须微微竖起，脸上带着微妙而狡黠的微笑，表明它认为自己比房间里的其他人都优越。

English Translation: A close-up photo of a smug cat. The cat is wearing square sunglasses. Its whiskers bristled slightly and it had a subtle, sly smile on its face, suggesting that it considered itself superior to everyone else in the room.



Open & Inclusive

<https://neo-babel.github.io>



NeoBabel: A Multilingual Open Tower for Visual Generation

Mohammad Mahdi Derakhshani², Dheeraj Varghese², Marzieh Fadaee^{1,†}, Cees G. M. Snoek^{2,†}

¹Cohere Labs, ²University of Amsterdam

[†]Principal senior advisors

 Paper

 Code

 Models

 Demo

 Pretraining Data

 Instruction Data

 Evaluation Data

Code and all data released. Models coming soon! 🕒🔥

Conclusion on NeoBabel

Multilingualism is a **catalyst**, not a trade-off.

NeoBabel is performant while being smaller and more **inclusive**.

NeoBabel **provides**: a curated dataset, a novel architecture, a progressive training strategy, and a rigorous evaluation framework.

We are releasing a **fully open** toolkit to democratize research: checkpoints, multilingual datasets, training scripts, multilingual evaluation suite.



Part II: Seeing and Reasoning

Reasoning with a Chain-of-Thought

Prompt LLM by a series of intermediate natural language **reasoning steps** to address complex arithmetic, common sense and symbolic reasoning tasks

Standard Prompting

Model Input

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?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain-of-Thought Prompting

Model Input

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?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

From a single to multiple thoughts

- No need to limit Chain-of-Thought reasoning to one solution path
- Decoder provides multiple reasoning paths
- Allows selecting most consistent one, or ensembling over answers

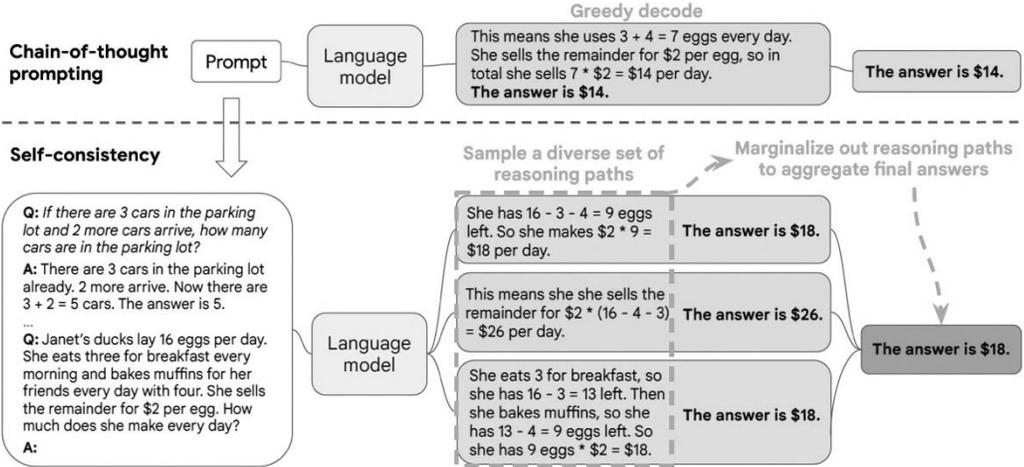
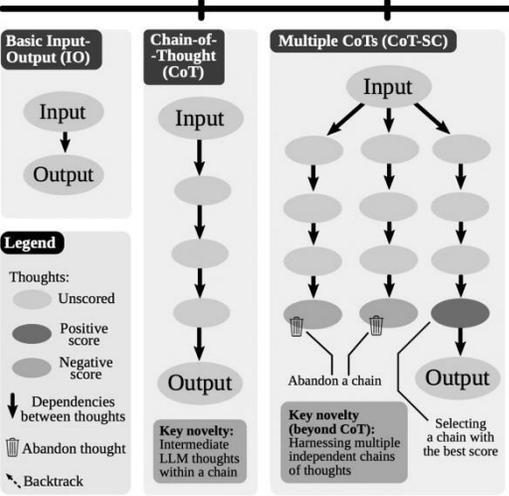


Figure by Besta et al. AAAI 2024

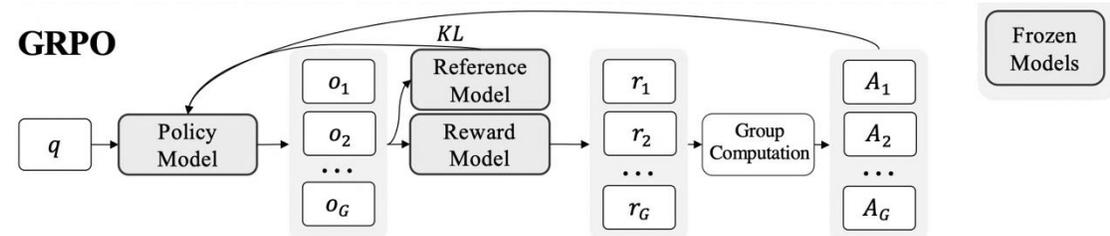
The DeepSeek 'aha moment'



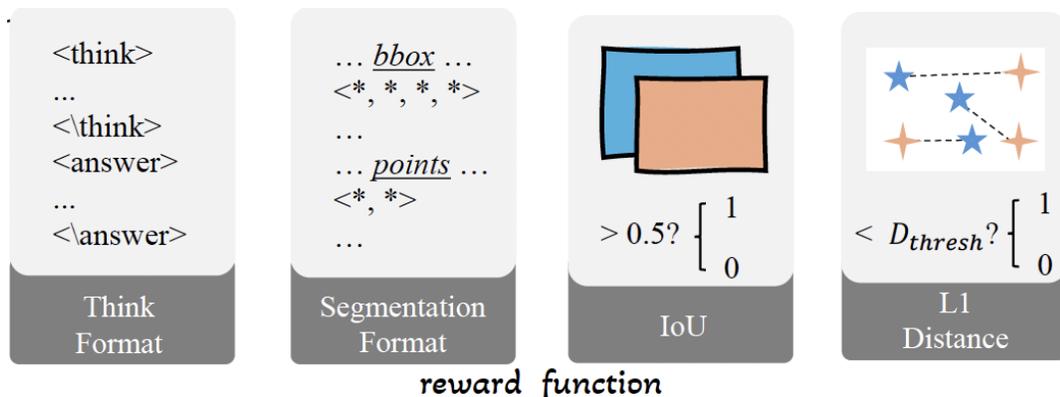
No need to supervise the reasoning process itself

GRPO-reinforcement learning, with accuracy and format reward signal

- Evaluate whether the final answer is correct
- Encapsulate reasoning process within `<think>` and `</think>` tags



Also suited for visual reasoning?



Seg-Zero: segmentation only

- Yuqi Liu et al. arXiv 3/2025

VisionReasoner: detection & segmentation

- Yuqi Liu et al. arXiv 5/2025

User: Find the older man in *a brown coat* wearing a santa hat and *holding a cell phone* in his hand.



Sample #1: **Award Score = 5**
 Think Format = 1 Seg Format = 1 IoU = 1 Bbox = 1 Points = 1

`<think>` Thought process:
 1. Identify the image elements:
 - Two people wearing Santa hats.
 - One person in a brown coat.
 ...
 3. Confirm the person in the brown coat is wearing a Santa hat.
 4. Confirm the person in the brown coat is holding a cell phone.
 5. Identify the most closely matched object in the image...`</think>`
`<answer>` {
 "bbox": [35, 60, 354, 423], "points_1": [183, 235], "points_2": [168, 305]}`</answer>`



Sample #2: **Award Score = 2**
 Think Format = 1 Seg Format = 0 IoU = 1 Bbox = 0 Points = 0

`<think>` The person in the brown coat is the one holding the cell phone and wearing a Santa hat.`</think>`
`<answer>` {
 "bbox": [35, 60, 354, 423], "bbox": [35, 60, 354, 423]}`</answer>`

Sample #3: **Award Score = 0**
 Think Format = 0 Seg Format = 0 IoU = 0 Bbox = 0 Points = 0

The people wearing a santa hat is on the right.
`<answer>` {
 "bbox": [25, 70, 300, 410]}`</answer>`



Single-turn dialogue 👍 / Multi-turn dialogue 👎

Input



Round1: Highlight the person in row 2 with glasses and head turned.



VisionReasoner

`<think>` The task is to highlight the person in row 2 with glasses and head turned. The image shows a crowd of people, and we need to identify the second row from the bottom. The person in the second row with glasses and their head turned is located towards the left side of the image. The person is wearing a suit and tie, and their head is turned slightly to the side, looking at something or someone off-camera. `<think>`

Prediction



Ground truth

Round1 Label



Single-turn dialogue 👍 / Multi-turn dialogue 👎



Round1: Highlight the person in row 2 with glasses and head turned.



VisionReasoner

<think> The task is to highlight the person in row 2 with glasses and head turned. The image shows a crowd of people, and we need to identify the second row from the bottom. The person in the second row with glasses and their head turned is located towards the left side of the image. The person is wearing a suit and tie, and their head is turned slightly to the side, looking at something or someone off-camera. </think>

Prediction



Ground truth

Round1 Label



Round2: Focus on the person who is talking to the one in R1.



VisionReasoner

<think> The task is to identify who is talking to the person in [176,102,300,389]. The reference box seems to mark someone in the crowd, but the instruction is vague. Likely the focus is the central figure in the purple suit, who appears to be the man subject and possibly the one being addressed. </think>



Round2 Label



Reference propagation across rounds is brittle, credit assignment becomes ambiguous and coordinate hallucinations are hard to detect.

Single-turn dialogue 👍 / Multi-turn dialogue 👎



Round1: Highlight the person in row 2 with glasses and head turned.



VisionReasoner

<think> The task is to highlight the person in row 2 with glasses and head turned. The image shows a crowd of people, and we need to identify the second row from the bottom. The person in the second row with glasses and their head turned is located towards the left side of the image. The person is wearing a suit and tie, and their head is turned slightly to the side, looking at something or someone off-camera. </think>

Prediction



Ground truth

Round1 Label



Round2: Focus on the person who is talking to the one in R1.



VisionReasoner

<think> The task is to identify who is talking to the person in [176,102,300,389]. The reference box seems to mark someone in the crowd, but the instruction is vague. Likely the focus is the central figure in the purple suit, who appears to be the man subject and possibly the one being addressed. </think>



Round2 Label



Round3: Recognize the person directly in front of R2.



VisionReasoner

<think> The task is to recognize the person directly in front of the woman in the purple jacket near the center foreground. The person directly in front of her appears to be a man wearing sunglasses and a dark suit, holding a phone in his hands. </think>



Round3 Label





RegionReasoner: Region-Grounded Multi-Round Visual Reasoning

Wenfang Sun*, Hao Chen*, Yingjun Du, Yefeng Zheng, and Cees G. M. Snoek



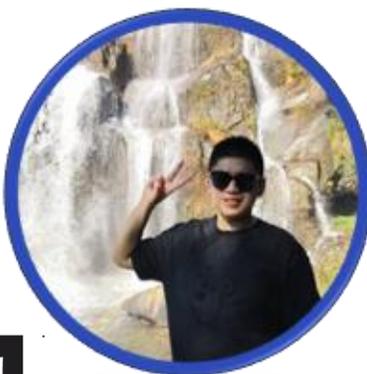
UNIVERSITY OF AMSTERDAM



安徽大學
Anhui University



WESTLAKE
UNIVERSITY



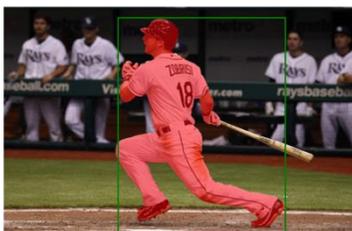
Proposal: Region-Grounded Multi-Round Reasoning



Hi RegionReasoner, please find the player wearing jersey number 18.



Output: **R1**



<scene> The image depicts a baseball game. A batter is at the center of the scene, holding a bat and wearing a white uniform. Several other players in white jerseys are standing in the dugout behind him . </scene>

<think> I first look at all the players. Most of them are in the dugout. The player **in the middle foreground** is actively batting, and his **jersey clearly shows the number 18**. </think>



RegionReasoner



Great, then please locating the standing spectator behind the **R1** on the left.



Output: **R2**



<scene> The image depicts a baseball game. A batter is at the center of the scene, holding a bat and wearing a white uniform. Several other players in white jerseys are standing in the dugout behind him . </scene>

<focus> The reference `bbox=[205, 22, 511, 412]` corresponds to the batter wearing jersey 18, holding a bat in the middle of the image. </focus>

<think> The query specifies: locate the standing spectator **behind the ref `bbox=[205, 22, 511, 412]` on the left**. I apply the spatial relation: the target must be behind the batter and shifted left. Looking into the dugout, there is a person in a white jersey **standing upright behind the batter's left side**. </think>



RegionReasoner

Reward signal I: Reference citation

To make the reasoning **verifiable and grounded**, the trace must explicitly cite the referenced boxes.



Great, then please locating the standing spectator behind the **R1** on the left.



Output: **R2**



<scene> The image depicts a baseball game. A batter is at the center of the scene, holding a bat and wearing a white uniform. Several other players in white jerseys are standing in the dugout behind him . </scene>

<focus> The reference bbox=[205, 22, 511, 412] corresponds to the batter wearing jersey 18, holding a bat in the middle of the image. </focus>

<think> The query specifies: locate the standing spectator **behind the ref bbox=[205, 22, 511, 412] on the left**. I apply the spatial relation: the target must be behind the batter and shifted left. Looking into the dugout, there is a person in a white jersey **standing upright behind the batter's left side**. </think>



We reward correct citation and penalize hallucinated coordinates.

Reward signal II: Global–local consistency

To keep the reasoning coherent with both global scene context and localized evidence, we **align reasoning with scene description and the focus**



Great, then please locating the standing spectator behind the **R1** on the left.



Output: **R2**



<scene> The image depicts a baseball game. A batter is at the center of the scene, holding a bat and wearing a white uniform. Several other players in white jerseys are standing in the dugout behind him . </scene>

<focus> The reference bbox=[205, 22, 511, 412] corresponds to the batter wearing jersey 18, holding a bat in the middle of the image. </focus>

<think> The query specifies: locate the standing spectator **behind the ref bbox=[205, 22, 511, 412] on the left**. I apply the spatial relation: the target must be behind the batter and shifted left. Looking into the dugout, there is a person in a white jersey **standing upright behind the batter's left side**. </think>



Instruction schema used during training

Instruction Schema

<image>

Task: “Please find {Question} with bboxes and points.”

Reference guidance: If a reference bbox is provided (e.g., above/ below/ to the left of/ to the right of/ inside/ overlapping with/ touching $\text{bbox}=[x1, y1, x2, y2]$), use it only as spatial guidance.

Steps: 1) In <scene> </scene>, give a concise global scene description.

2) If a reference bbox exists, in <focus> </focus> describe ONLY what is visible inside that bbox (do not output the final answer or target label here).

3) In <think> </think>, reason over the whole image by combining the global scene and the reference bbox relation. Explicitly state which spatial relation from the question you apply (e.g., “target is above the reference”), and use it to constrain the search over the scene to locate the target object(s). If multiple candidates exist, compare them and pick the closest match.

4) In <answer> </answer>, output the bbox(es) and point(s) for the target object(s) in JSON.

Format: <scene> global scene description </scene>

<focus> description of reference bbox content (if provided) </focus>

<think> reasoning that applies the spatial relation to the scene and narrows to the final target(s) </think>

<answer>{Answer}</answer>

RegionReasoner is an auto-regressive VLM policy π_θ

Initialized from Qwen2.5-VL-7B

Generates the **structured text action**: <scene>, <focus>, <think>, <answer>

$$\pi_\theta(y_t \mid I, q_t, \mathcal{B}_t^{\text{ref}}, \mathcal{M}_{t-1}) = \prod_{n=1}^{N_t} \pi_\theta(w_{t,n} \mid I, q_t, \mathcal{B}_t^{\text{ref}}, \mathcal{M}_{t-1}, w_{t,<n}).$$

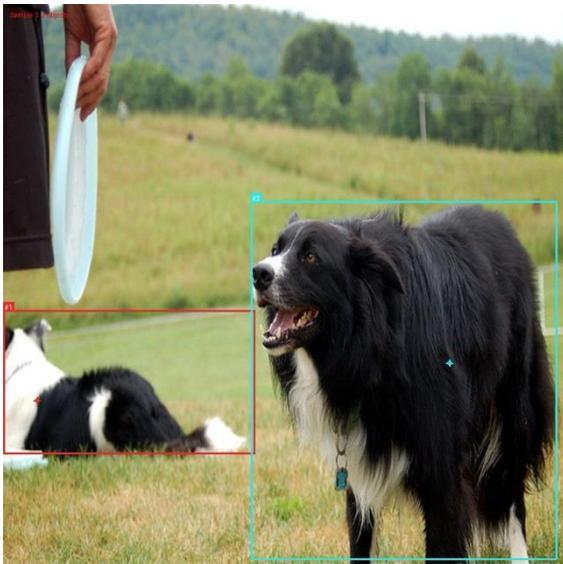
Optimize GRPO on autoregressive likelihood of the structured action, with reward:

$$R(t) = R_{\text{base}}(t) + \alpha R_{\text{ref}}(t) + \beta R_{\text{cons}}(t),$$

Benchmark

Multi-round training data construction

We extend ~7k single-turn VisionReasoner samples into ~**10k dialogue samples**, by decomposing multi-object instructions into sequential sub-queries



a lying dog and a standing dog

👤 **Round1:** Find a black and white dog laying down, looking away from the camera

💡 **label1:** bbox_2d=[0, 457, 374, 672], point_2d=[50, 592]

👤 **Round2:** Also find the *standing dog*, next to **bbox=[0, 457, 374, 672]**.

💡 **label2:** bbox_2d=[368, 294, 823, 830], point_2d=[663, 536]

- ↑ above
- ↓ below
- ← to the left of
- to the right of

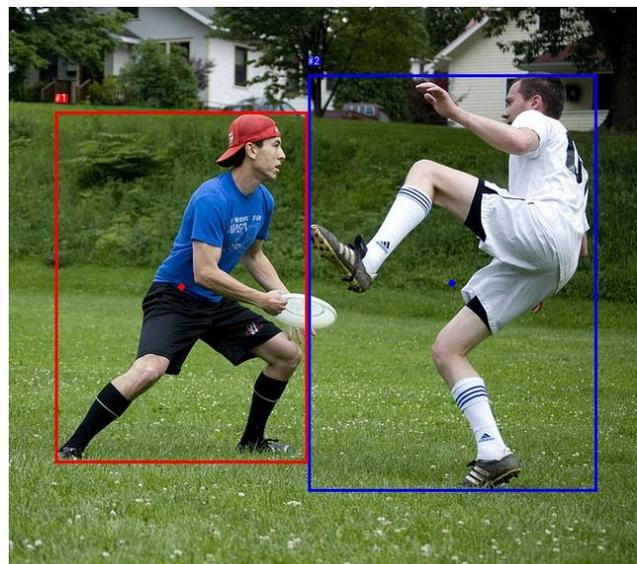
- 📦 inside
- 🗨️ overlapping
- 👉 touching
- 📍 next to

RegionDial-Bench

	Image	Round
RefCOCO+ multi round	715	2289
RefCOCOg multi round	1580	4115

First curated multi-round benchmark for reference-grounded reasoning

Grouped RefCOCO(+/g) samples manually rewritten with region references



 **Round1:** Who is wearing a blue shirt?

 **label1:**
`bbox_2d=[47, 107, 303, 466]`

 **Round2:** Who's the man next to `bbox=[47, 107, 303, 466]`?

 **label2:**
`bbox_2d=[304, 69, 597, 495]`

Results

Multi-turn detection results

Method	RefCOCO+ Multi-turn (AP \uparrow)							Avg
	R1	R2	R3	R4	R5	R6	R7	
Qwen2-VL-7B	6.2	8.5	6.5	5.4	7.5	3.6	3.5	6.7
Qwen2.5-VL-7B	65.5	49.0	48.1	36.5	30.0	38.2	25.9	49.9
Seg-Zero-7B	90.5	71.2	73.6	59.6	48.8	58.2	48.2	73.1
VisionReasoner-7B	88.3	74.7	75.8	64.2	56.3	57.3	47.0	74.8
RegionReasoner -7B	89.3	83.2	81.6	69.6	61.9	69.1	64.7	80.7

Same trend on RefCOCOg multi-turn.

RegionReasoner detection more stable in later rounds

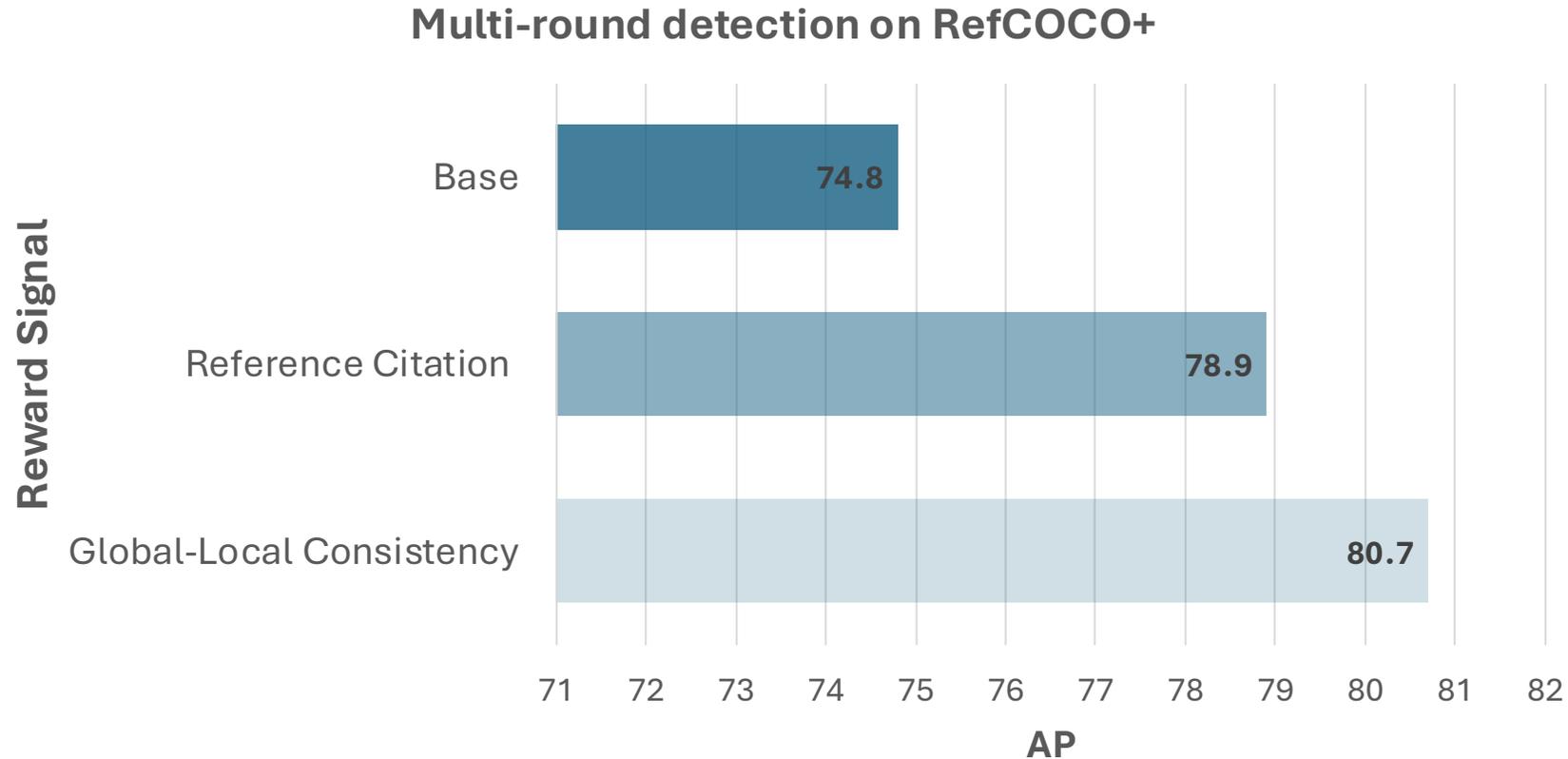
Multi-turn segmentation results

Method	RefCOCOg Multi-turn (gIoU \uparrow)							Avg
	R1	R2	R3	R4	R5	R6	R7	
Qwen2-VL-7B	8.5	11.6	8.8	10.0	7.0	6.4	4.4	9.3
Qwen2.5-VL-7B	53.8	36.3	35.5	31.6	37.3	36.8	28.6	42.1
Seg-Zero-7B	72.3	52.3	53.5	55.4	59.4	59.5	58.3	60.5
SegLLM-7B	68.9	55.3	50.5	47.7	47.3	37.8	25.4	56.7
VisionReasoner-7B	69.5	52.7	55.4	56.0	57.8	64.1	57.6	59.9
RegionReasoner -7B	73.9	62.9	60.7	58.9	64.4	66.8	63.3	66.5

Same trend on RefCOCO+ multi-turn.

RegionReasoner segmentation more stable in later rounds

Ablation of reward signal



Reference citation provides largest gain, consistency reward improves stability

Success & Failure

RegionReasoner: dialogue success



Hi RegionReasoner, who is the man in the green shirt?



Output: **R1**



<scene> The image shows a restaurant setting where several people are sitting at a table. Two pizzas are placed on plates, beer bottles are on the table, and cutlery with napkins is arranged. The background shows more people and tables. </scene>
<think> Looking at the image, there is only one person wearing a green shirt. **He is seated at the center of the image, behind a pizza.** Since no other man with a green shirt is present, the target must be this person. </think>



RegionReasoner



Which slice of pizza is **R1** about to eat?



Output: **R2**



<scene> The image shows a restaurant setting where several people are sitting at a table. Two pizzas are placed on plates, beer bottles are on the table, and cutlery with napkins is arranged. The background shows more people and tables. </scene>

<focus> Inside the reference bbox [241, 1, 472, 165], a man in a green shirt is visible. He is holding a fork, seated behind a pizza. </focus>

<think> The query asks: "Which slice of pizza is [241, 1, 472, 165] about to eat?". The reference bbox [241, 1, 472, 165] corresponds to **the man in the green shirt**. He is holding utensils and is directly positioned behind the pizza in front of him. Using the spatial relation "the target pizza is in front of the reference person," **the pizza located on the plate in front of this man is the one he is about to eat.** </think>



RegionReasoner

RegionReasoner: dialogue success



Which slice of pizza is **R1** about to eat?



Output: **R2**



<scene> The image shows a restaurant setting where several people are sitting at a table. Two pizzas are placed on plates, beer bottles are on the table, and cutlery with napkins is arranged. The background shows more people and tables. </scene>

<focus> Inside the reference bbox [241, 1, 472, 165], a man in a green shirt is visible. He is holding a fork, seated behind a pizza. </focus>

<think> The query asks: "Which slice of pizza is [241, 1, 472, 165] about to eat?". The reference bbox [241, 1, 472, 165] corresponds to the man in the green shirt. He is holding utensils and is directly positioned behind the pizza in front of him. Using the spatial relation "the target pizza is in front of the reference person," the pizza located on the plate in front of this man is the one he is about to eat. </think>



RegionReasoner



Who is the person next to **R1**?



Output: **R3**



<scene> The image shows a restaurant setting where several people are sitting at a table. Two pizzas are placed on plates, beer bottles are on the table, and cutlery with napkins is arranged. The background shows more people and tables. </scene>

<focus> Inside the reference bbox [241, 1, 472, 165], there is a man in a green shirt, sitting at the table with a fork in his hands. </focus>

<think> The query asks: "Who is the person next to [241, 1, 472, 165]?" The reference bbox [241, 1, 472, 165] corresponds to the man in the green shirt. Looking at the spatial arrangement, "next to" refers to the person seated directly beside him. The image shows a woman with long hair sitting immediately to his left. Therefore, the person next to the reference is this woman. </think>



RegionReasoner

RegionReasoner: dialogue failure



Hi RegionReasoner, please find the black bear with a white patch on the front.



Output: **R1**



```
<scene>
The image shows four stuffed teddy bears sitting in a row on chairs outdoors, each wearing different clothes.
</scene>
<think>
The image shows multiple stuffed bears. The black bear with a white patch on its front is unique among them, which identifies the target.
</think>
```



RegionReasoner



Where is the gray plush toy on the left of **R1**?



Output: **R2**



```
<scene>
The image shows four stuffed teddy bears sitting in a row on chairs outdoors, each wearing different clothes.
</scene>

<focus>
The reference bbox [416, 88, 628, 338] contains a black bear with a white patch on its front.
</focus>

<think>
The question asks to find the gray plush toy located to the left of the reference bbox [308, 0, 541, 147]. The reference bbox [308, 0, 541, 147] corresponds to a black bear. The plush toy positioned farthest to its left is clearly gray, and therefore matches the description.
</think>
```



RegionReasoner

Conclusion on RegionReasoner

RegionReasoner: An RL framework for multi-round visual reasoning that produces interpretable, region-grounded reasoning trajectories.

Training Signals: Uses reference–citation grounding and global–local consistency rewards to align reasoning with cited regions and captions.

Benchmark & Results: Introduces RegionDial-Bench and shows consistent gains on RefCOCO+ and RefCOCOg, especially in deeper reasoning rounds.

Overall conclusion

World understanding needs **vision, language and reasoning interplay**

Obvious next step for AI is to start **acting** in a **multi-sensory video** world

Thank you

谢谢

Bedankt

Merci

धन्यवाद

متشکرم