ELLIOT

NeoBabel: A Multilingual Open Foundation Model for Visual Generation

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

Cohere Labs

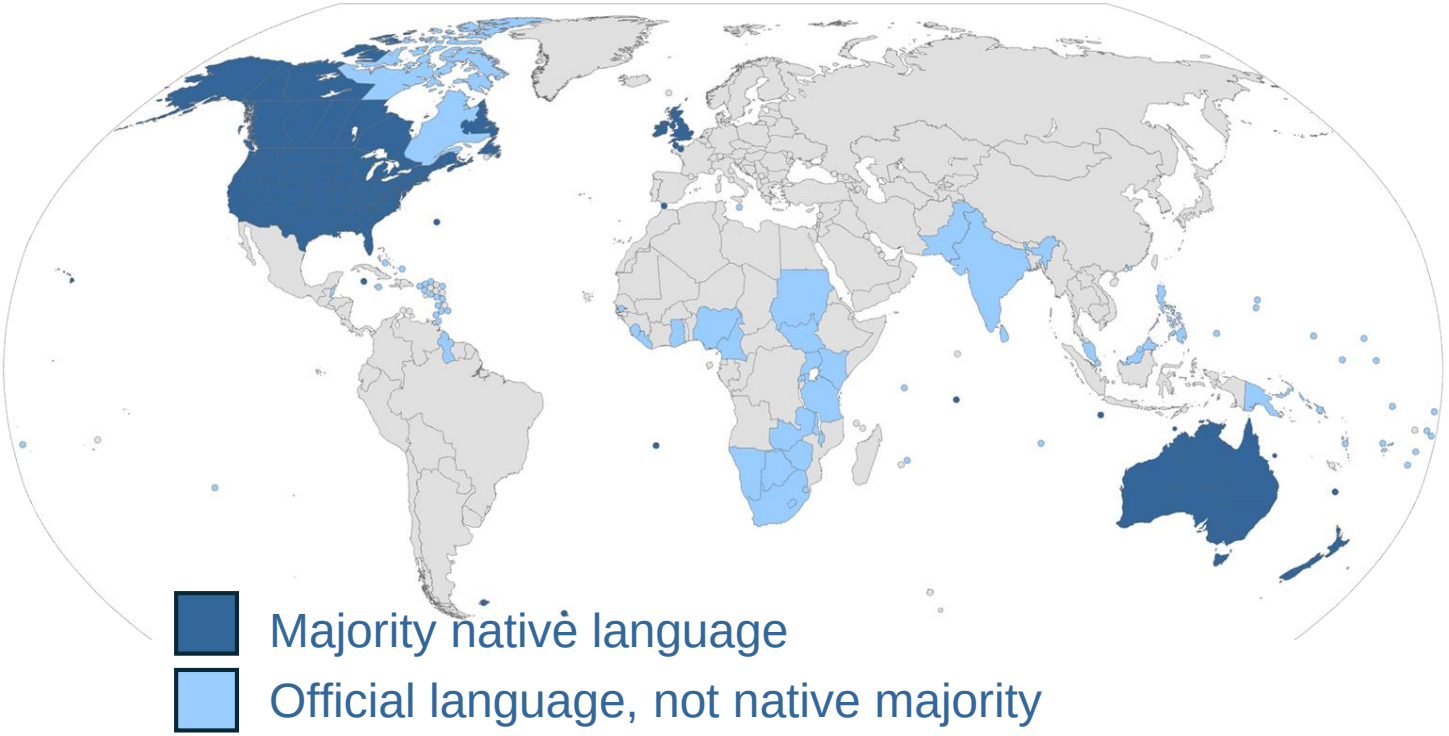
University of Amsterdam

Funded by the European Union

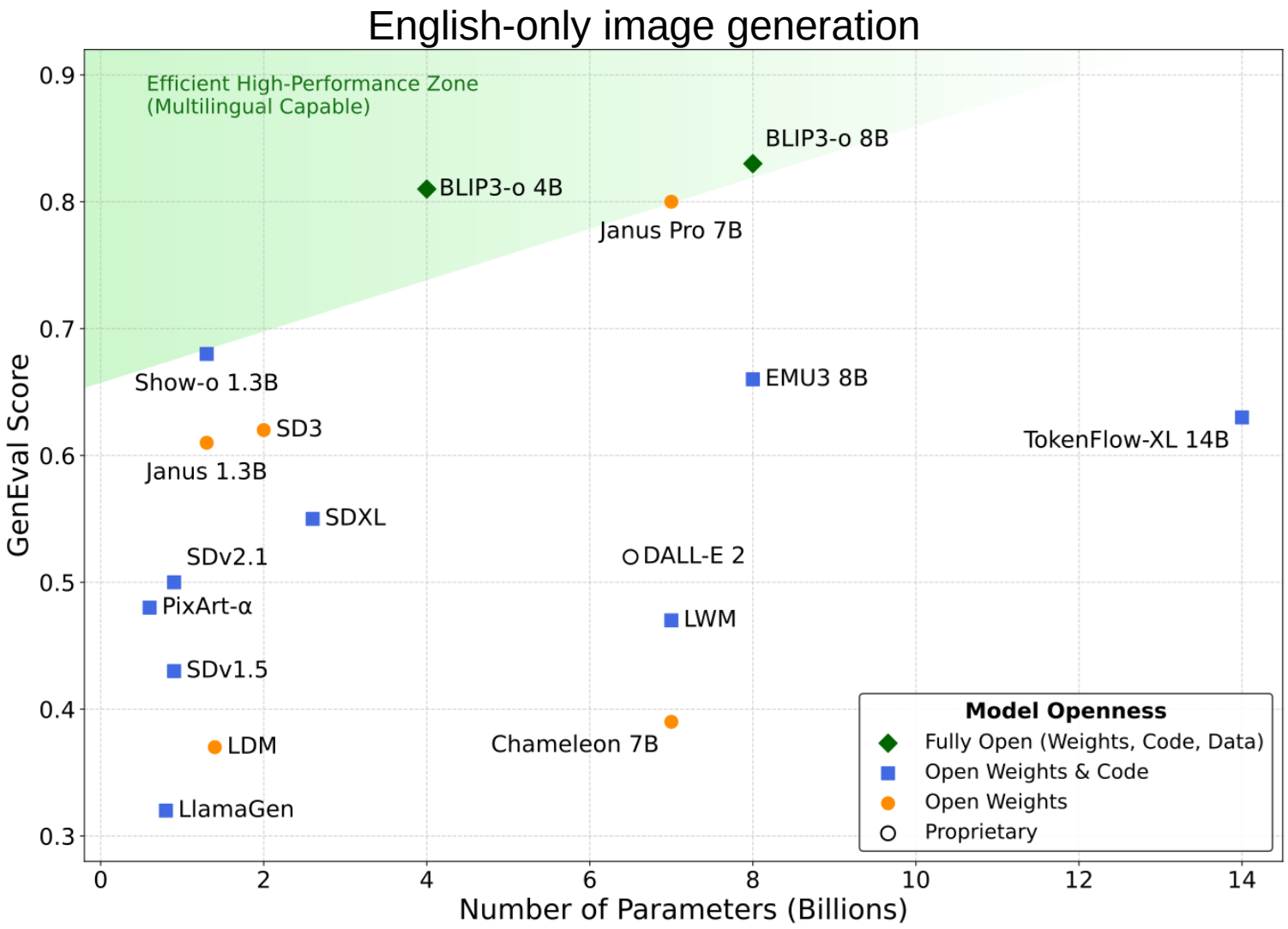
Problem: A monolingual AI world

Text-to-Image generation is English-centric.

- State-of-the-art models serve **~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



Field moves fast

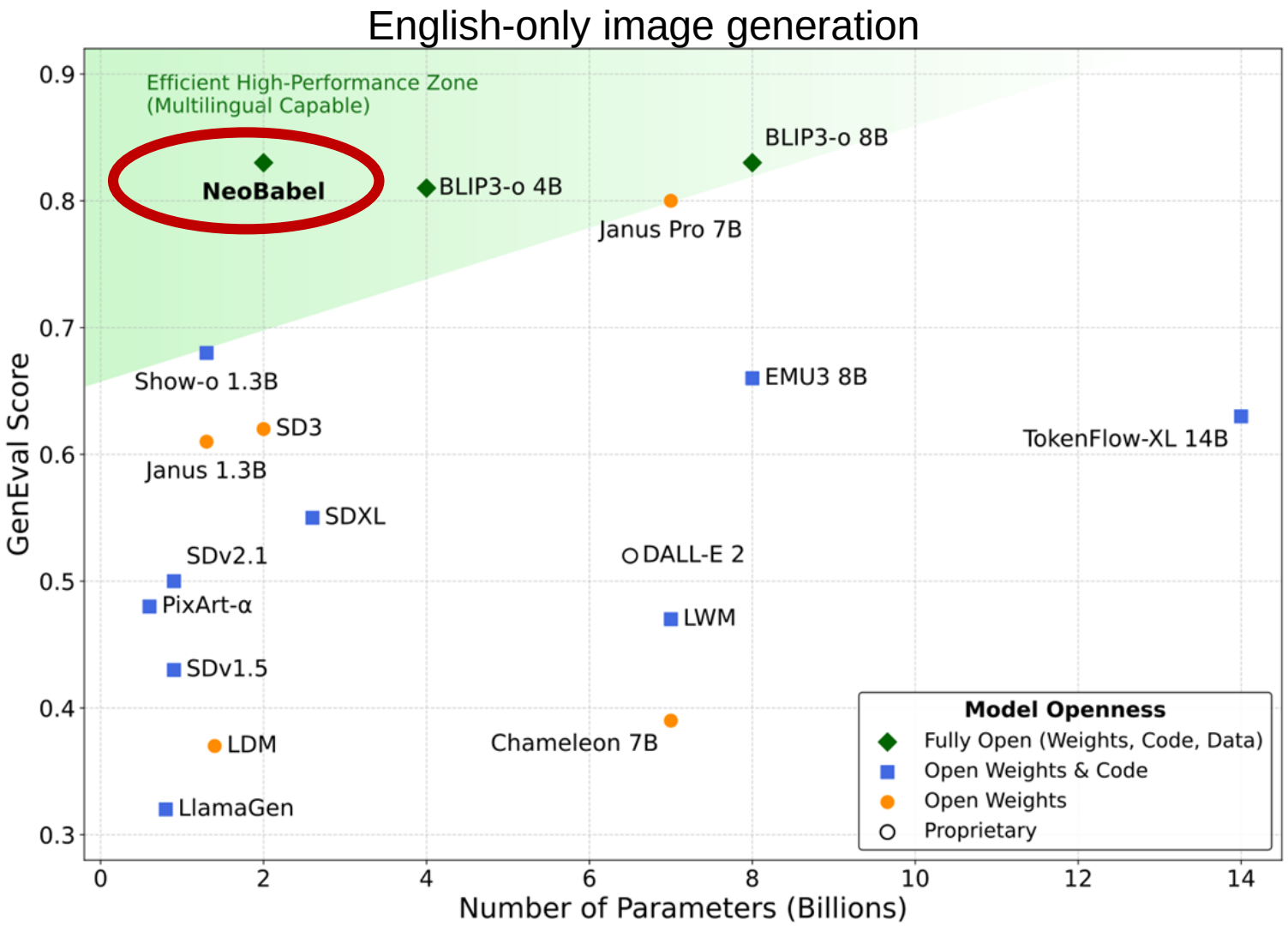
Almost all models English-only

Some rely on multilingual LLM

No native multilinguality

Limited openness

NeoBabel: Natively multilingual, efficient & open



NeoBabel is SOTA for English

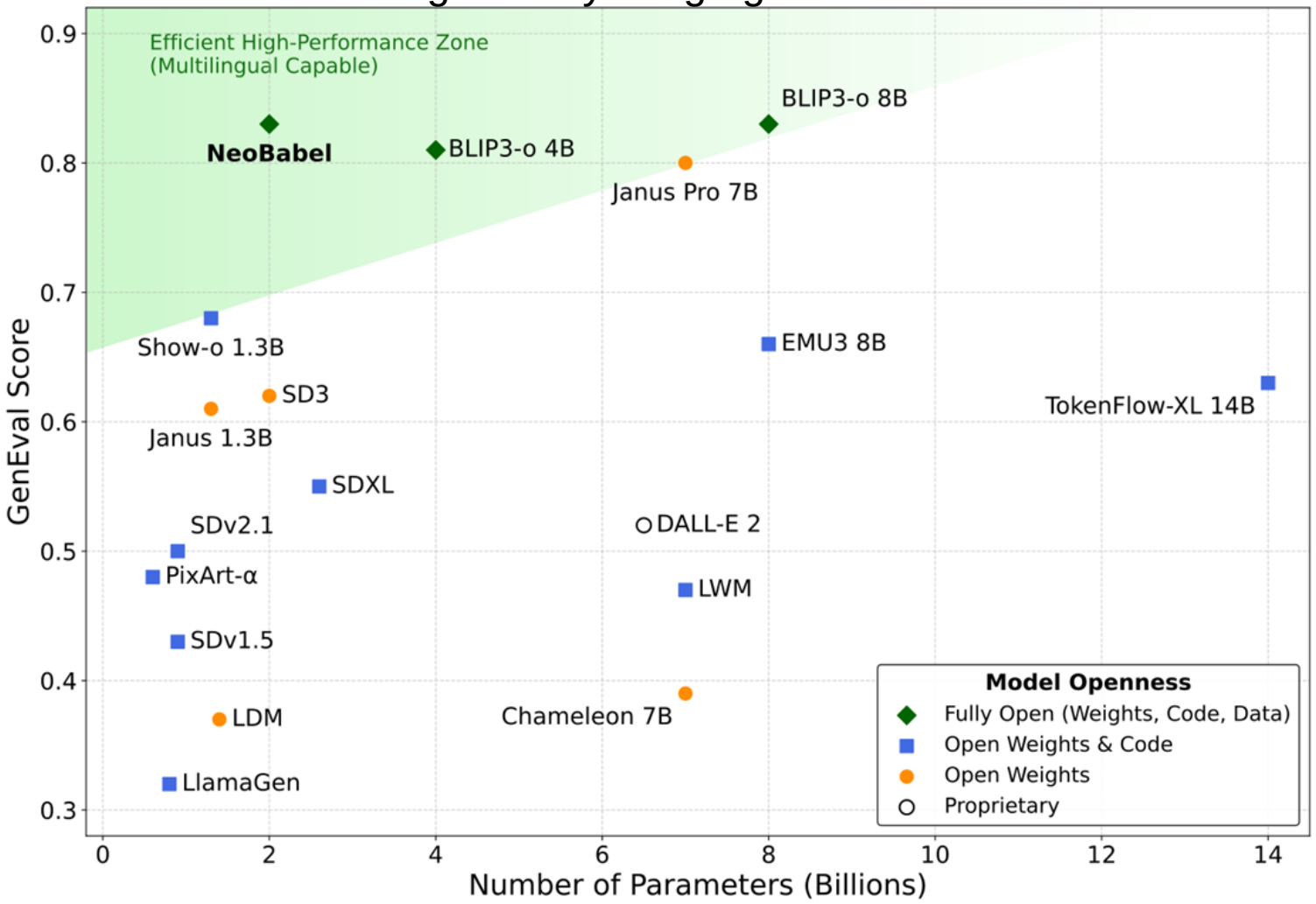
Model size 2-4x smaller

Fully open

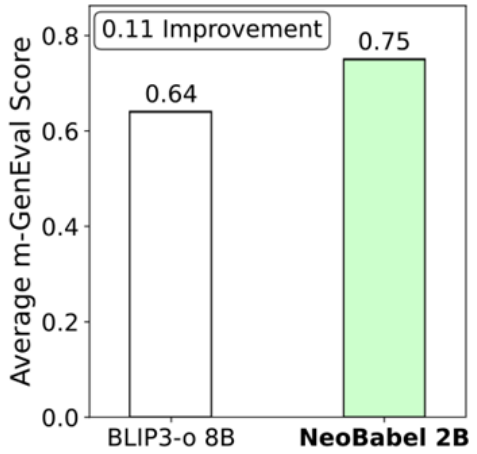
Native support for six languages

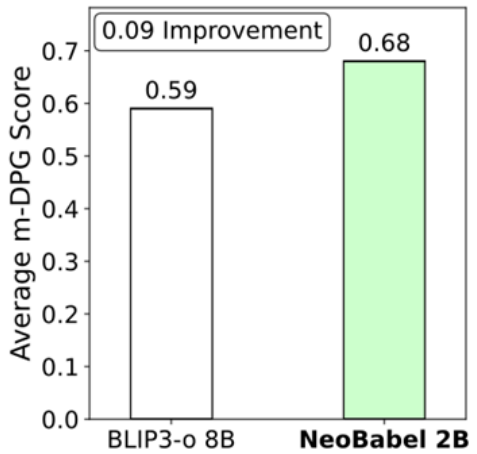
NeoBabel: Natively multilingual, efficient & open

English-only image generation



Multilingual image generation





English

Chinese

Dutch

French

Hindi

Persian

NeoBabel contributions

First natively multilingual text to image generation framework.

- 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
Chinese caption: 有轨电车
Dutch caption: tram
French caption: tramway
Hindi caption: स्ट्रीटकार
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着,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 हैमबर्गर आदमी की समुद्र तट पर स्टिल्ट्स पर चलते हुए, समृद्ध विवरण और नरम प्रकाश के साथ।

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

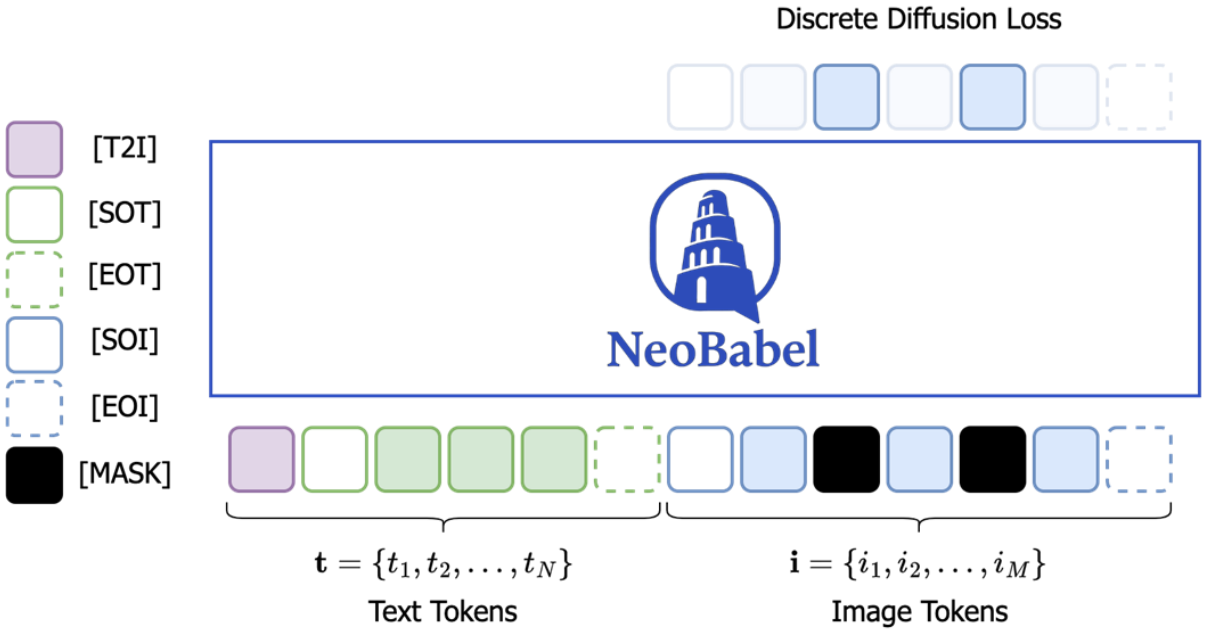
NeoBabel multilingual dataset

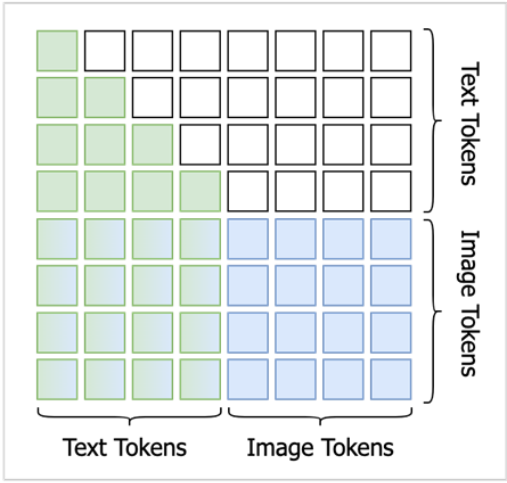
We increase the total size from 39M to 124M image–caption/label pairs

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 to Image Attention Pattern

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:

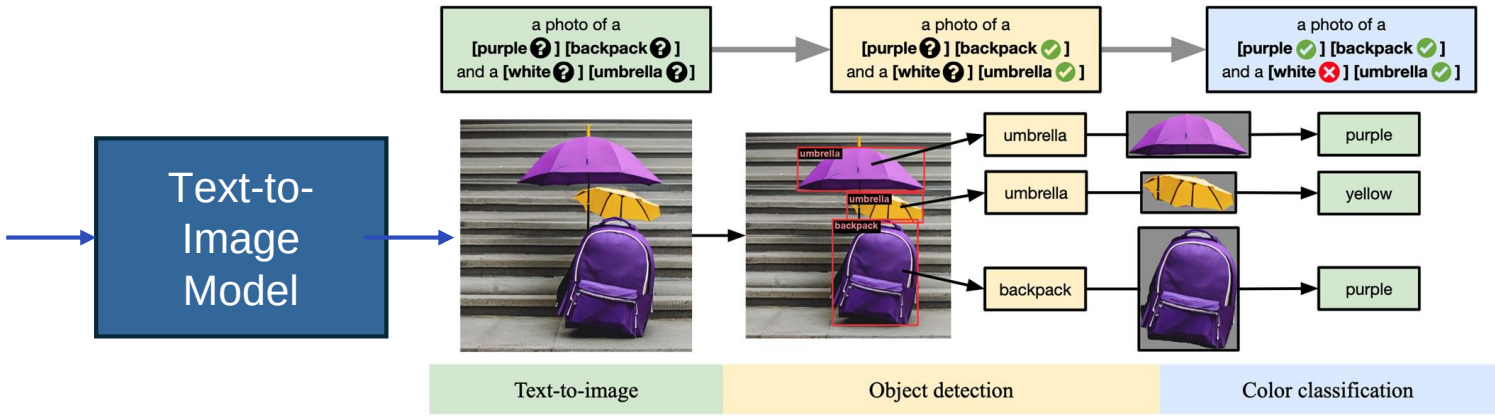
- **Multilingual extensions** of GenEval & DPG to 6 languages.

- **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

a photo of a purple backpack and a white umbrella



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

English prompt: a photo of a purple backpack and a white umbrella

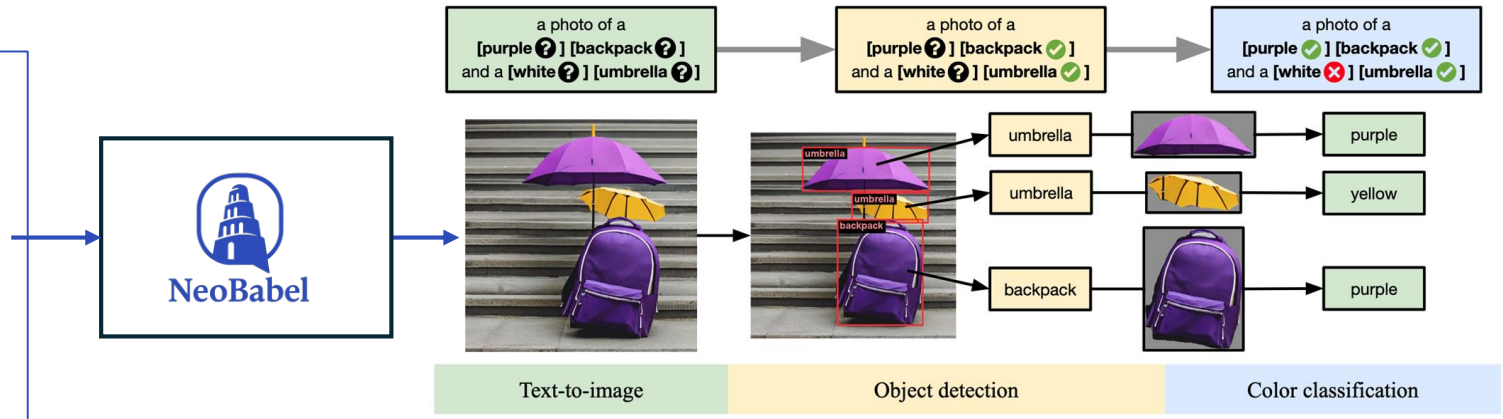
Chinese prompt: 一张紫色背包和白色雨伞的照片

Dutch prompt: Een foto van een paarse rugzak en een witte paraplu

French prompt: Une photo d'un sac à dos violet et d'un parapluie blanc

Hindi prompt: एक बैंगनी बैग और एक सफेद छाते की तस्वीर

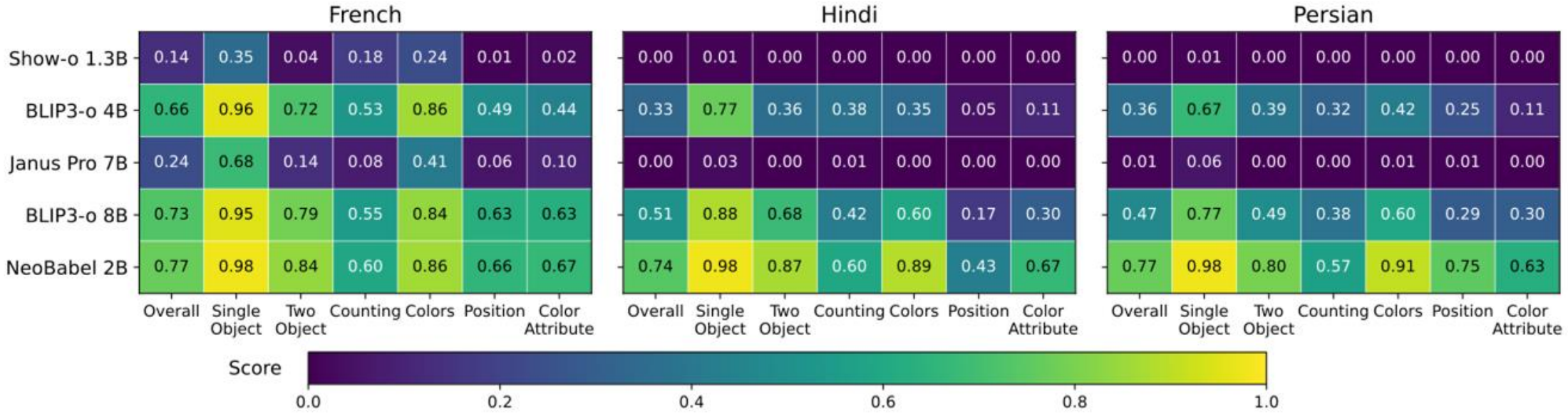
Persian prompt: عکسی از یک کوله‌پشتی بنفش و یک چتر سفید



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.

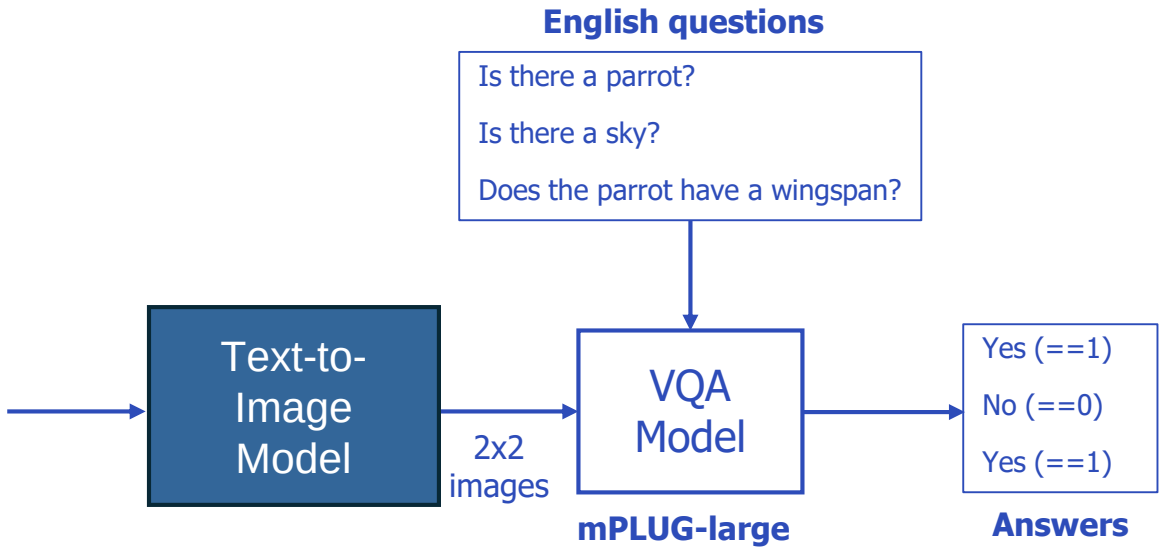
Qualitative results for Dutch

	One object	two objects	counting	colors	position	color attribute
	Een foto van een bankje	Een foto van een eettafel en een beer	Een foto van vier honden	Een foto van een paarse wortel	Een foto van een rugzak rechts van een sandwich	Een foto van een paars wijnglas en een zwarte appel
BLIP3-o						
NeoBabel						

23

From DPG

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

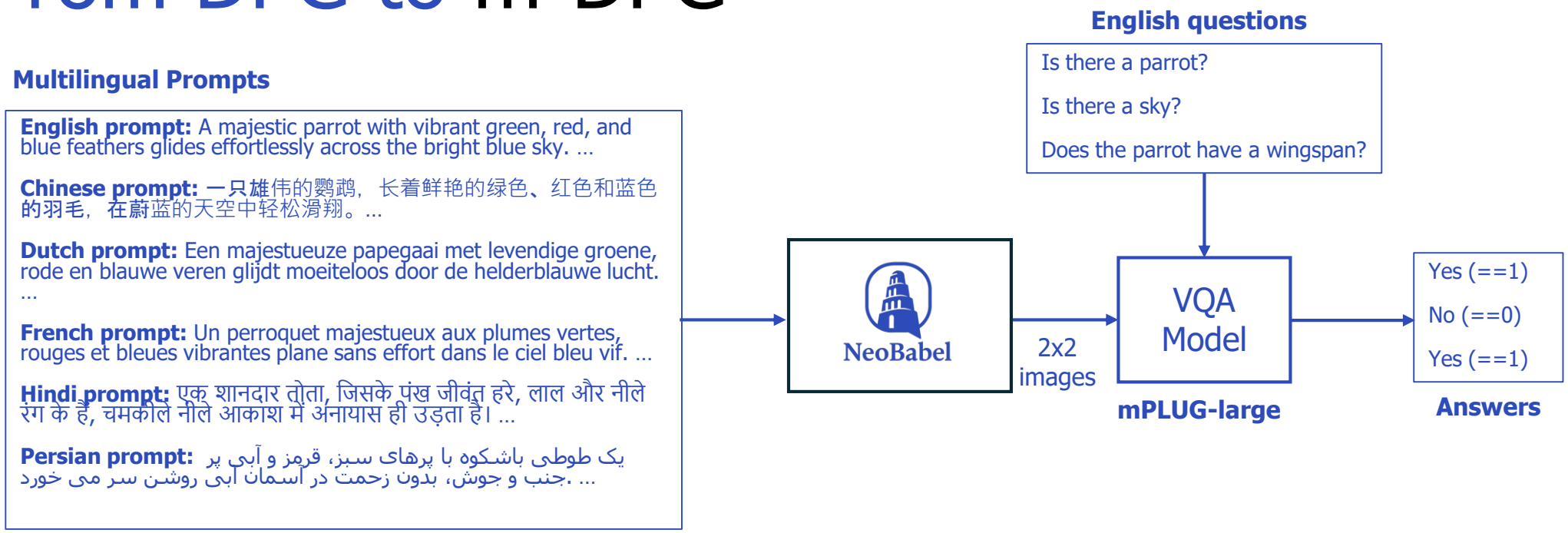


For each dense prompt (out of 1K), 4 images are generated in a grid, paired with natural-language questions.

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



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.

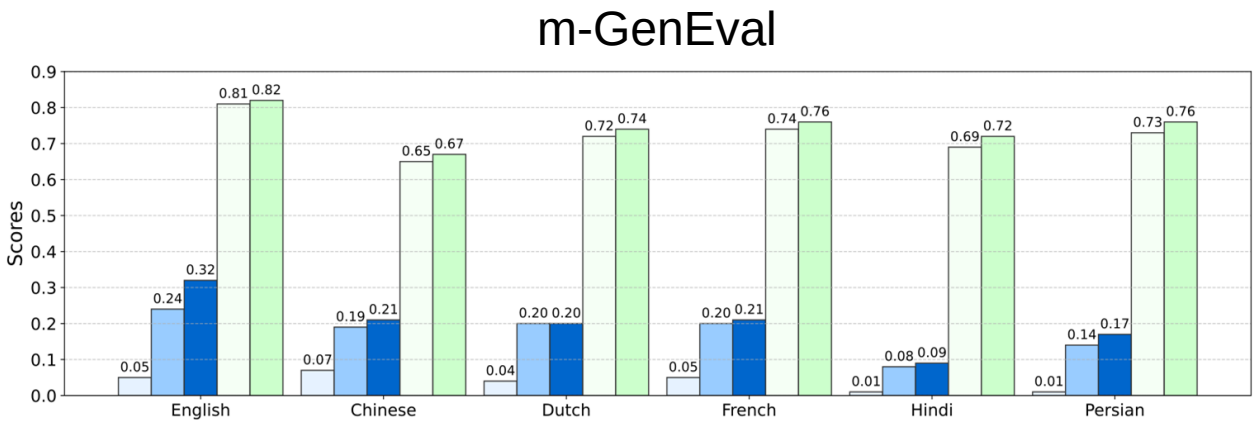
Final score = average of valid question scores across all selected images, per language.

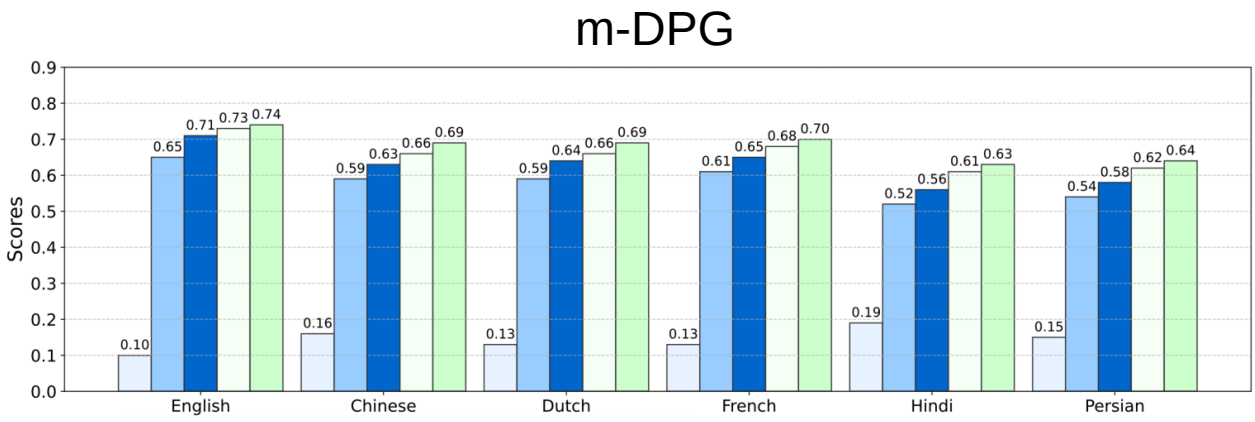
m-DPG comparison

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





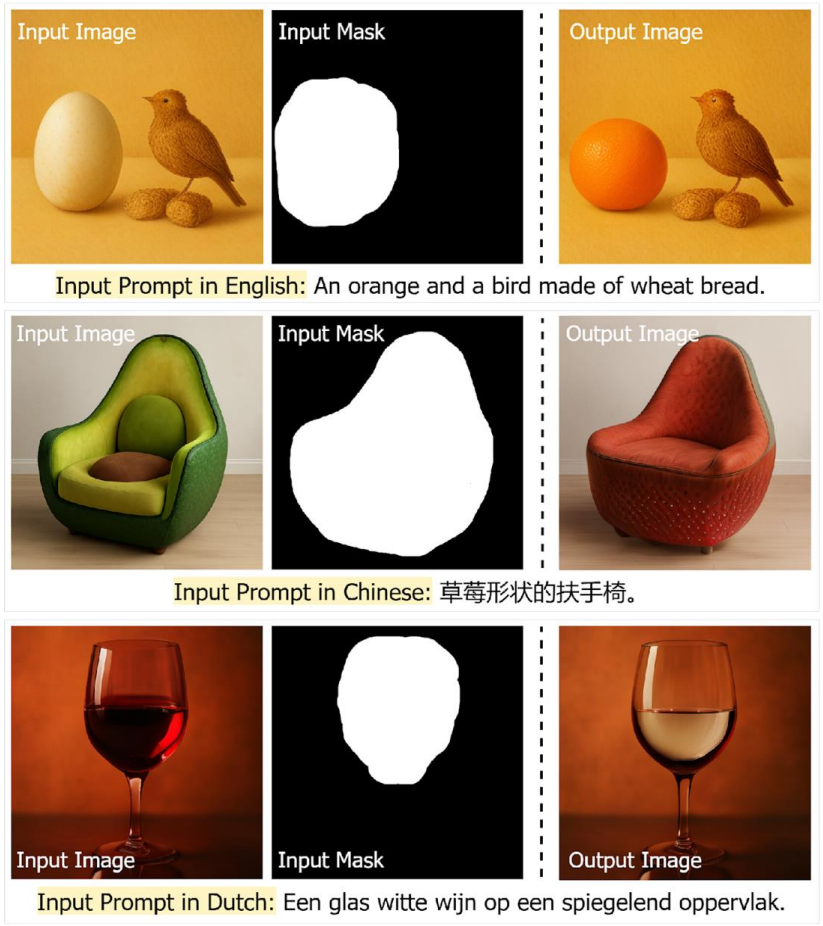
Progressive Pretraining
- Stage 1 – Pixel Dependency Learning
- Stage 2 – Scaling Alignment with Large-Scale Multilingual Data
- Stage 3 – Refined Multilingual Pretraining

Progressive Instruction Tuning
- Stage 1 – Initial Multilingual Instruction Alignment
- Stage 2 – Instruction Refinement

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

Application

Multilingual Inpainting



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. گربه عینک آفتابی مربعی به چشم دارد. 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.

English translations below each prompt for audience convenience, not used as input.

Open & Inclusive

https://neo-babel.github.io

NeoBabel: A Multilingual Open Tower for Visual Generation

Mohammad Mahdi Derakhshani[2], Dheeraj Varghese[2], Marzieh Fadaee[1,†], Cees G. M. Snoek[2,†]

[1]Cohere Labs, [2]University of Amsterdam

[†]Principal senior advisors

Paper Code Models

Pretraining Data Instruction Data Evaluation Data

More details will be released soon!

Conclusion

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.

Acknowledgements

- **Cohere Labs** for valuable feedback and for providing generous computing resources for conducting and analyzing our experiments.
- Dutch Research Council (**NWO**) in The Netherlands for awarding this project access to the **LUMI** supercomputer, owned by the **EuroHPC** Joint Undertaking, hosted by **CSC** (Finland) and the LUMI consortium through the Computing Time on National Computer Facilities call.
- **NWO** for providing access to **Snellius**, hosted by **SURF** through the Computing Time on National Computer Facilities call for proposals.
- Cees Snoek is (partially) funded by the Horizon Europe project **ELLIOT** (GA No. 101214398)