Multimodal Data Foundation at Industry Scale

Hu Xu and Shang-Wen Li

Research Scientist FAIR, Meta

Meta AI

About Us

- Research Scientist, FAIR, Meta.
- Foundational Data Research
- Leads Meta CLIP, VideoCLIP etc.
- Foundation for Llama, DINO, Perception Encoder, SAM 3, Web-SSL, Smart Glasses etc.

Motivation

- Share with the community our observations and insights on data.
- Why data matters as a foundation for research.

Foundation for Research and Production at Meta

Meta CLIP

Foundation for Research and Production at Meta

MLLM

Meta CLIP

Foundation for Research and Production at Meta

Llama 3

Meta CLIP

Foundation for Research and Production at Meta

Llama 3

Segmentation

Meta CLIP

Foundation for Research and Production at Meta

Llama 3

SAM 3

Meta CLIP

Foundation for Research and Production at Meta

Meta CLIP

Llama 3

SAM 3

Vision Encoding

Foundation for Research and Production at Meta

Meta CLIP

Llama 3

SAM 3

DINO/Perception Encoder

Foundation for Research and Production at Meta

Meta CLIP

Llama 3

SAM 3

DINO/Perception Encoder

Video Generation

Foundation for Research and Production at Meta

Meta CLIP

Llama 3

SAM 3

DINO/Perception Encoder

MovieGen

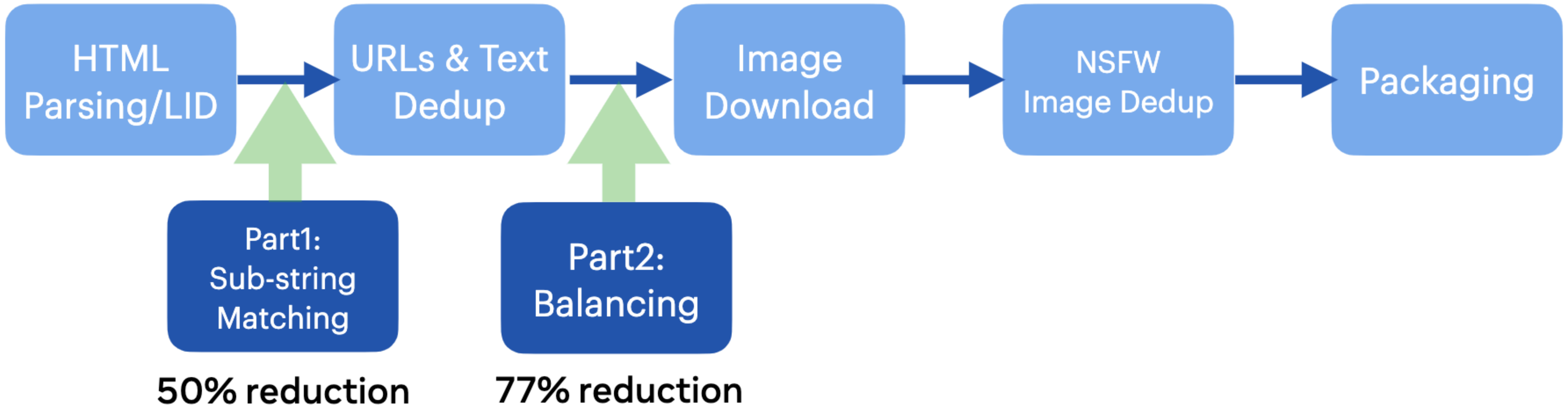
Foundation for Research and Production at Meta

Meta CLIP

- Llama 3
- SAM 3
- DINO/Perception Encoder
- MovieGen
- Recommendation

Foundation for Research and Production at Meta

Meta CLIP



Data pipeline built from scratch, processing 100B+ scale image-text pairs.

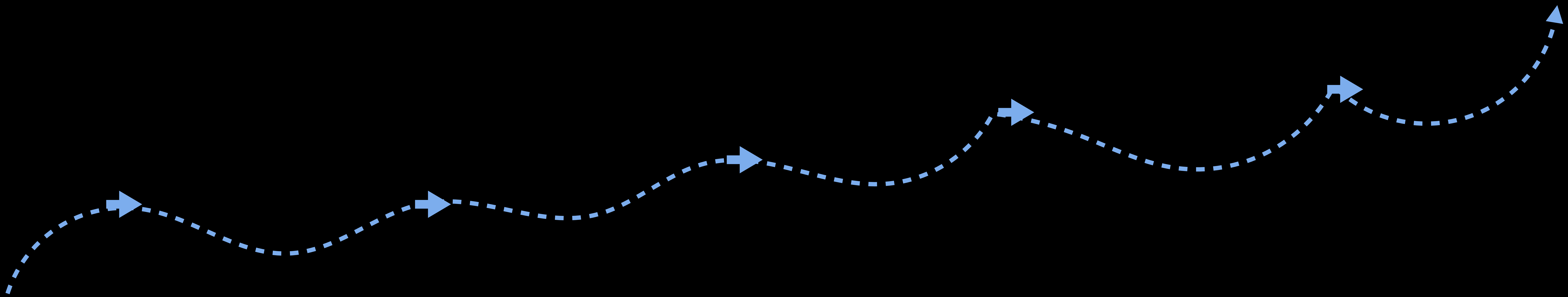
Outline

- Data, Supervision and Bottleneck
- Meta CLIP
- Meta CLIP 2
- Future Bottlenecks (Our Estimation)

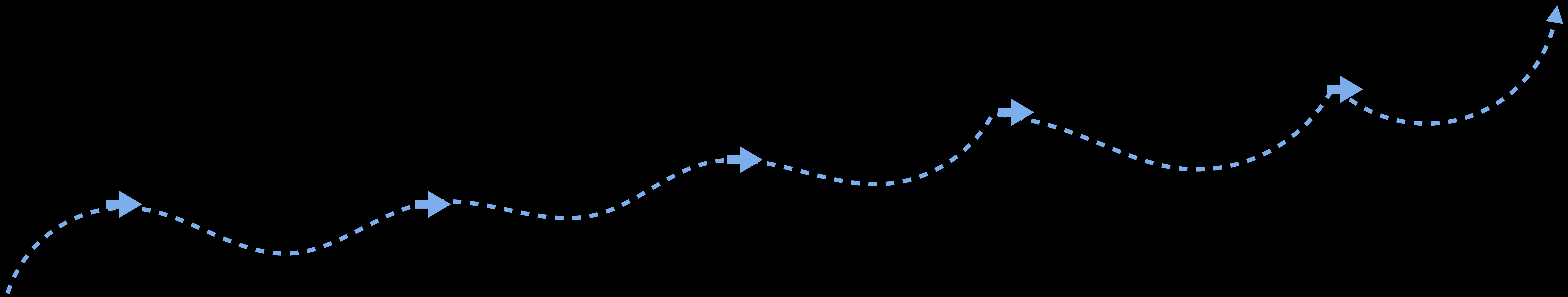
01 Data, Supervision and Bottleneck

What is Data?

History of ALL Processes Ordered by Timestamp



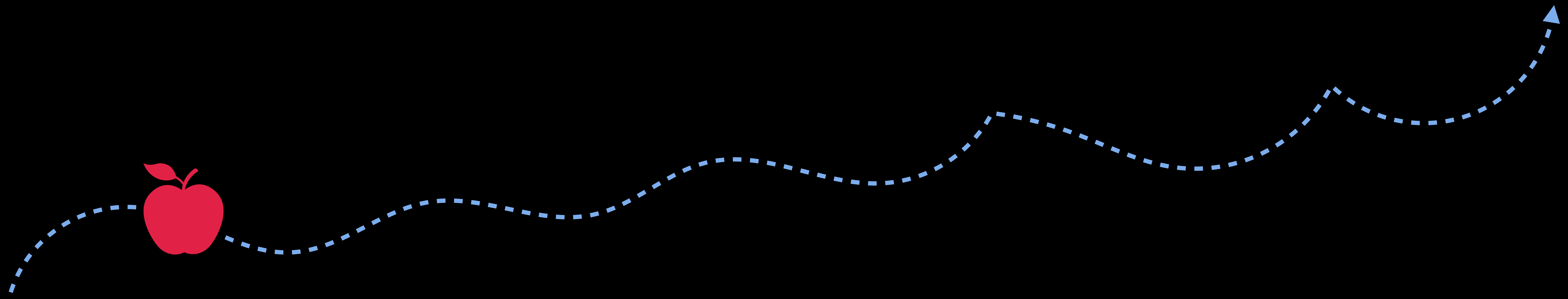
History of ALL Processes Ordered by Timestamp



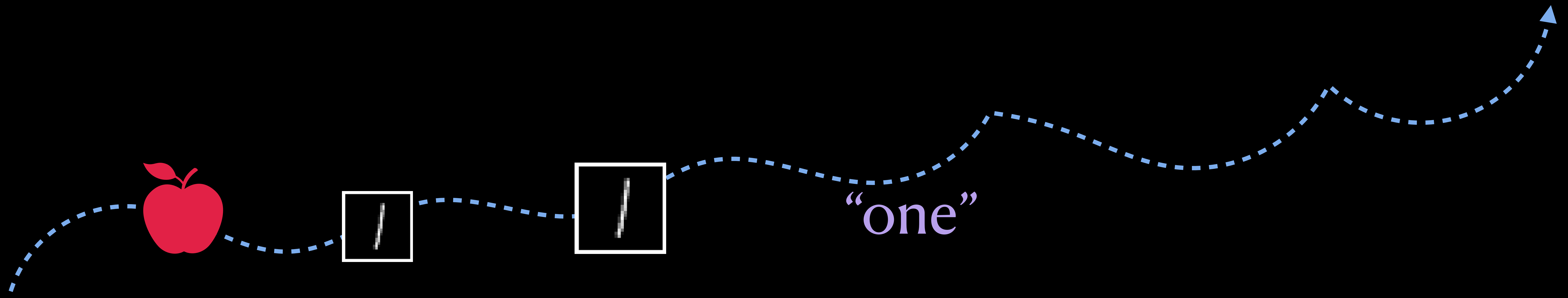
“You cannot step into the same river twice.”

Heraclitus

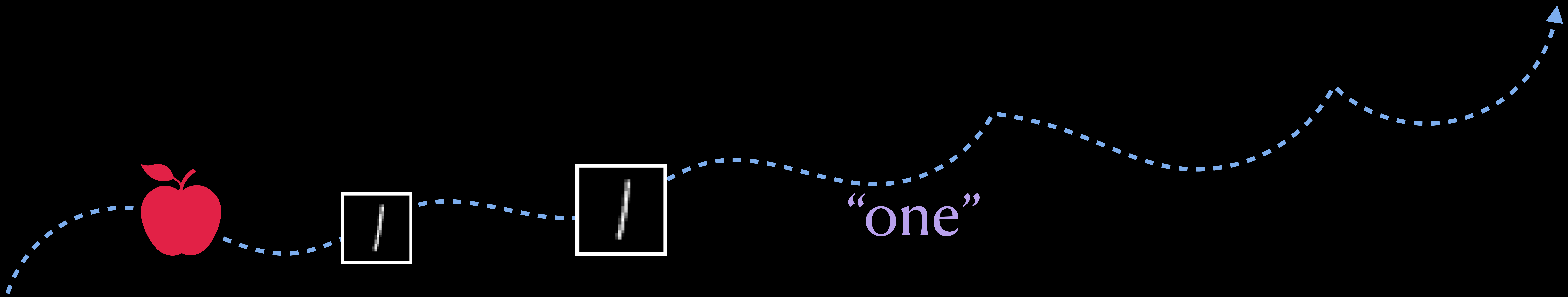
Observation



After some Hidden Processes, More Observation

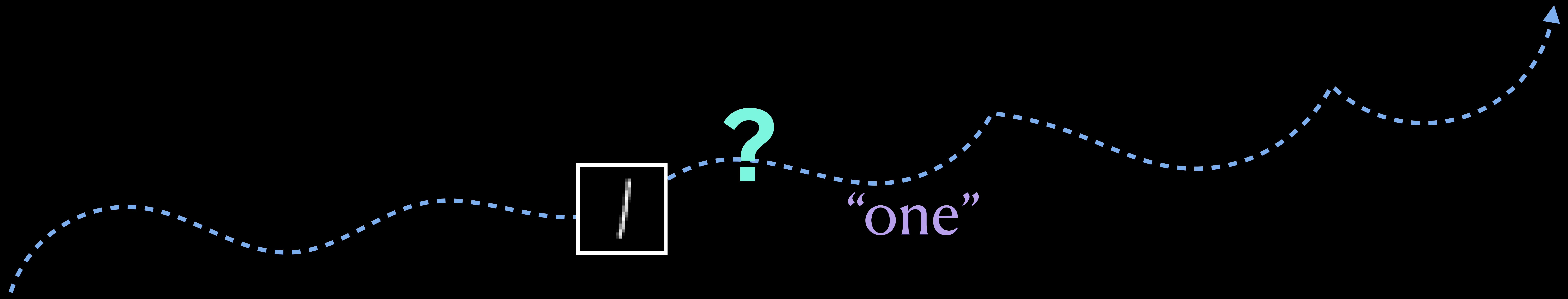


After some Hidden Processes, More Observation

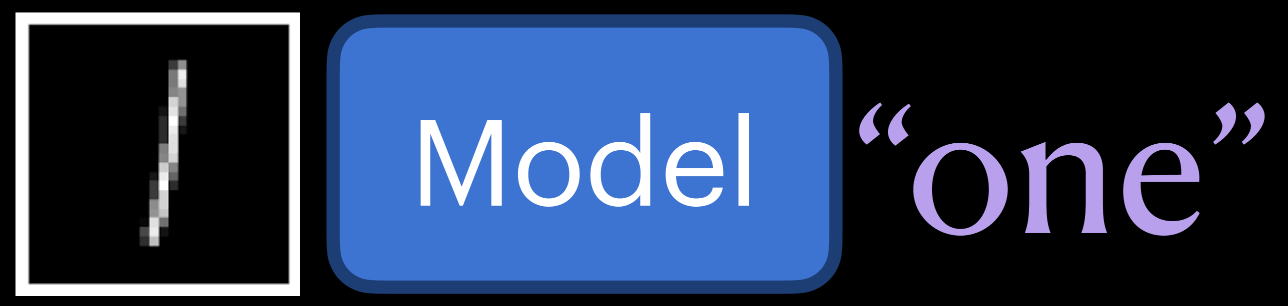


Data are partial observations.

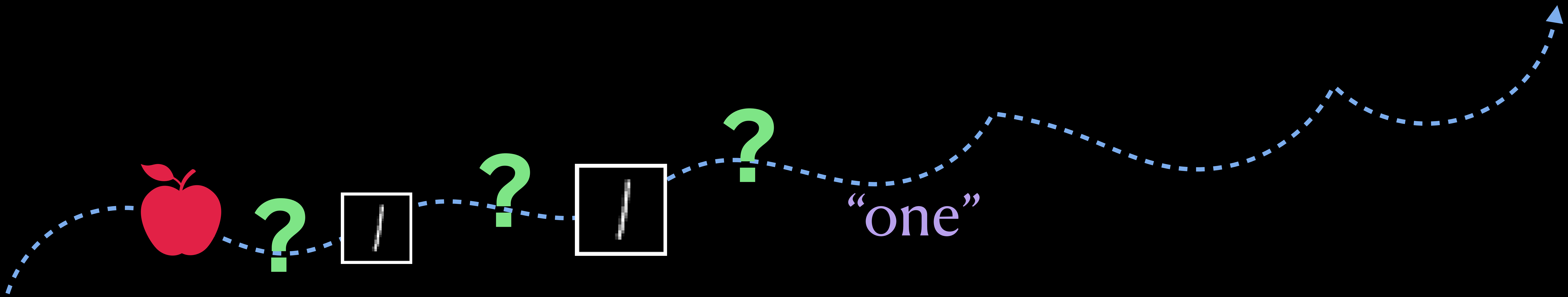
From Processes with Hidden Structures



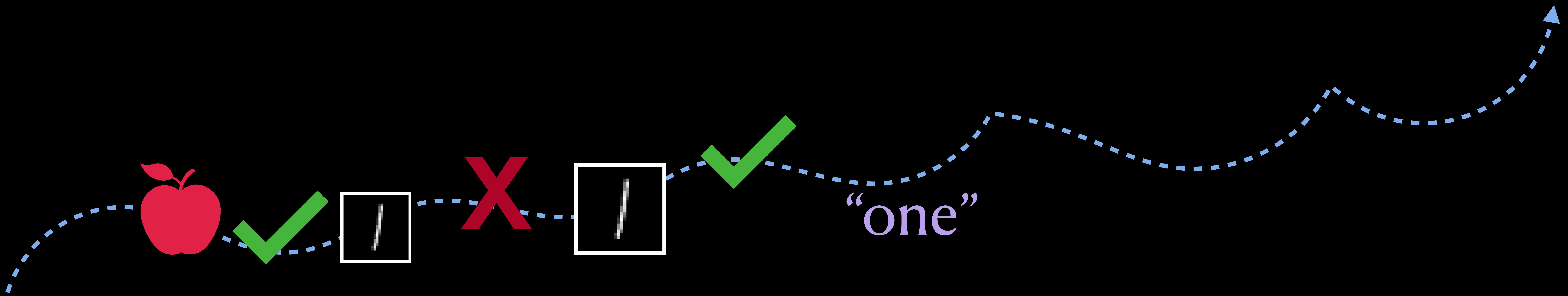
To a Model (that Approximates the Hidden Structure)



Processes are NOT Equally Valuable



Select Processes with Dense Supervision ?



Scaling Processes with Dense Supervision ?

Scaling Processes with Dense Supervision ?

Resources are always limited; cannot scale arbitrarily !!!

What is Bottleneck and Why Finding it Matters ?

Shortest Plank Theory

Shortest Plank Theory

Shortest Plank Theory

Shortest Plank Theory

Bottleneck of AI

- (1990s-late 2000s)
- Big Data
- Small Model
- SVM's fixed non-linear kernel

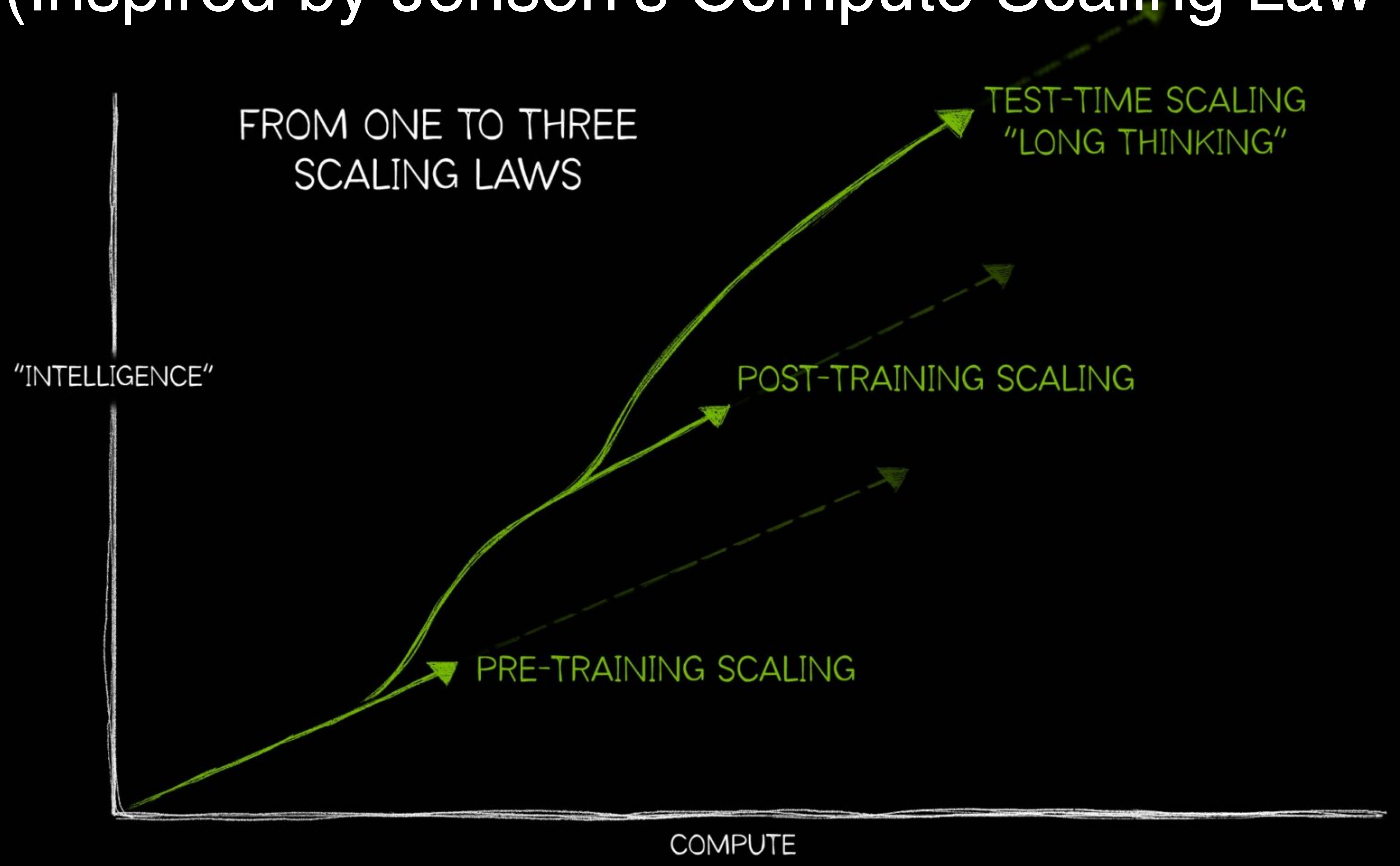
Bottleneck of AI

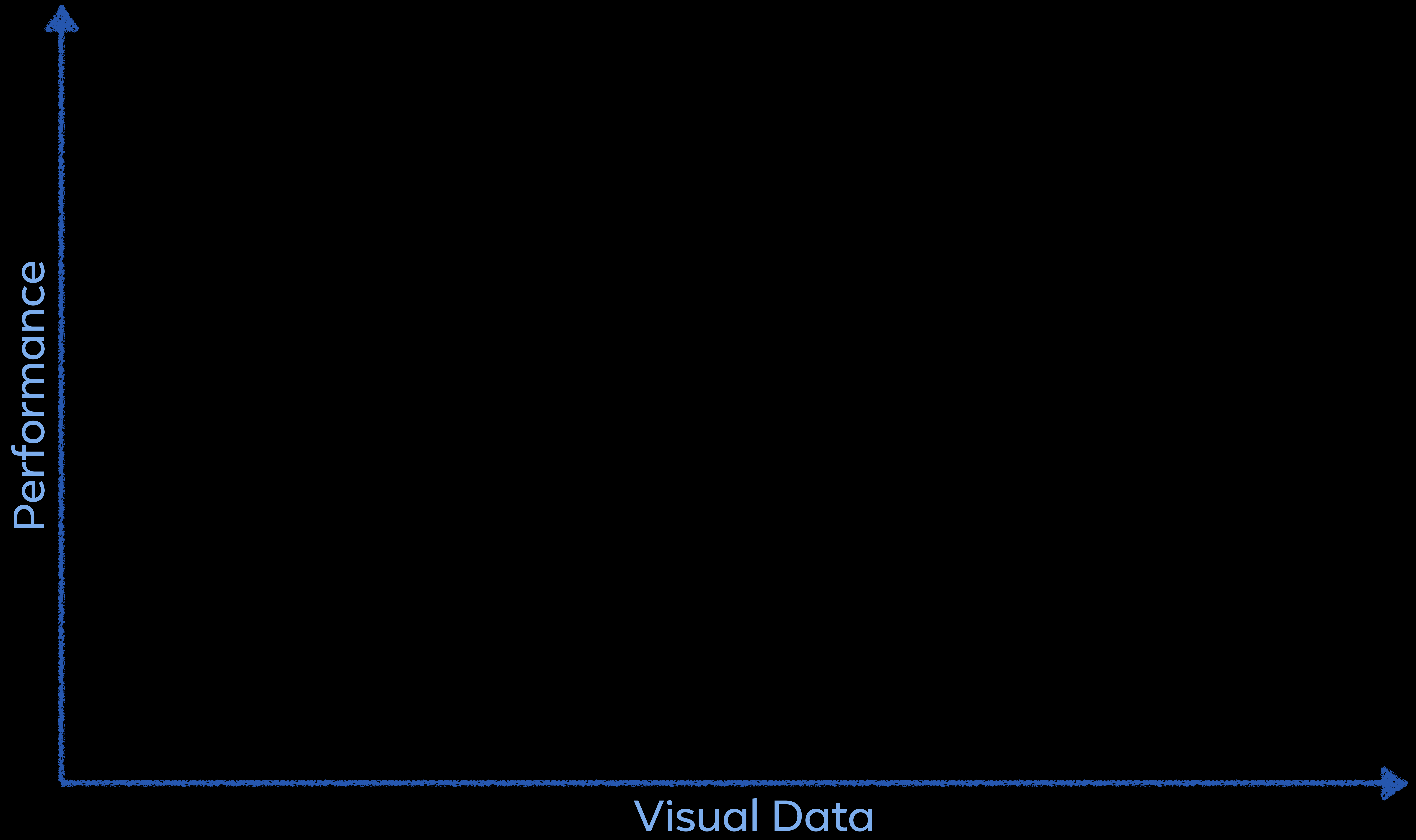
- (2012~2025)
- Big Model (Neural Network)
-   Learnable Non-linear Transformation
- More data ?

Bottleneck of AI

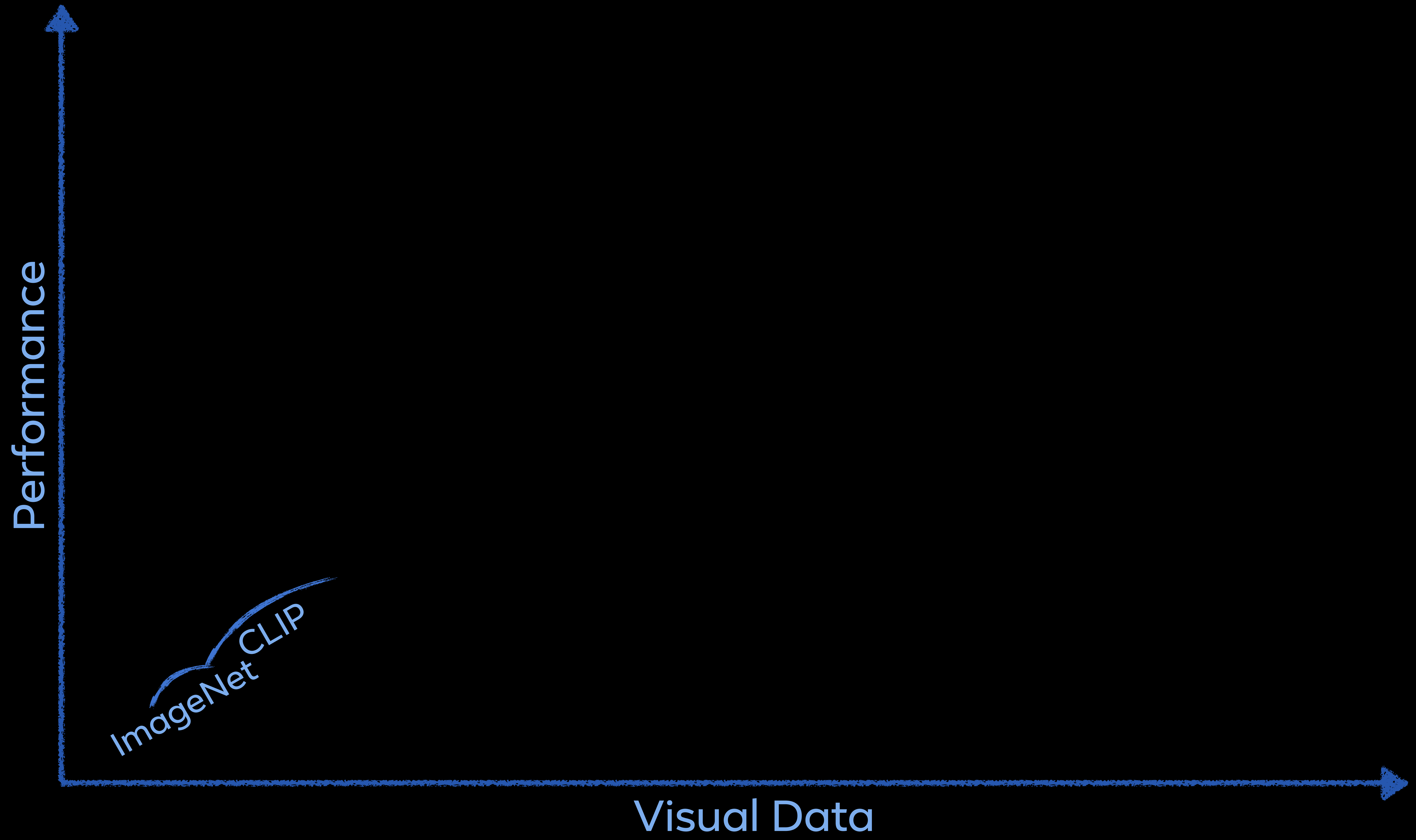
- (2012~2025)
- Big Model (Neural Network)
- Learnable Non-linear Transformation
- Data Filters and Data Walls ?

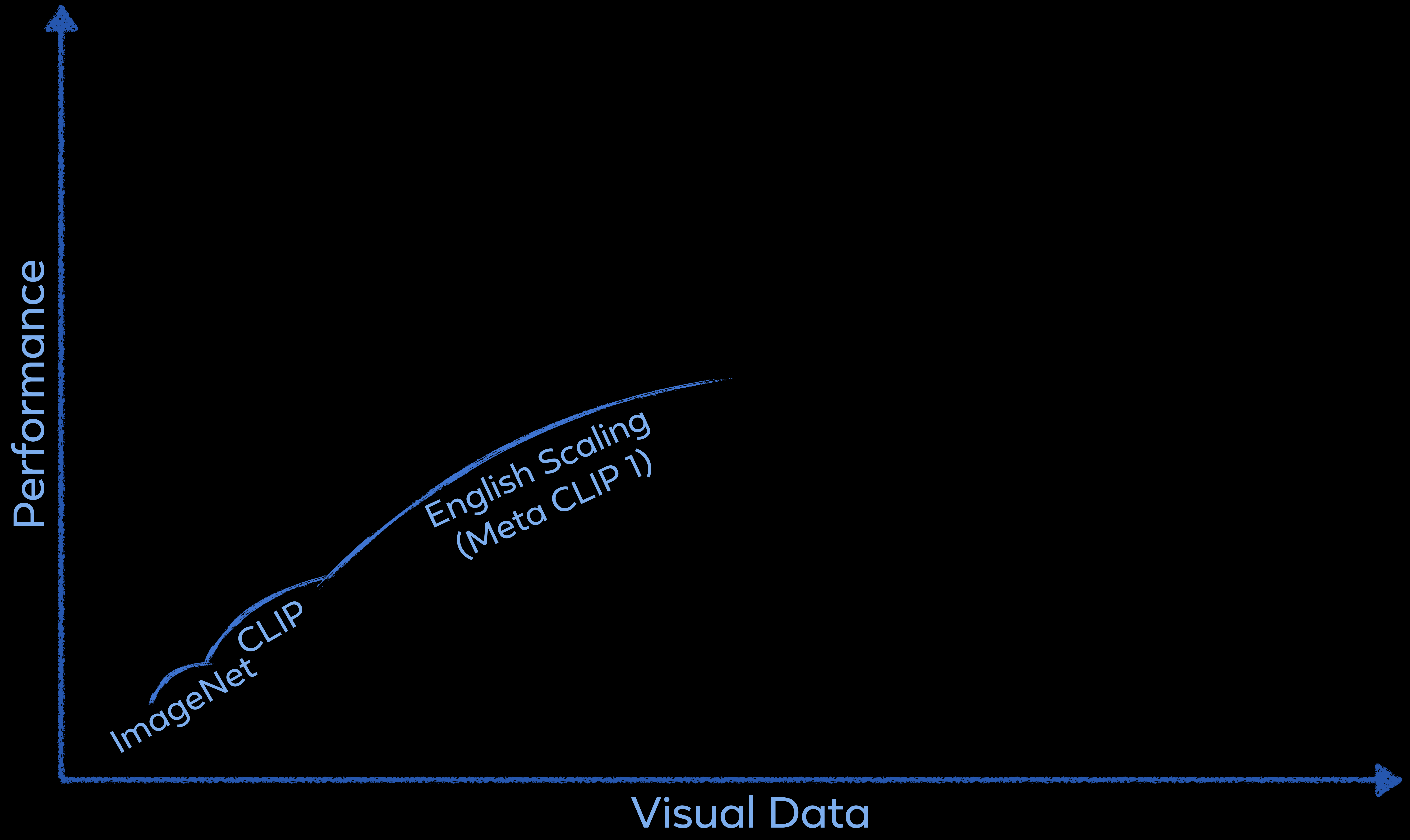
(Inspired by Jensen's Compute Scaling Law …)



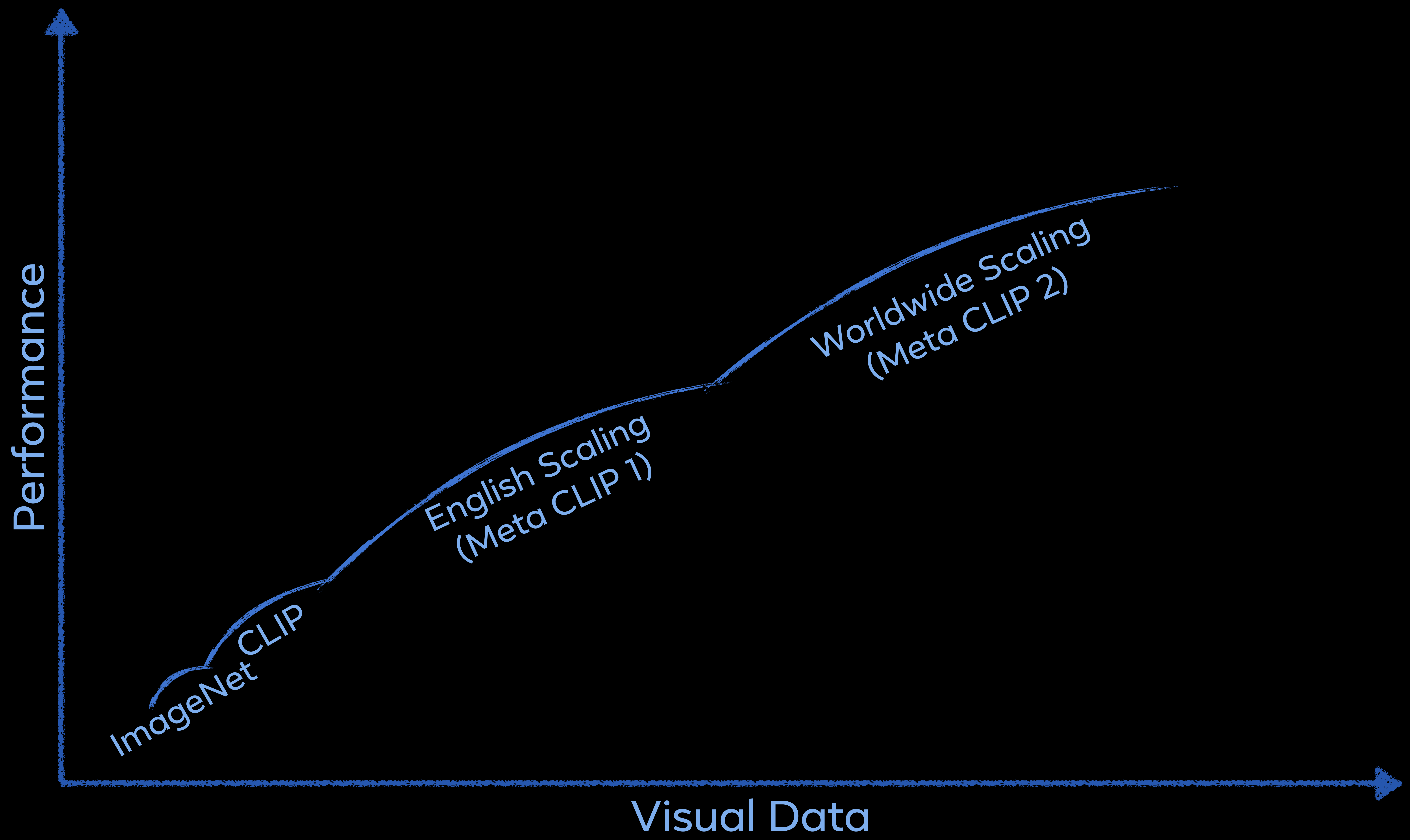

Performance
Visual Data

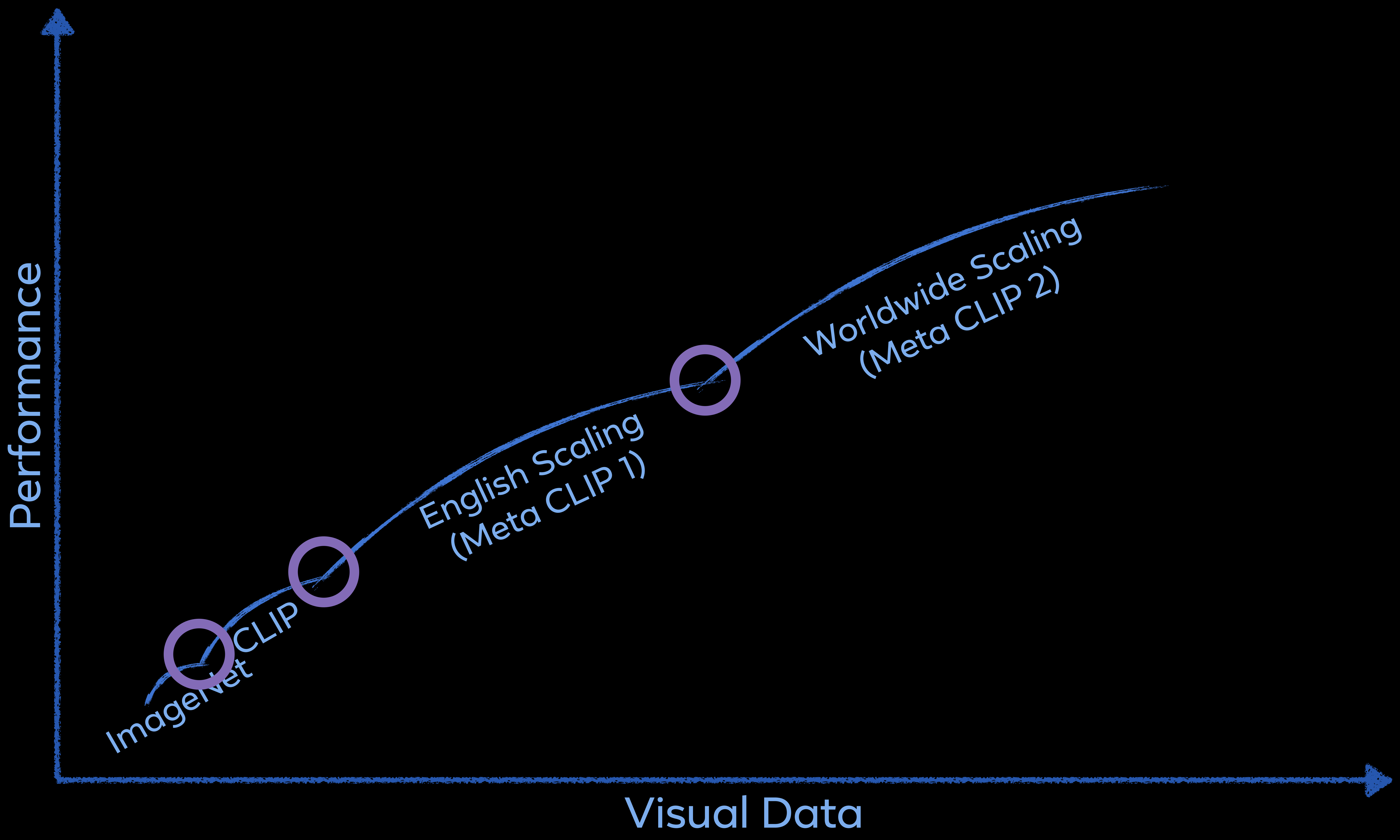

Performance
ImageNet
Visual Data

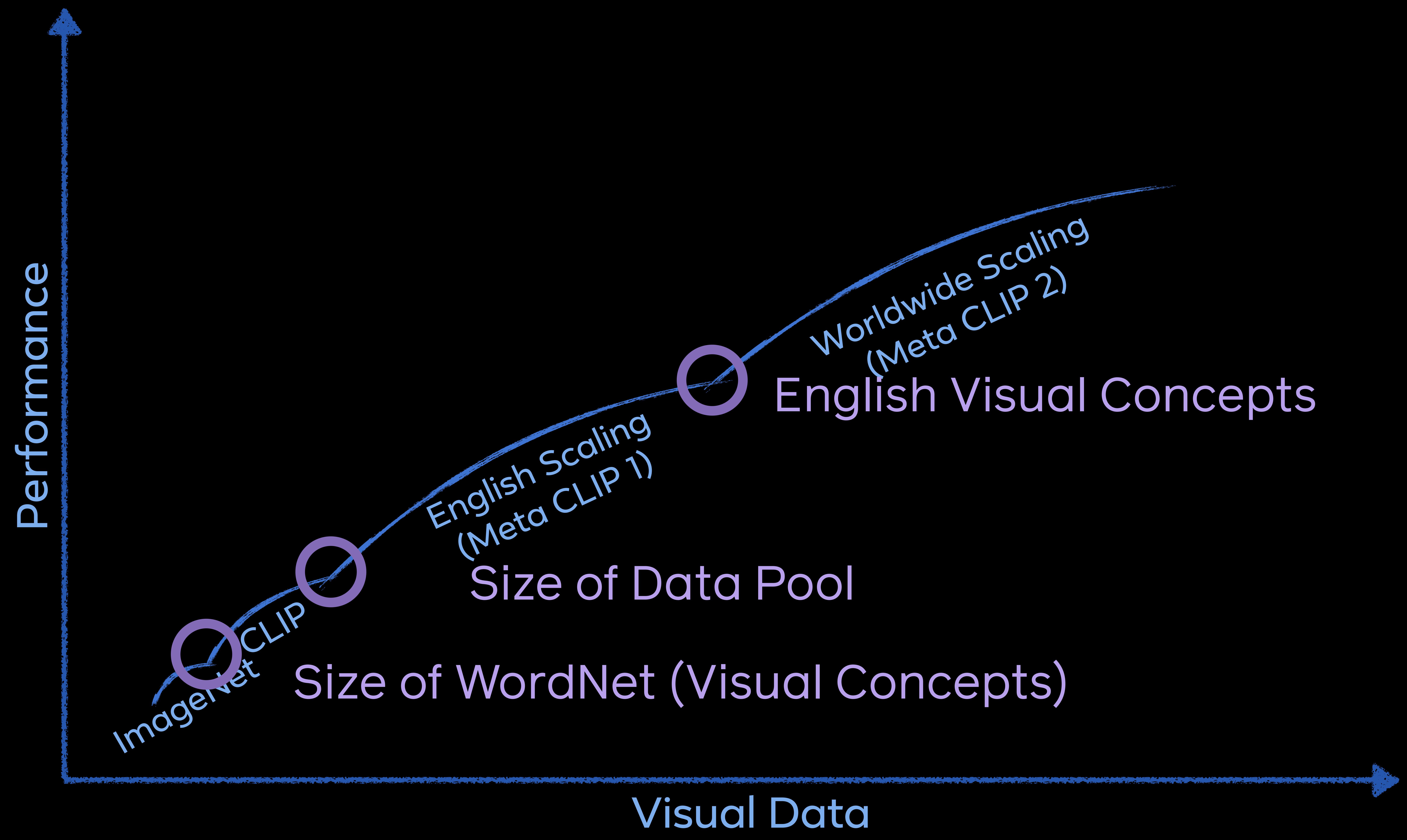

Performance
ImageNet
CLIP
Visual Data

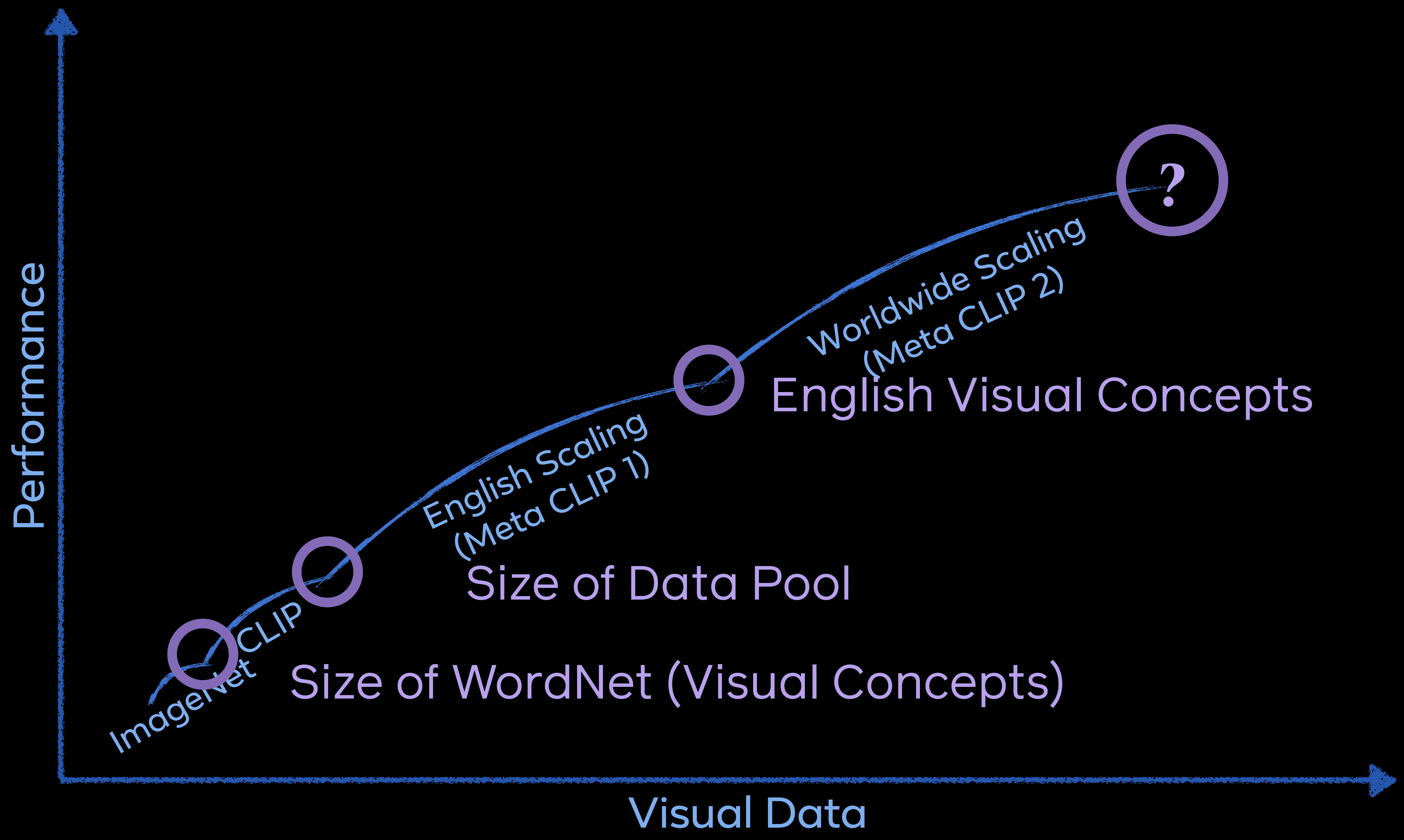


Performance
ImageNet
CLIP
English Scaling
(Meta CLIP 1)
Visual Data


Performance
ImageNet
CLIP
English Scaling
(Meta CLIP 1)
Worldwide Scaling
(Meta CLIP 2)
Visual Data


Performance
ImageNet
CLIP
English Scaling
(Meta CLIP 1)
Worldwide Scaling
(Meta CLIP 2)
Visual Data


Performance
Worldwide Scaling
(Meta CLIP 2)
English Visual Concepts
English Scaling
(Meta CLIP 1)
Size of Data Pool
CLIP
ImageNet
Size of WordNet (Visual Concepts)
Visual Data


Performance
Visual Data
ImageNet
CLIP
English Scaling
(Meta CLIP 1)
Worldwide Scaling
(Meta CLIP 2)
?
Size of WordNet (Visual Concepts)
Size of Data Pool
English Visual Concepts

02 Meta CLIP

Main Contribution

- A formal data algorithm:
- no OpenAI or Google Image Search dependency;

Main Contribution

- A formal data algorithm:
- no OpenAI or Google Image Search dependency;
- Scaling CLIP data to billions from all CommonCrawl image-text pairs;
- Dense Concept Supervision.

Main Contribution

- A formal data algorithm:
- no OpenAI or Google Image Search dependency;
- Scaling CLIP data to billions from all CommonCrawl image-text pairs;
- Dense Concept Supervision, wide adoption by research and production.
- **No Filter Philosophy**:
- CLIP filter / file name filter / date filter etc. are unnecessary or harmful;
- Short-term gains, long-term bottlenecks: **bitter lessons**.

Main Contribution

- A formal data algorithm:
    - no OpenAI or Google Image Search dependency;
- Scaling CLIP data to billions from all CommonCrawl image-text pairs;
    - Dense Concept Supervision, wide adoption by research and production.
- **No Filter Philosophy**:
    - CLIP filter / file name filter / date filter etc. are unnecessary or harmful;
    - Short-term gains, long-term bottlenecks: **bitter lessons**.
- Online Curation: training-on-distribution:
    - NOT a finite dataset.

From a Description in CLIP paper

"To address this, we constructed a new dataset of 400 million (image, text) pairs collected from a variety of publicly available sources on the Internet. To attempt to cover as broad a set of visual concepts as possible, we *search* for (image, text) pairs as part of the construction process whose text includes one of a set of *500,000 queries* We approximately class balance the results by including *up to 20,000 (image, text) pairs per query*."

Radford et al. Learning Transferable Visual Models From Natural Language Supervision

To Data Algorithm

Algorithm 1: Pseudo-code of Curation Algorithm in Python/NumPy style.

```
# D: raw image-text pairs;
# M: metadata;
# t: max matches per entry in metadata;
# D_star: curated image-text pairs;

D_star = []
# Part 1: sub-string matching: store entry indexes in text.matched_entry_ids and
    output counts per entry in entry_count.
entry_count = substr_matching(D, M)
# Part 2: balancing via indepenent sampling
entry_count[entry_count < t] = t
entry_prob = t / entry_count
for image, text in D:
   for entry_id in text.matched_entry_ids:
      if random.random() < entry_prob[entry_id]:
         D_star.append((image, text))
         break
```

To Data Algorithm

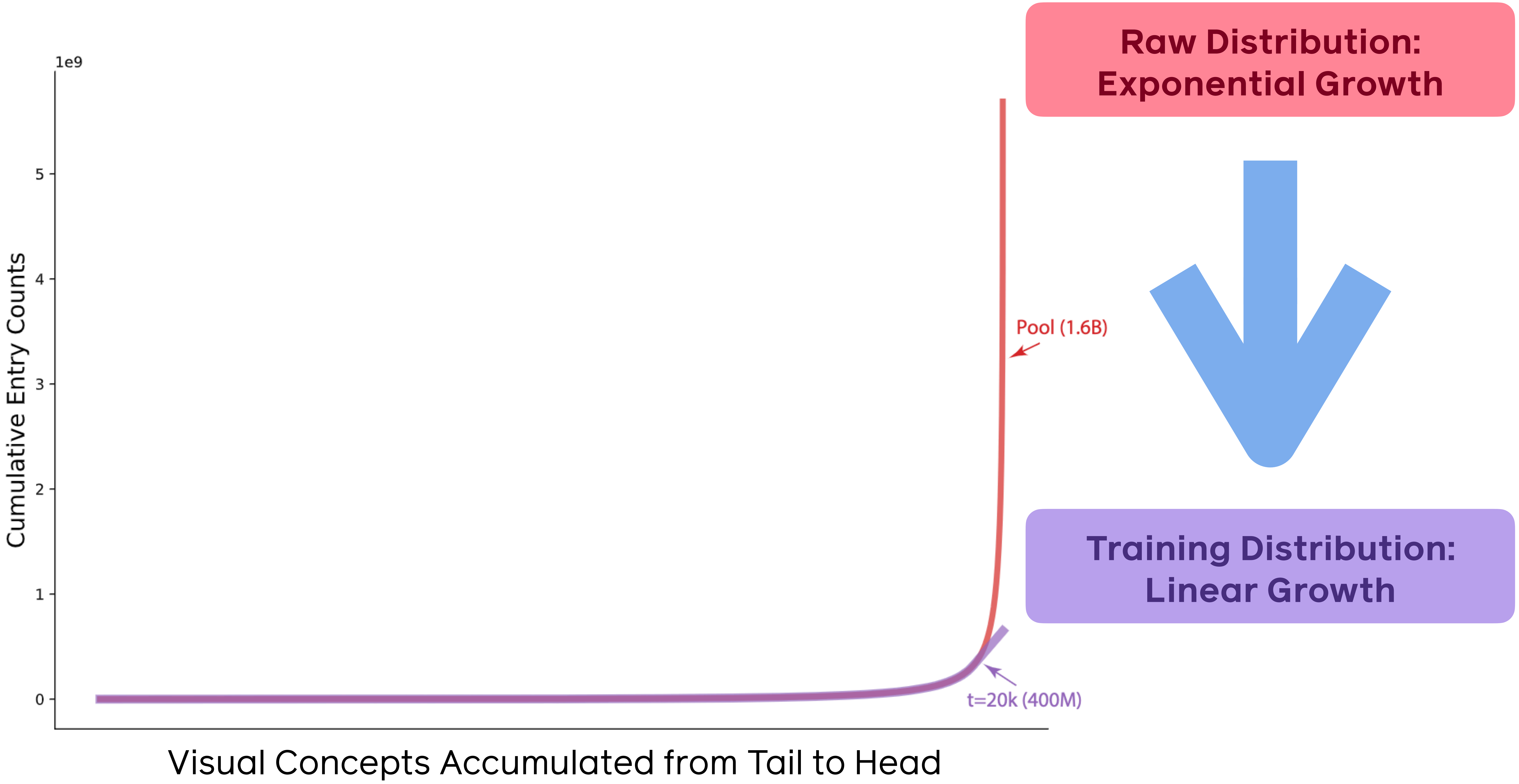
Algorithm 1: Pseudo-code of Curation Algorithm in Python/NumPy style.

Global Operation

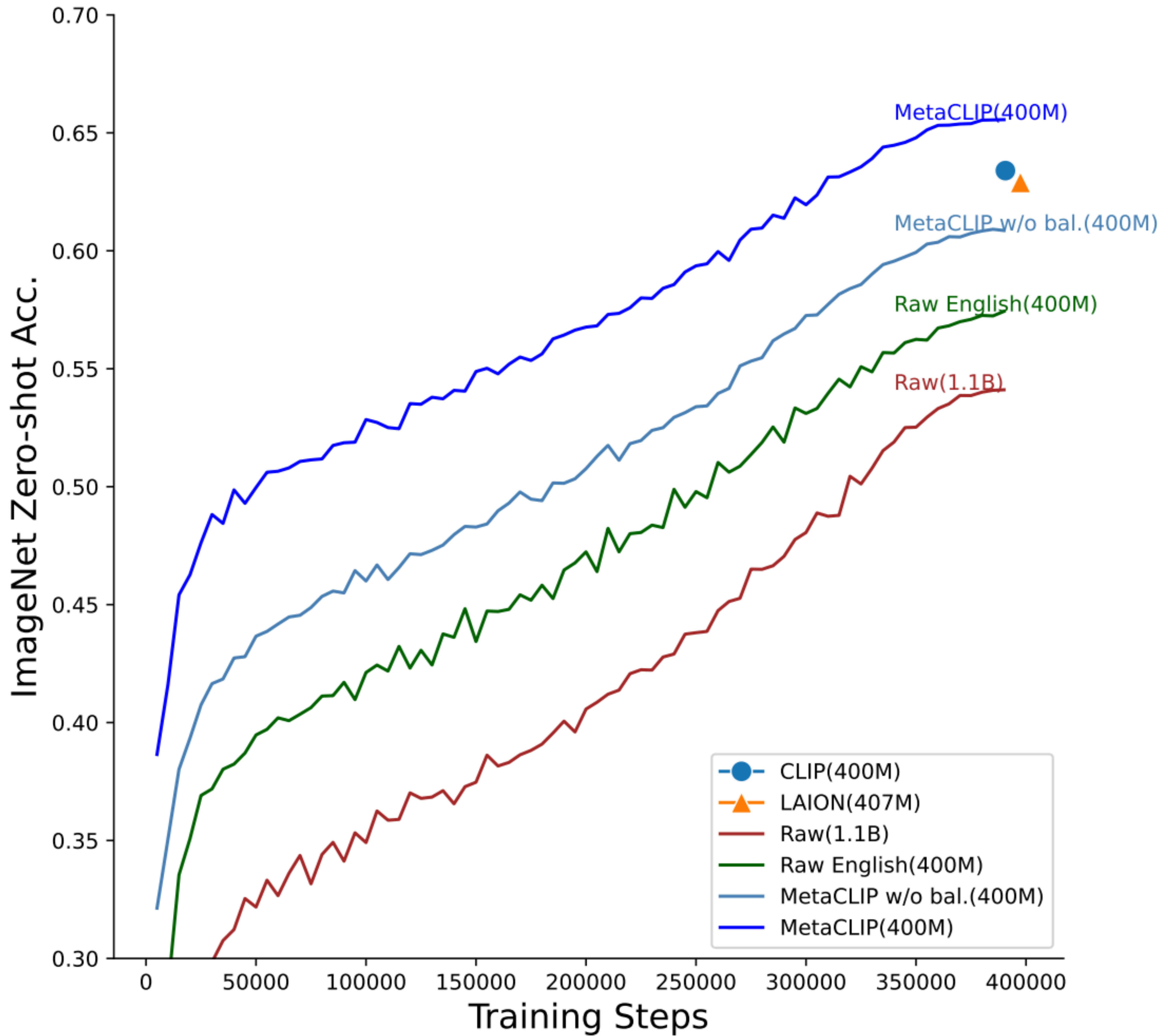
```
# D: raw image-text pairs;
# M: metadata;
# t: max matches per entry in metadata;
# D_star: curated image-text pairs;

D_star = []
# Part 1: sub-string matching: store entry indexes in text.matched_entry_ids and
    output counts per entry in entry_count.
entry_count = substr_matching(D, M)
# Part 2: balancing via indepenent sampling
entry_count[entry_count < t] = t
entry_prob = t / entry_count
for image, text in D:
   for entry_id in text.matched_entry_ids:
      if random.random() < entry_prob[entry_id]:
         D_star.append((image, text))
         break
```

- Minimal global operation, mostly async operations to scale on workers.

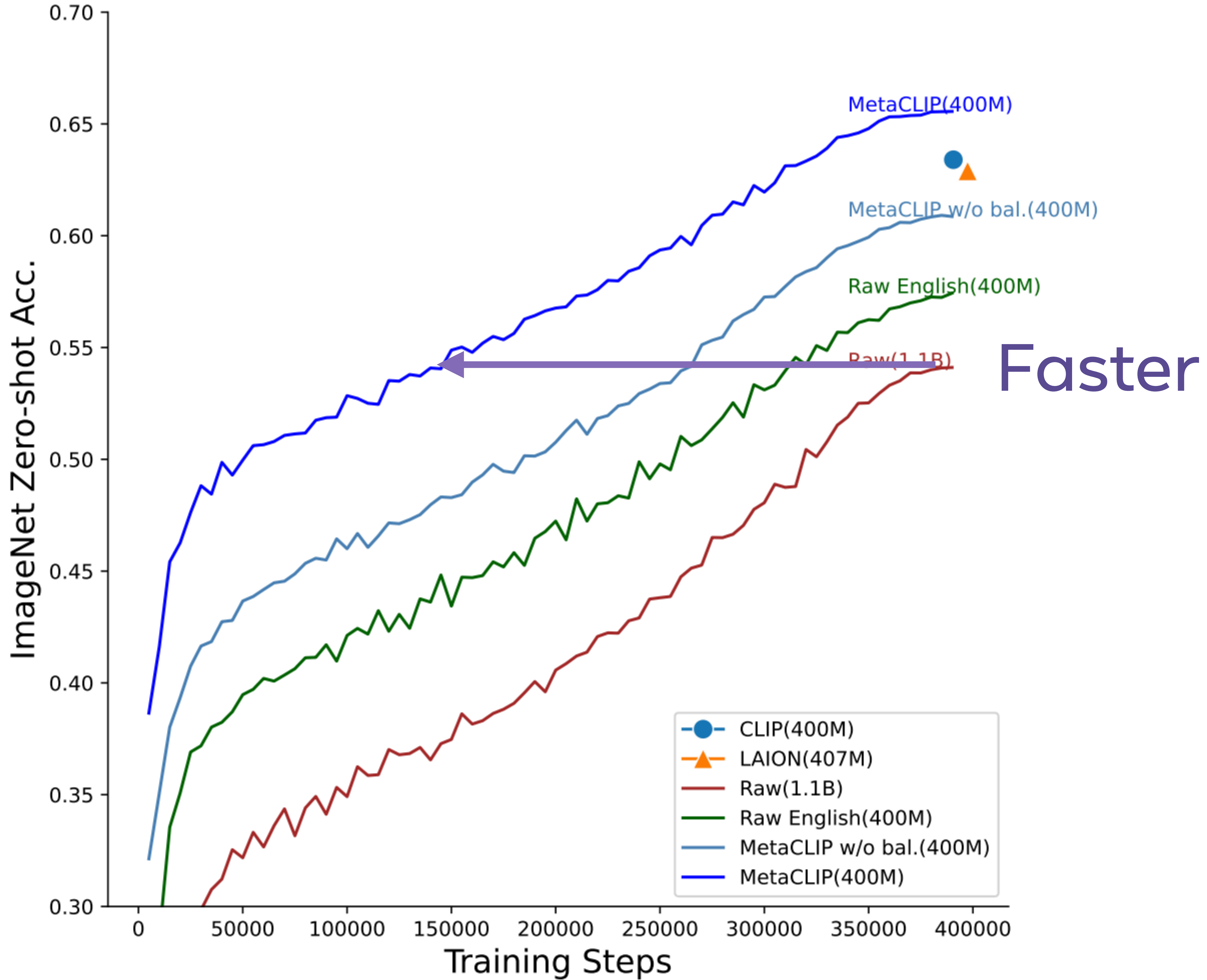
Balancing


1e9
5
4
3
2
1
0
Cumulative Entry Counts
Pool (1.6B)
t=20k (400M)
Visual Concepts Accumulated from Tail to Head
Raw Distribution:
Exponential Growth
Training Distribution:
Linear Growth

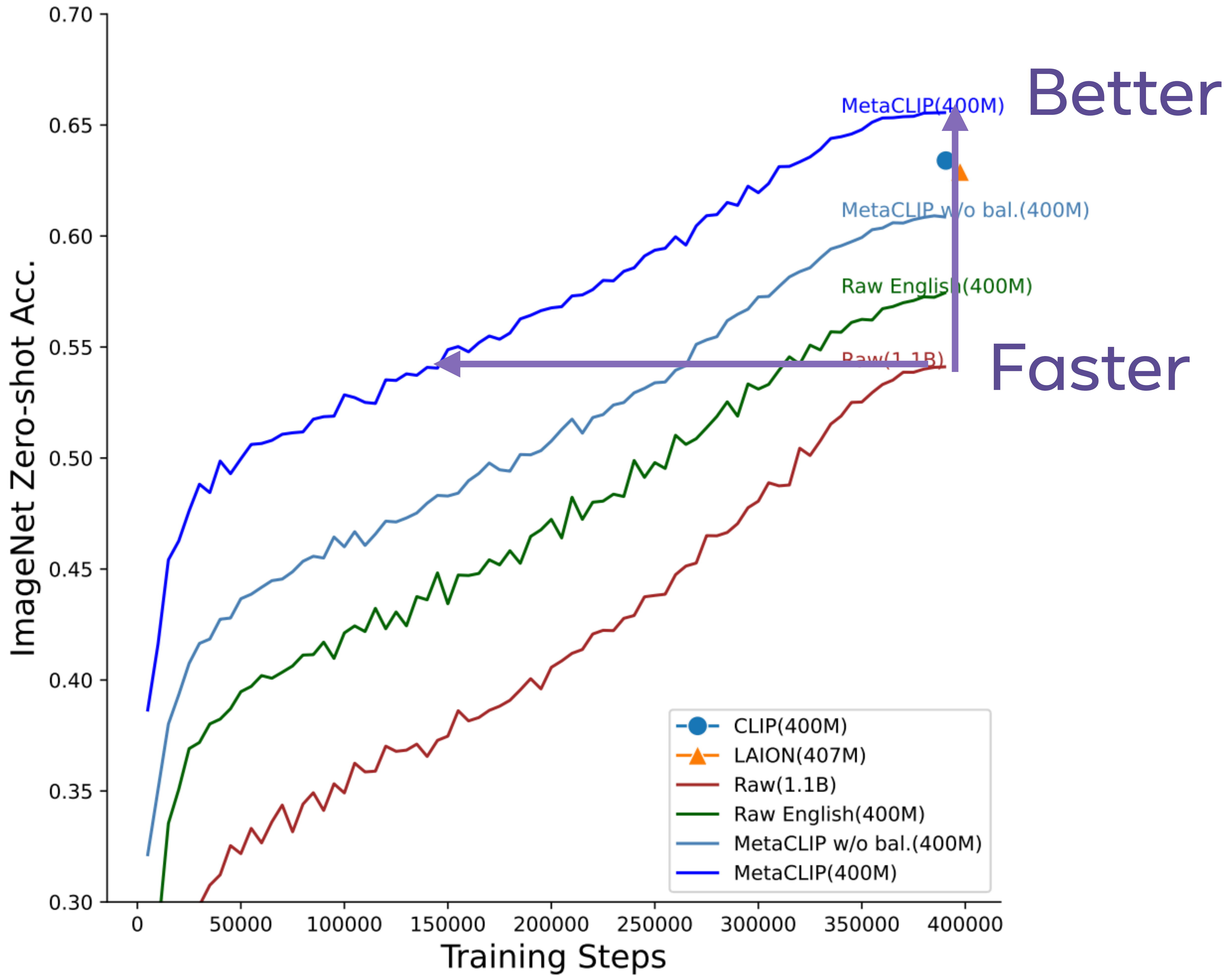
Bending the Curve



Bending the Curve



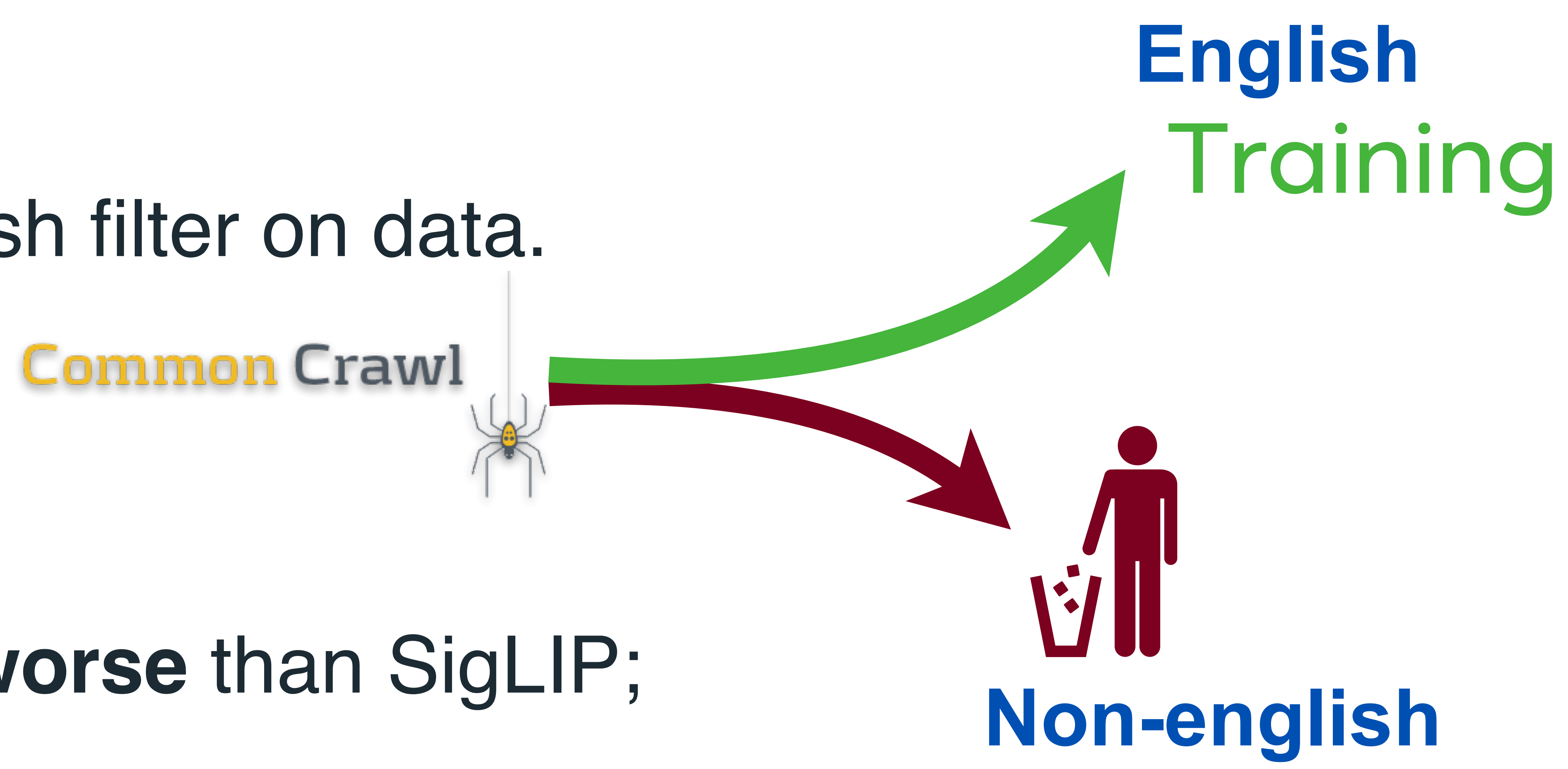
Bending the Curve



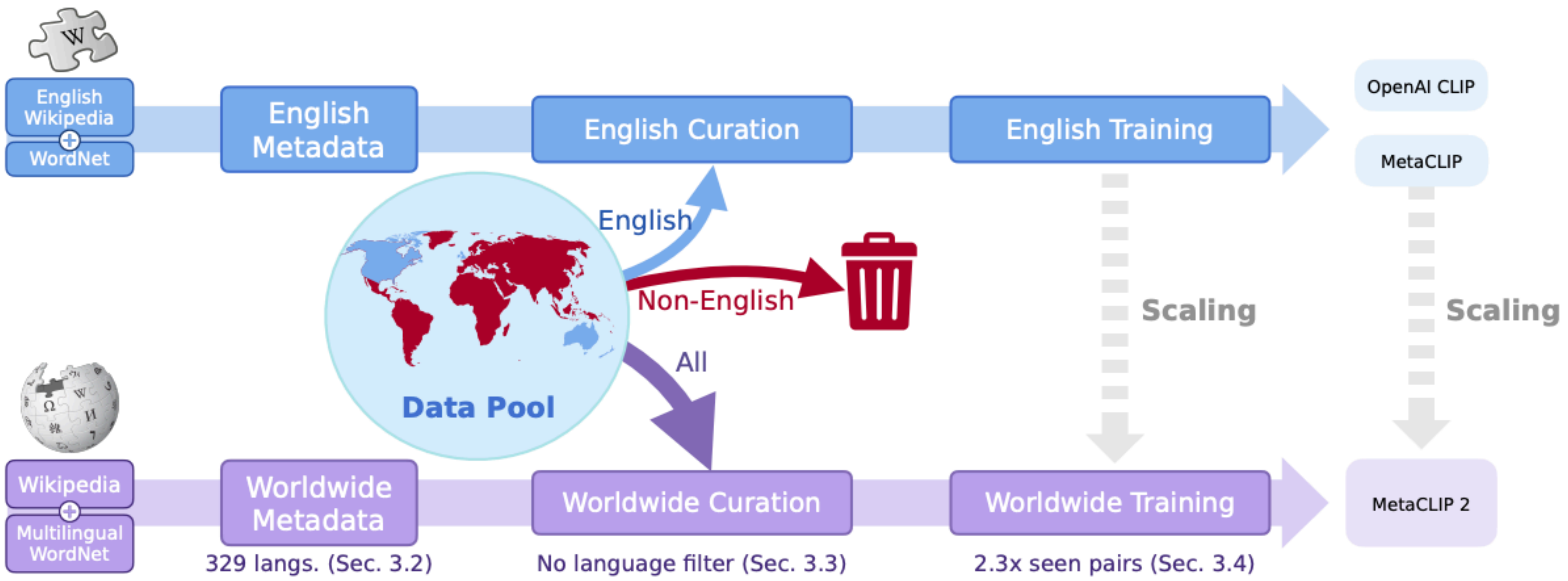
03 Meta CLIP 2

Motivation

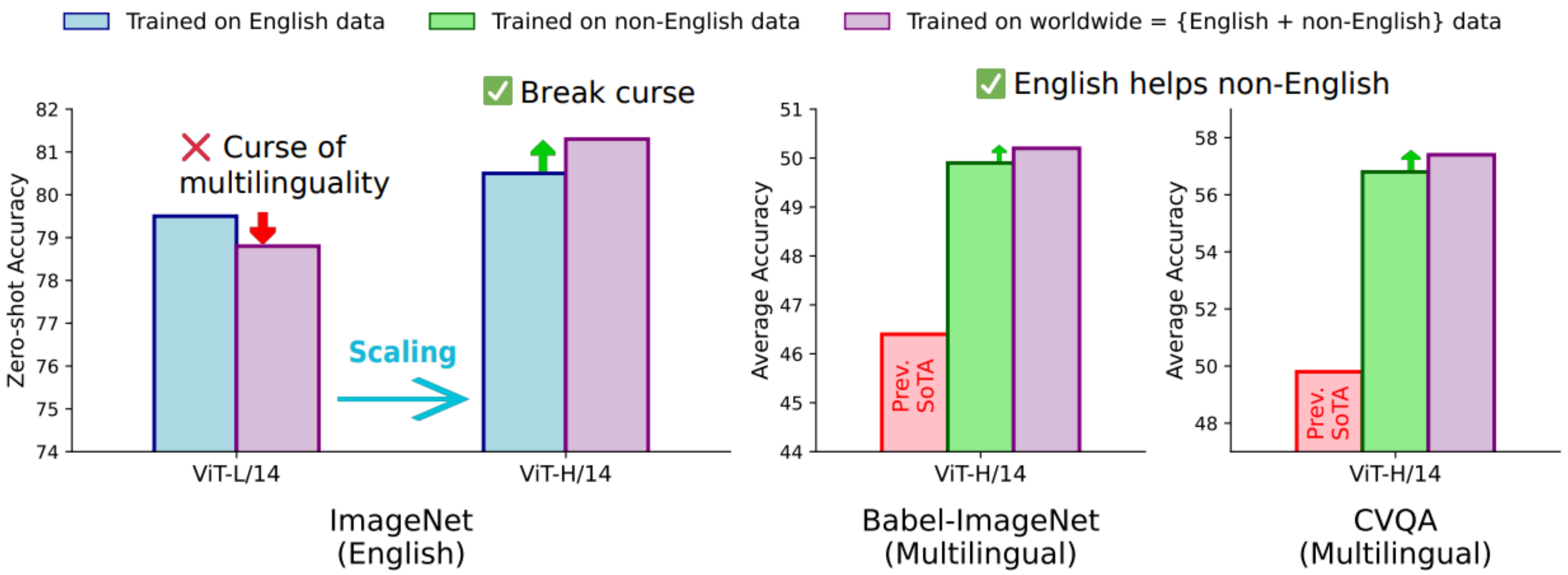
- CLIP is English only, with an implicit English filter on data.
- Dropped 50%+ non-English pairs.
- Curse of Multilinguality:
- eg English performance in mSigLIP is **worse** than SigLIP;
- Hindering wide adoption (English as the major use case).
- Reduce language bias and culture bias.
- If **no filter philosophy for CLIP, so as to languages.**



Meta CLIP 2 Scaling



Break the Curse of Multilinguality



Algorithm 2.0

```
# Stage 1: sub-string matching.
entry_counts = {lang: np.zero(len(M[lang])) for lang in M}
for image, text in D:
   # call substr_match which returns matched entry ids.
   text.matched_entry_ids = substr_match(text, M[text.lang])
   entry_counts[text.lang][text.matched_entry_ids] += 1

# Stage 2: compute t for each langauge.
p = t_to_p(t_en, entry_counts["en"]); t = {}
for lang in entry_counts:
   t[lang] = p_to_t(p, entry_counts[lang])

# Stage 3: balancing via indepenent sampling per language.
entry_probs = {}
for lang in entry_counts:
   entry_counts[lang][entry_counts[lang] < t[lang]] = t[lang]
   entry_probs[lang] = t[lang] / entry_counts[lang]

D_star = []
for image, text in D:
   for entry_id in text.matched_entry_ids:
      if random.random() < entry_probs[text.lang][entry_id]:
         D_star.append((image, text))
         break
```

Scaling Both model (ViT-H) and Seen Pairs (2.3x)

Model	ViT Size (Res.)	Data	Seen Pairs	English Benchmarks			Multilingual Benchmarks					
				IN val	SLIP 26 avg.	DC 37 avg.	Babel -IN	XM3600 T→I I→T	CVQA EN LOC	Flicker30k -200 T→I I→T	XTD-10 T→I I→T	XTD-200 T→I I→T
XLM-CLIP(Ilharco et al., 2021)	H/14(224)	LAION-5B	32B (2.5×)	77.0	69.4	65.5	34.0	50.4 / 60.5	56.1 / 48.2	43.2 / 46.2	87.1 / 88.4	42.5 / 45.2
mSigLIP(Zhai et al., 2023)	B/16(256)	WebLI(12B)	40B (3.0×)	75.1	63.8	60.8	40.2	44.5 / 56.6	51.8 / 45.7	34.0 / 36.0	80.8 / 84.0	37.8 / 40.6
mSigLIP(Zhai et al., 2023)	SO400M(256)	WebLI(12B)	40B (3.0×)	80.6	69.1	65.5	46.4	50.0 / 62.8	56.8 / 49.8	39.9 / 42.0	85.6 / 88.8	42.5 / 45.2
SigLIP 2(Tschannen et al., 2025)	SO400M(256)	WebLI(12B)	40B (3.0×)	83.2	73.7	69.4	40.8	48.2 / 59.7	58.5 / 49.0	36.6 / 40.3	86.1 / 87.6	40.3 / 44.5
Meta CLIP(Xu et al., 2024)	L/14(224)	English(2.5B)	13B (1.0×)	79.2	69.8	65.6	-	- -	- -	- -	- -	- -
	H/14(224)	English(2.5B)	13B (1.0×)	80.5	72.4	66.5	-	- -	- -	- -	- -	- -
Meta CLIP 2	L/14(224)	English	13B (1.0×)	79.5	69.5	66.0	-	- -	- -	- -	- -	- -
		Worldwide	29B (2.3×)	78.8	67.2	63.5	44.2	45.3 / 58.2	59.2 / 55.1	41.9 / 45.8	82.8 / 85.0	41.9 / 44.8
Meta CLIP 2	H/14(224)	English	13B (1.0×)	80.4	72.6	68.7	-	- -	- -	- -	- -	- -
		Non-Eng.	17B (1.3×)	71.4	63.1	61.7	49.9	46.9 / 59.9	59.8 / 56.8	47.5 / 50.5	83.2 / 85.7	46.6 / 49.2
		Worldwide	13B (1.0×)	79.5	71.1	67.2	47.1	49.6 / 62.6	59.9 / 56.0	49.1 / 52.1	85.2 / 87.1	47.0 / 49.7
		Worldwide	29B (2.3×)	81.3	74.5	69.6	50.2	51.5 / 64.3	61.5 / 57.4	50.9 / 53.2	86.1 / 87.5	48.9 / 51.0

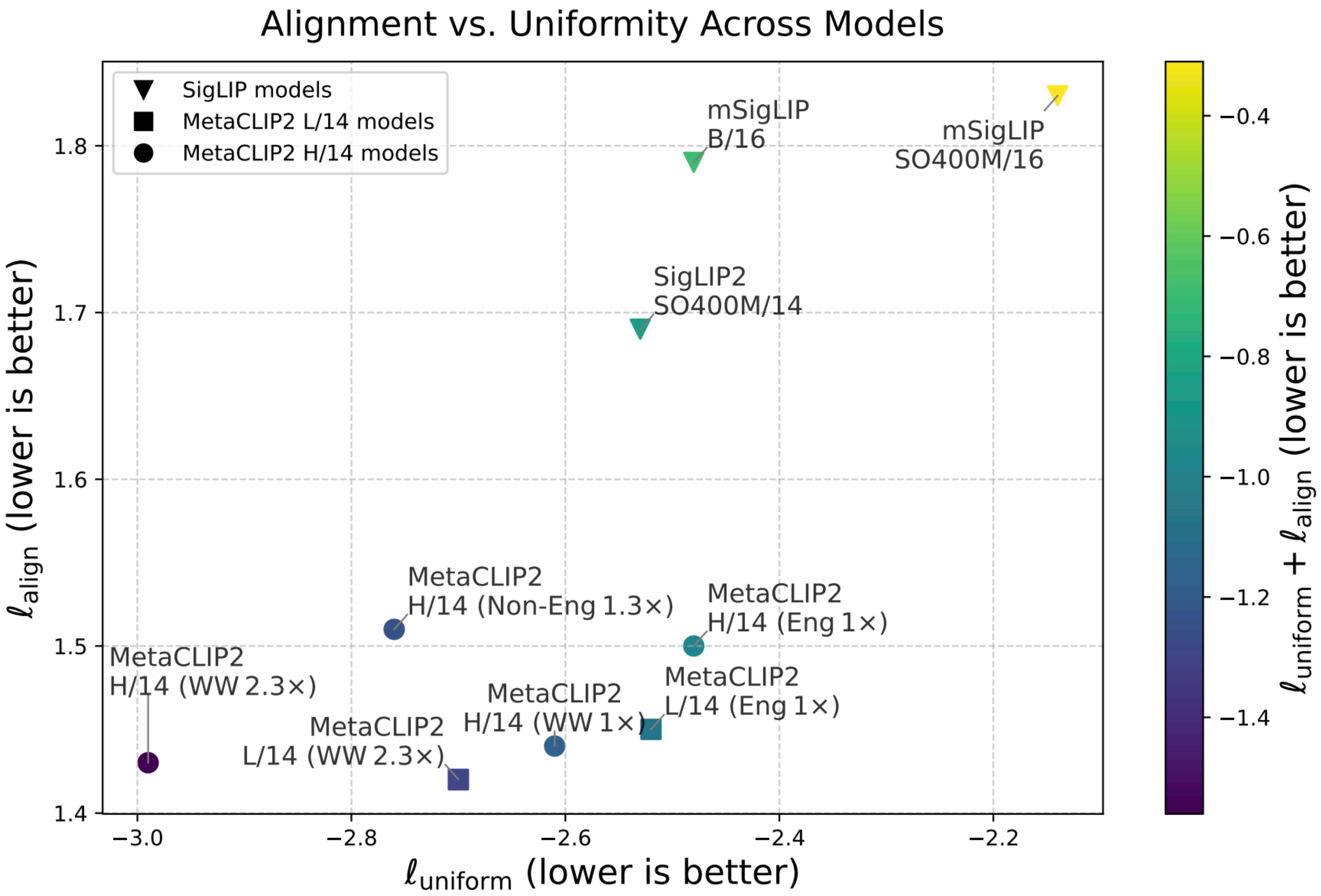
Table 1 Main ablation: Meta CLIP 2 breaks the curse of multilinguality when adopting ViT-H/14, with seen pairs scaled (2.3×) proportional to the added non-English data. Meta CLIP 2 outperforms mSigLIP with fewer seen pairs (72%), lower resolution (224px vs. 256px), and comparable architectures (H/14 vs. SO400M). We grey out baselines those are SoTA-aiming systems with confounding factors. Here, numbers of seen pairs are rounded to the nearest integer (e.g., 12.8B->13B).

To Break the Curse of Multilinguality

Model	ViT Size (Res.)	Data	Seen Pairs	English Benchmarks: IN val	SLIP 26 avg.	DC 37 avg.	Multilingual Benchmarks: Babel -IN	XM3600 T→I I→T	CVQA EN LOC	Flicker30k -200 T→I I→T	XTD-10 T→I I→T	XTD-200 T→I I→T
XLM-CLIP(Ilharco et al., 2021)	H/14(224)	LAION-5B	32B (2.5×)	77.0	69.4	65.5	34.0	50.4 / 60.5	56.1 / 48.2	43.2 / 46.2	87.1 / 88.4	42.5 / 45.2
mSigLIP(Zhai et al., 2023)	B/16(256)	WebLI(12B)	40B (3.0×)	75.1	63.8	60.8	40.2	44.5 / 56.6	51.8 / 45.7	34.0 / 36.0	80.8 / 84.0	37.8 / 40.6
mSigLIP(Zhai et al., 2023)	SO400M(256)	WebLI(12B)	40B (3.0×)	80.6	69.1	65.5	46.4	50.0 / 62.8	56.8 / 49.8	39.9 / 42.0	85.6 / 88.8	42.5 / 45.2
SigLIP 2(Tschannen et al., 2025)	SO400M(256)	WebLI(12B)	40B (3.0×)	83.2	73.7	69.4	40.8	48.2 / 59.7	58.5 / 49.0	36.6 / 40.3	86.1 / 87.6	40.3 / 44.5
Meta CLIP(Xu et al., 2024)	L/14(224)	English(2.5B)	13B (1.0×)	79.2	69.8	65.6	-	- -	- -	- -	- -	- -
	H/14(224)	English(2.5B)	13B (1.0×)	80.5	72.4	66.5	-	- -	- -	- -	- -	- -
Meta CLIP 2	L/14(224)	English	13B (1.0×)	79.5	69.5	66.0	-	- -	- -	- -	- -	- -
		Worldwide	29B (2.3×)	78.8	67.2	63.5	44.2	45.3 / 58.2	59.2 / 55.1	41.9 / 45.8	82.8 / 85.0	41.9 / 44.8
Meta CLIP 2	H/14(224)	English	13B (1.0×)	80.4	72.6	68.7	-	- -	- -	- -	- -	- -
		Non-Eng.	17B (1.3×)	71.4	63.1	61.7	49.9	46.9 / 59.9	59.8 / 56.8	47.5 / 50.5	83.2 / 85.7	46.6 / 49.2
		Worldwide	13B (1.0×)	79.5	71.1	67.2	47.1	49.6 / 62.6	59.9 / 56.0	49.1 / 52.1	85.2 / 87.1	47.0 / 49.7
		Worldwide	29B (2.3×)	81.3	74.5	69.6	50.2	51.5 / 64.3	61.5 / 57.4	50.9 / 53.2	86.1 / 87.5	48.9 / 51.0

Table 1 Main ablation: Meta CLIP 2 breaks the curse of multilinguality when adopting ViT-H/14, with seen pairs scaled (2.3×) proportional to the added non-English data. Meta CLIP 2 outperforms mSigLIP with fewer seen pairs (72%), lower resolution (224px vs. 256px), and comparable architectures (H/14 vs. SO400M). We grey out baselines those are SoTA-aiming systems with confounding factors. Here, numbers of seen pairs are rounded to the nearest integer (e.g., 12.8B->13B).

Alignment and Uniformity



Culture Diversity

Model	Data	Seen Pairs	Dollar Street Top-1	Dollar Street Top-5	GLDv2	GeoDE
mSigLIP (Zhai et al., 2023)	WebLI(12B) (Chen et al., 2023b)	40B (3.0×)	36.0	62.5	45.3	94.5
SigLIP 2 (Tschannen et al., 2025)	WebLI(12B) (Chen et al., 2023b)	40B (3.0×)	36.7	61.9	48.5	95.2
Meta CLIP 2	English	13B (1.0×)	37.2	63.3	52.8	93.4
	Non-English	17B (1.3×)	35.7	61.3	68.6	91.7
	Worldwide	13B (1.0×)	37.2	63.7	65.8	94.3
	Worldwide	29B (2.3×)	37.9	64.0	69.0	93.4

Table 4 Zero-shot classification accuracy on cultural diversity benchmarks. Meta CLIP 2 models are in ViT-H/14 and mSigLIP/SigLIP 2 are in ViT-SO400M. mSigLIP/SigLIP 2 are SoTA-aiming systems with many factors changed and thus greyed out.

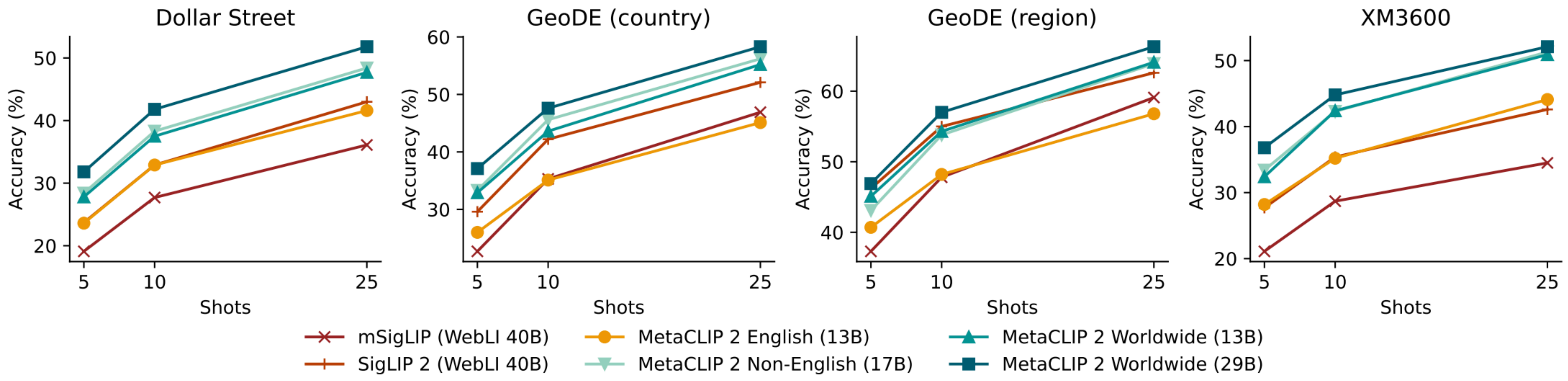
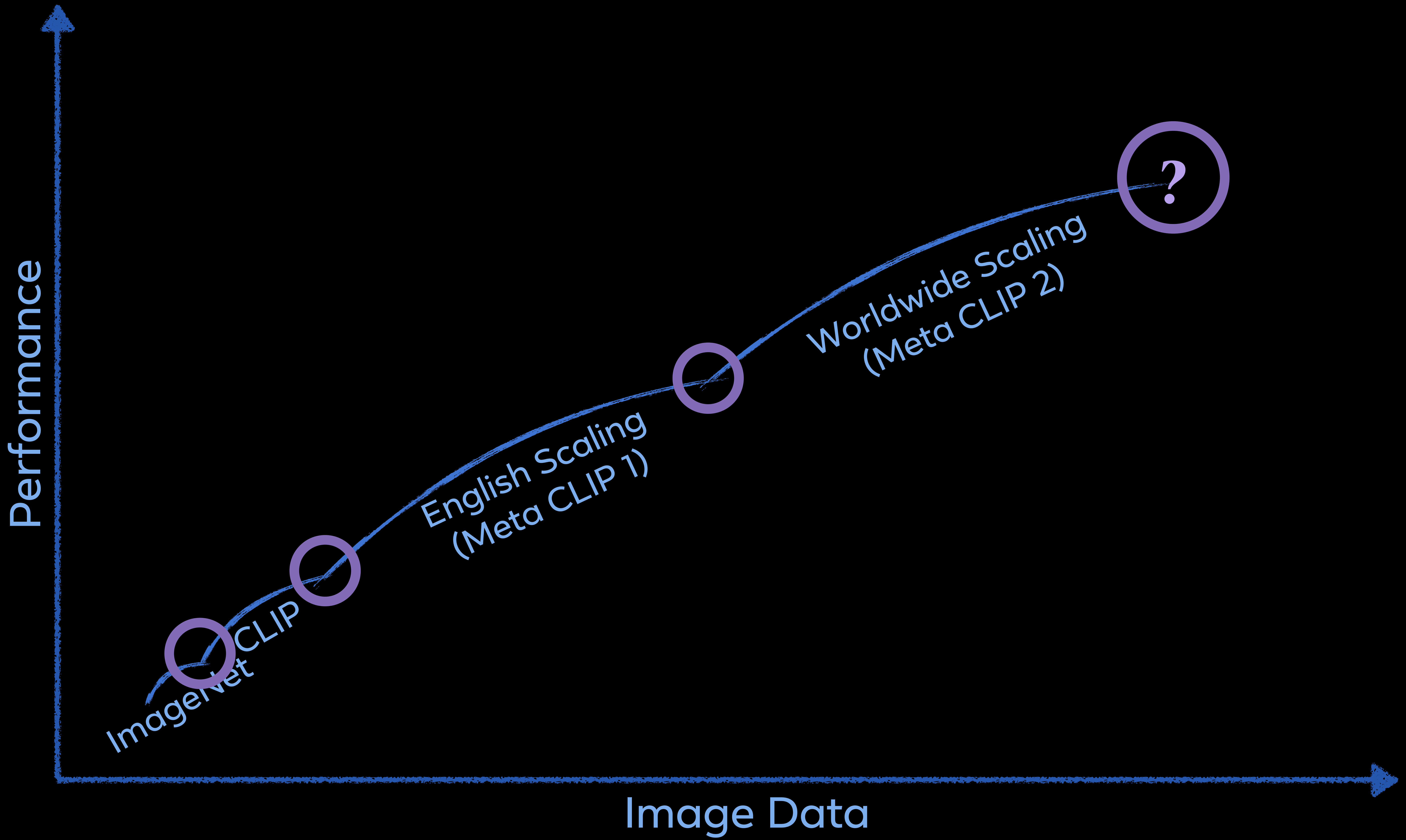
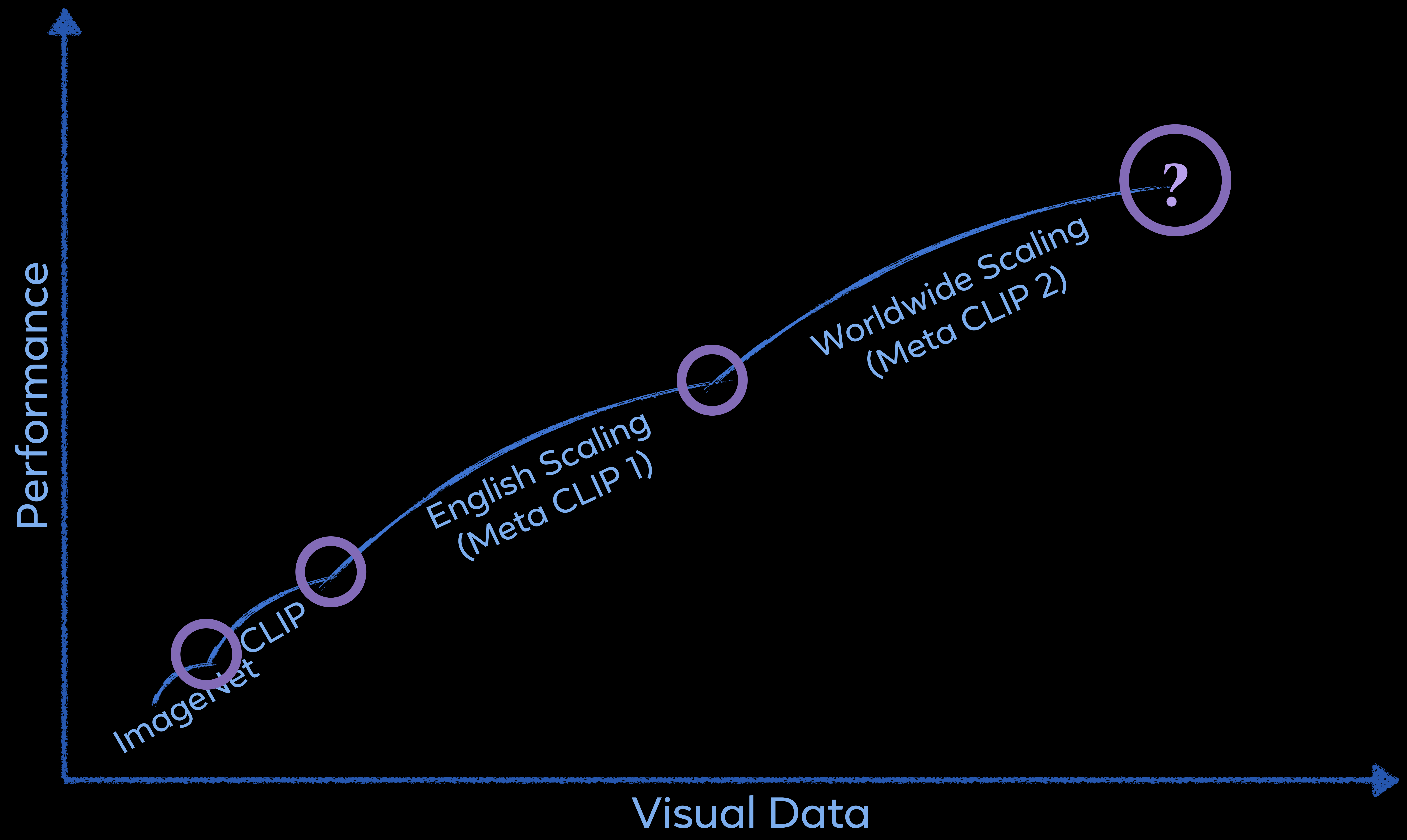
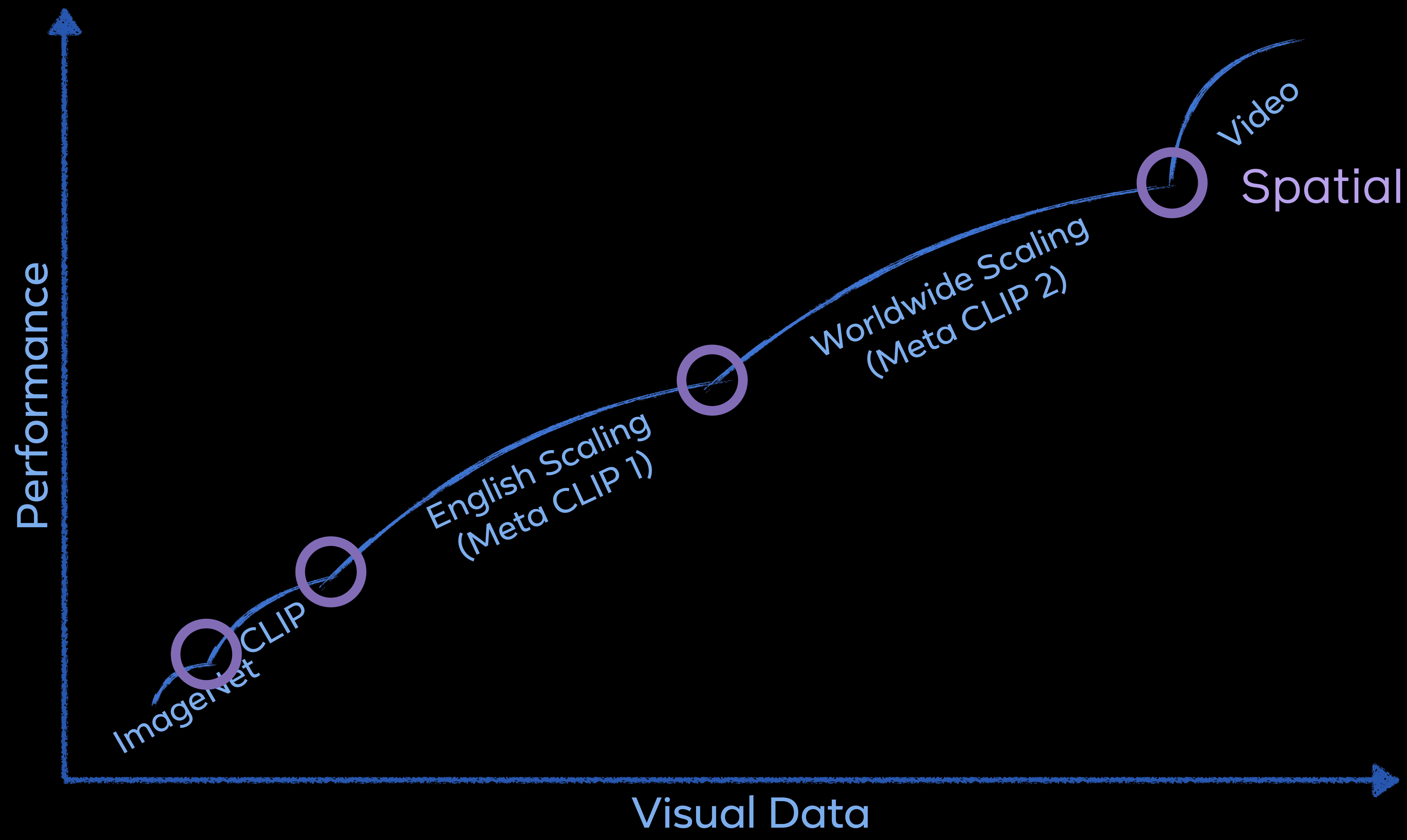


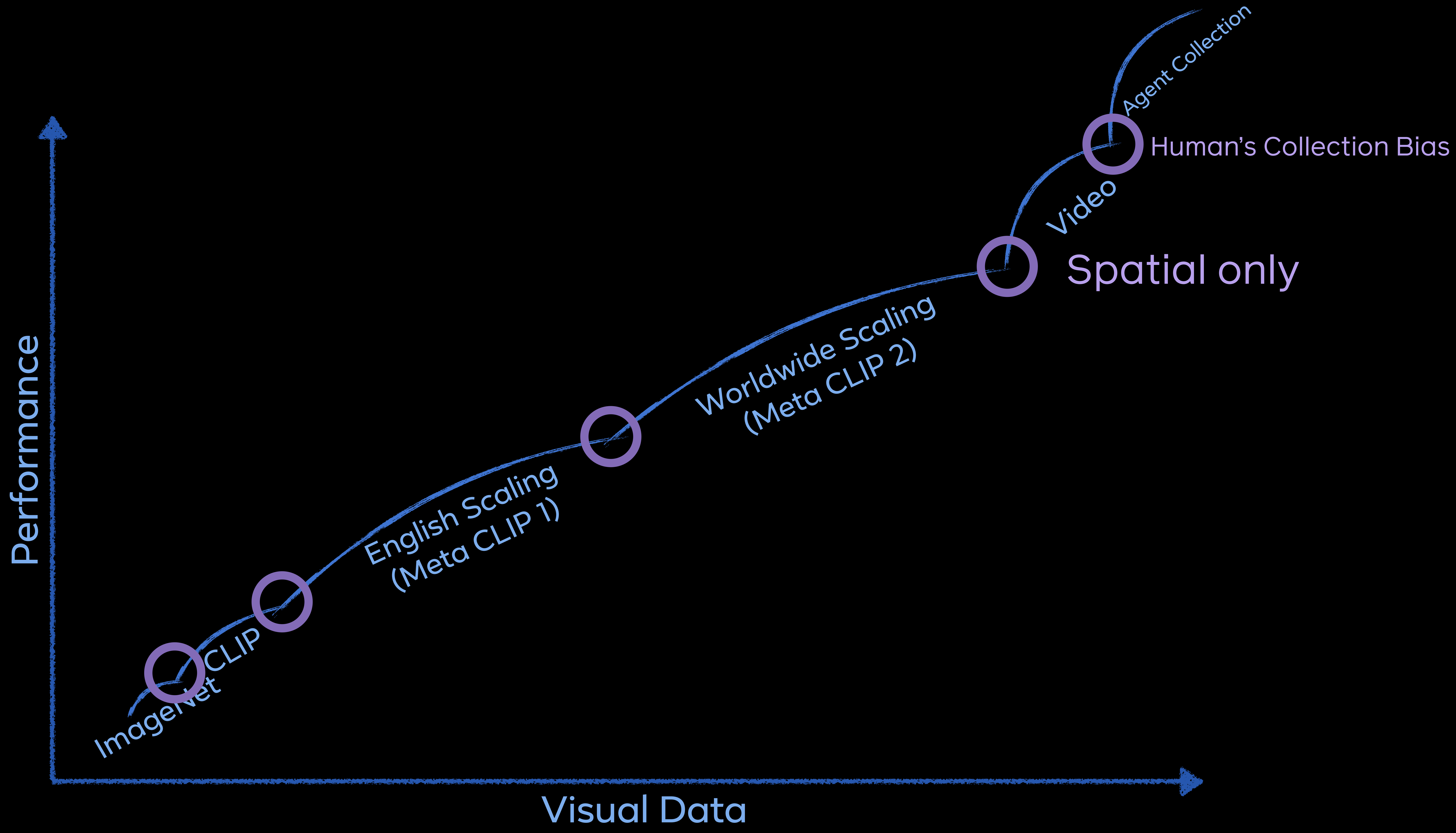
Figure 3 Few-shot geo-localization accuracy on cultural diversity benchmarks.

04 Future Bottlenecks (Estimation)


Performance
ImageNet
CLIP
English Scaling
(Meta CLIP 1)
Worldwide Scaling
(Meta CLIP 2)
?
Image Data


Performance
?
Worldwide Scaling
(Meta CLIP 2)
English Scaling
(Meta CLIP 1)
CLIP
ImageNet
Visual Data


Performance
Video
Spatial only
Worldwide Scaling
(Meta CLIP 2)
English Scaling
(Meta CLIP 1)
CLIP
ImageNet
Visual Data



Performance
Visual Data
ImageNet
CLIP
English Scaling
(Meta CLIP 1)
Worldwide Scaling
(Meta CLIP 2)
Spatial only
Video
Human's Collection Bias
Agent Collection

- Metadata, Code and Model:
- https://github.com/facebookresearch/MetaCLIP
- https://meta-clip.github.io

- For more information, visit Meta Booth, or

- Exhibit Hall C,D,E #4913
- Wed 3 Dec 11 a.m. PST — 2 p.m. PST