Searching for Multimodal Intelligence

Shengbang Tong

New York University

PhD Thesis Defense

We live in a visual world

We see

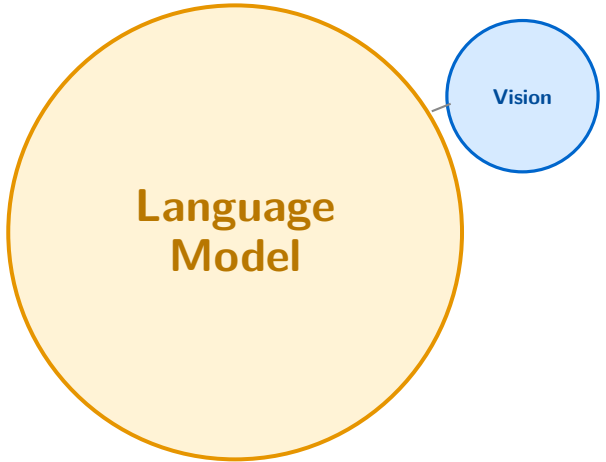
We see **We act**

We see We act **We imagine**

We see We act We imagine **We create**

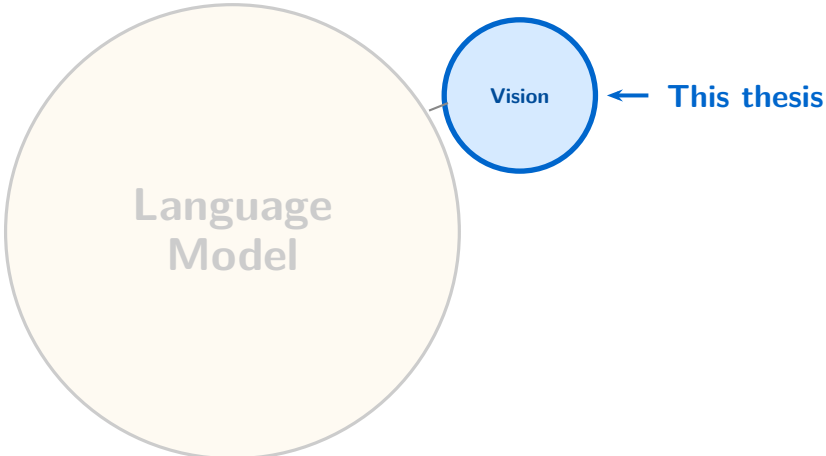
If we turn to machines, we see a different picture

AI has flourished in the text world

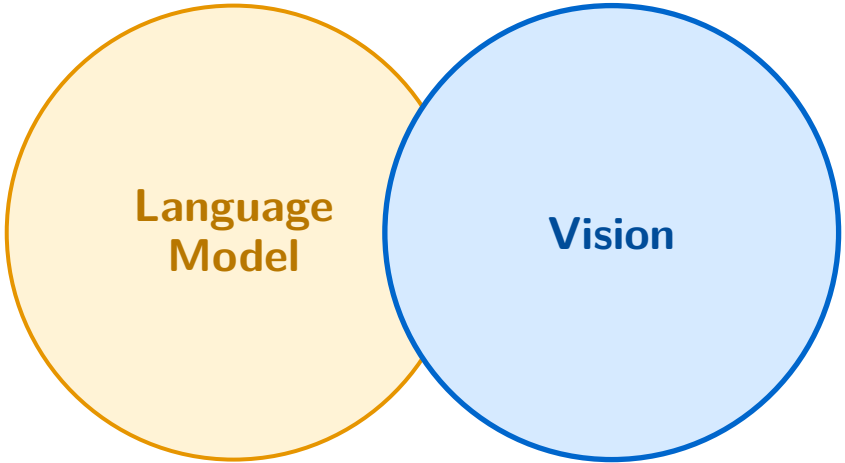


Vision is an additional module built on language

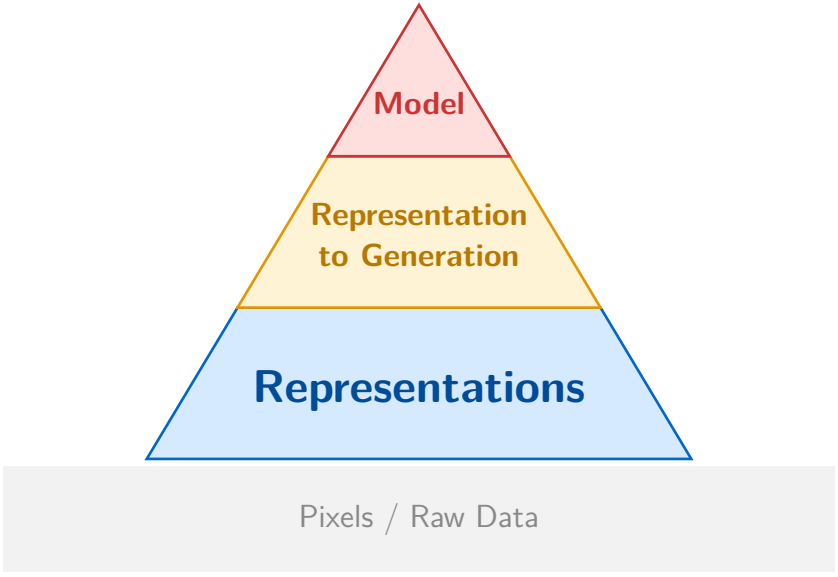
Multimodal models are built on top of pretrained language models

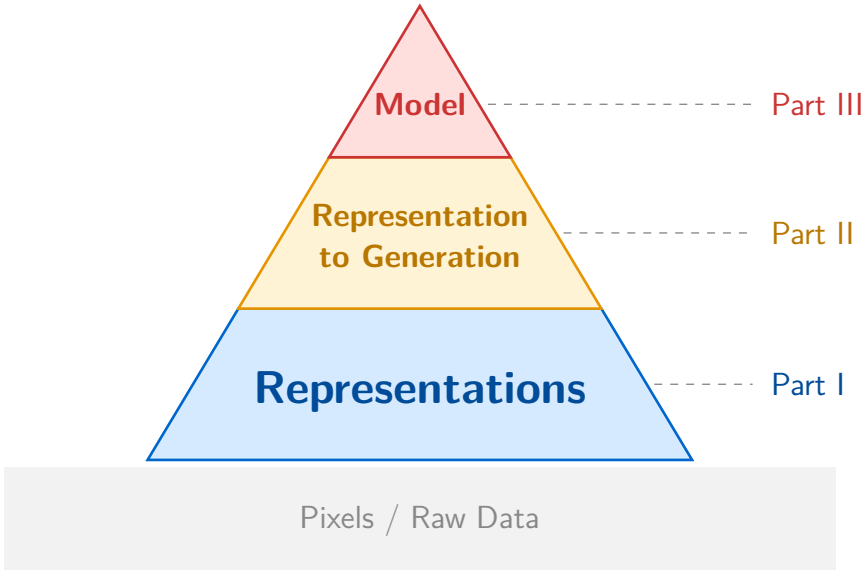


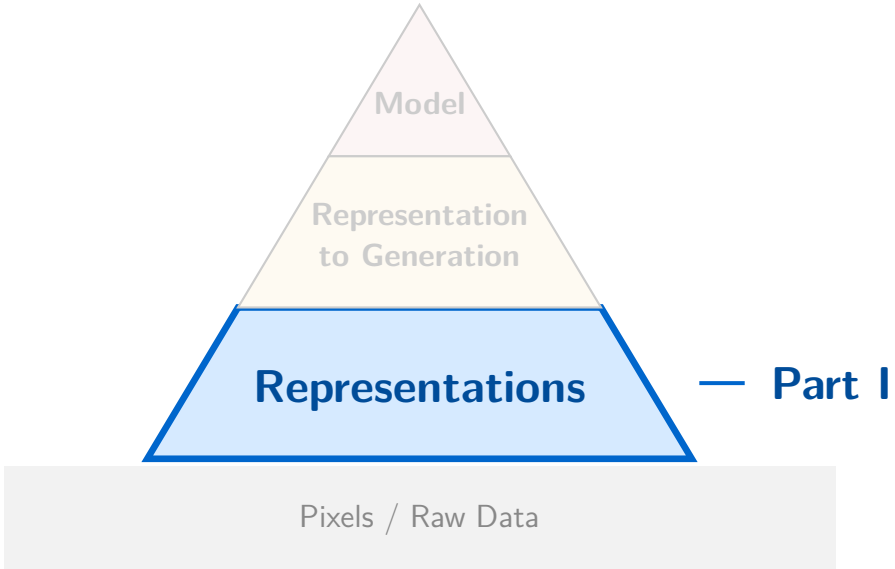
In this thesis, we focus on the vision side



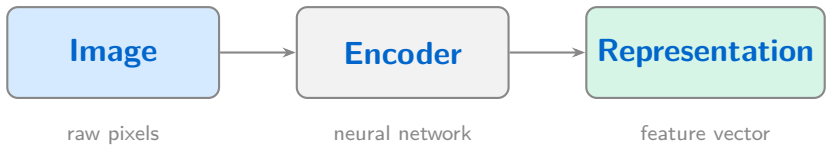
And gradually elevate it


Model
Representation
to Generation
Representations
Pixels / Raw Data


Model
Part III
Representation
to Generation
Part II
Representations
Part I
Pixels / Raw Data


Model
Representation
to Generation
Representations
— Part I
Pixels / Raw Data

Visual representations



An encoder maps an image into a **representation**

Two ways to learn visual representations

Self-Supervised

Learn from **images alone**

DINO, DINOv2, MAE, I-JEPA

No text, no labels

Language-Supervised

Learn from **image-text pairs**

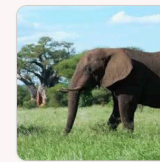
CLIP, SigLIP, OpenCLIP

Web-scale captions

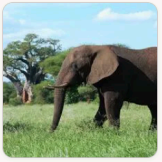
Back in 2023, **CLIP** is the dominant vision encoder in multimodal LLMs

LLaVA, InstructBLIP, MiniGPT-4, Qwen-VL, ...

All built on CLIP.

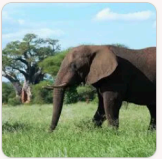


How many legs are shown on the elephant in this image? (a) 3 (b) 4



How many legs are shown on the elephant in this image? (a) 3 (b) 4

Looking at the image, I can see **(b) 4** legs on the elephant. All four legs are visible as the elephant stands in the grassy savanna.



How many legs are shown on the elephant in this image? (a) 3 (b) 4

Looking at the image, I can see **(b) 4** legs on the elephant. All four legs are visible as the elephant stands in the grassy savanna.

The answer is (a) 3 — but Claude Sonnet 4.6 (March 2026) still gets this wrong.

How do you diagnose a vision encoder?

Finding systematic failures in a representation is hard.

You cannot manually inspect every image.

Idea: use a reference model to find what CLIP misses.

CLIP-blind pairs

Find image pairs where:

CLIP says: "same"

cosine similarity > 0.95

DINOv2 says: "different"

cosine similarity < 0.6

Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs, Tong et al.

CLIP-blind pairs

Find image pairs where:

CLIP says: "same"

cosine similarity > 0.95

DINOv2 says: "different"

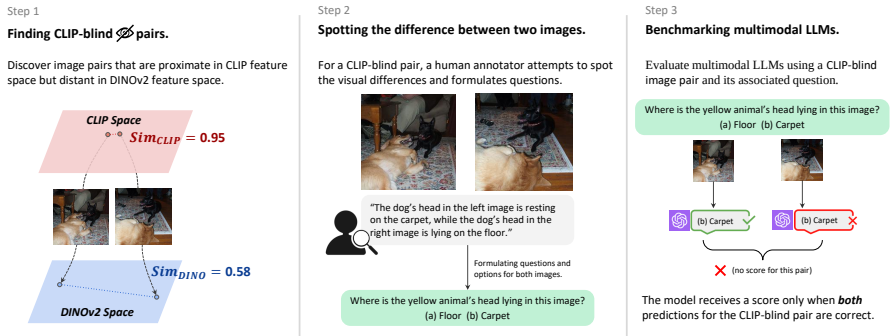
cosine similarity < 0.6

These are **CLIP-blind pairs**.

CLIP literally cannot tell them apart.

Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs, Tong et al.

From pairs to benchmark: MMVP

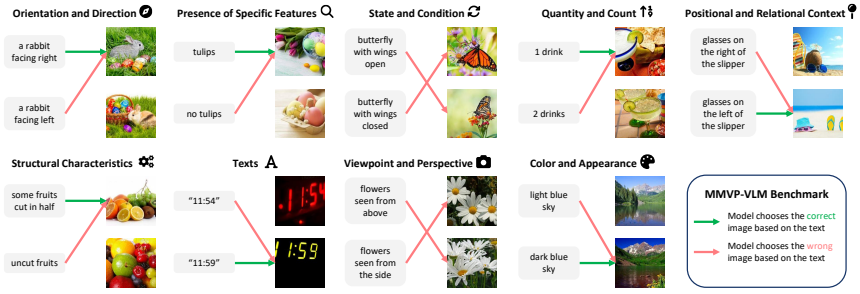


For each CLIP-blind pair, craft a VQA question targeting the visual difference.

150 pairs, 300 questions. A pair is correct only if **both** answers are right.

Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs, Tong et al.

What exactly is CLIP missing?



Orientation, counting, spatial relations, color, text, viewpoint...

Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs, Tong et al.

CLIP captures **semantics** — categories, concepts, meaning

but misses **visual details** — orientation, counting, spatial layout

Language-supervised models are **semantic**, not **visual**.

Two ways to learn visual representations

Self-Supervised

Learn from **images alone**

DINO, DINOv2, MAE, I-JEPA

No text, no labels

?

Language-Supervised

Learn from **image-text pairs**

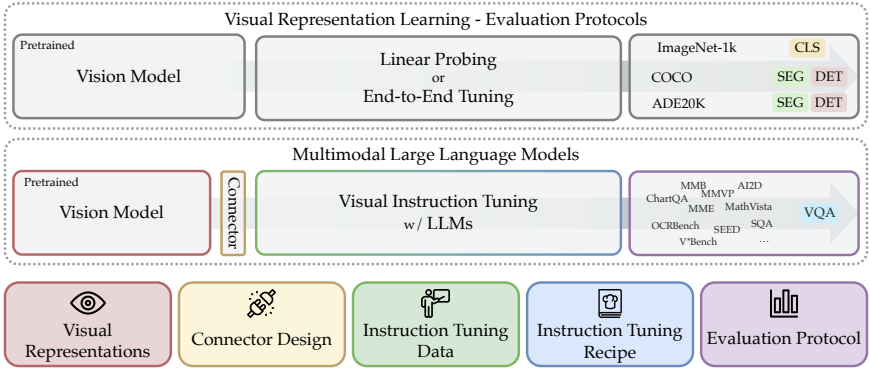
CLIP, SigLIP, OpenCLIP

Web-scale captions

$\rightarrow$ **Semantic**

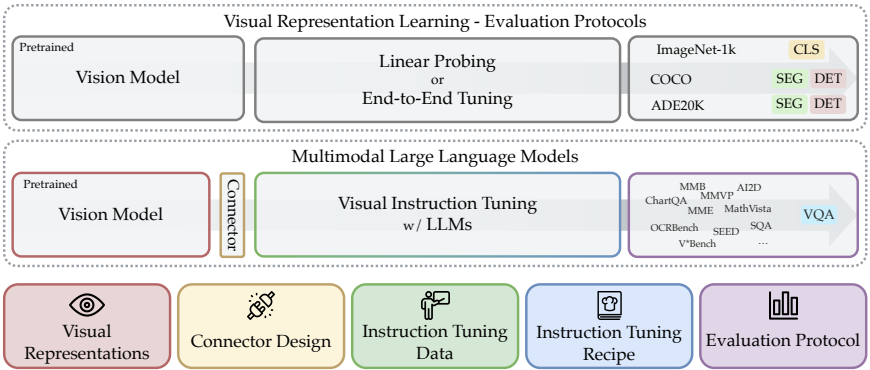
What else characterizes these representations?

Beyond conventional evaluation



Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs, Tong et al.

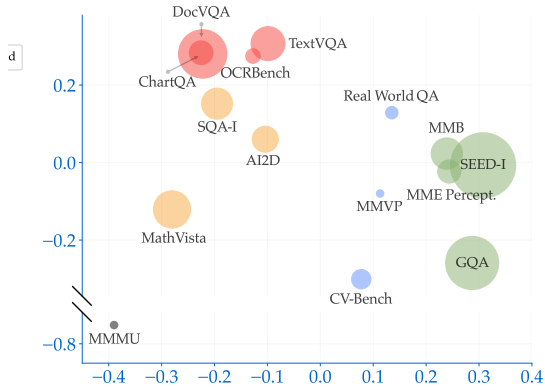
Beyond conventional evaluation



Freeze the encoder, attach an LLM — the representation is the only variable

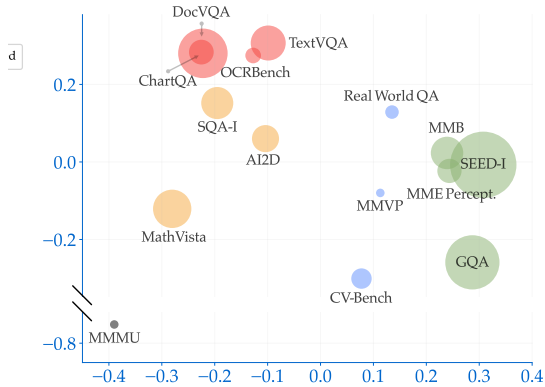
Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs, Tong et al.

Benchmarks cluster into 4 groups



Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs, Tong et al.

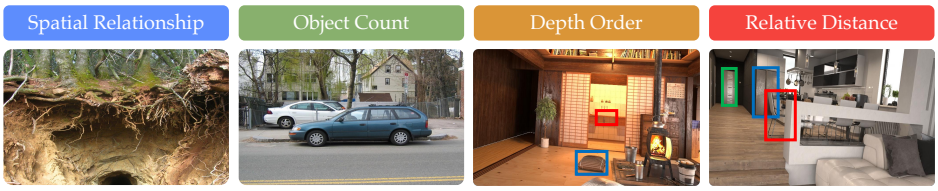
Benchmarks cluster into 4 groups



General **Knowledge** **Chart & OCR** **Vision-Centric**

Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs, Tong et al.

Cambrian Vision-Centric Benchmark (CV-Bench)



Repurpose classical CV tasks into VQA format
— 2,638 examples across 4 vision-centric tasks

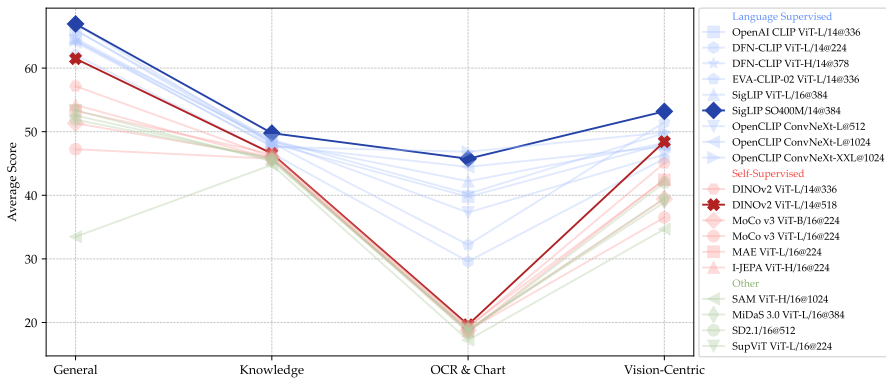
Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs, Tong et al.

Comparing 23 vision encoders

SSL and language-supervised encoders
plus specialized encoders like diffusion, depth, segmentation

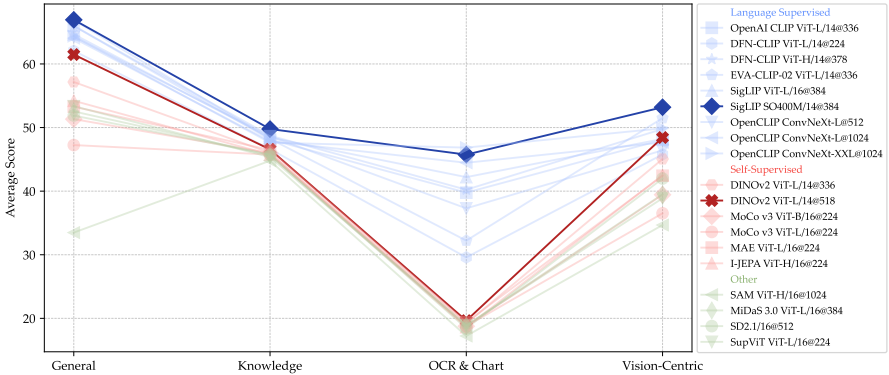
Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs, Tong et al.

What do representations actually capture?



Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs, Tong et al.

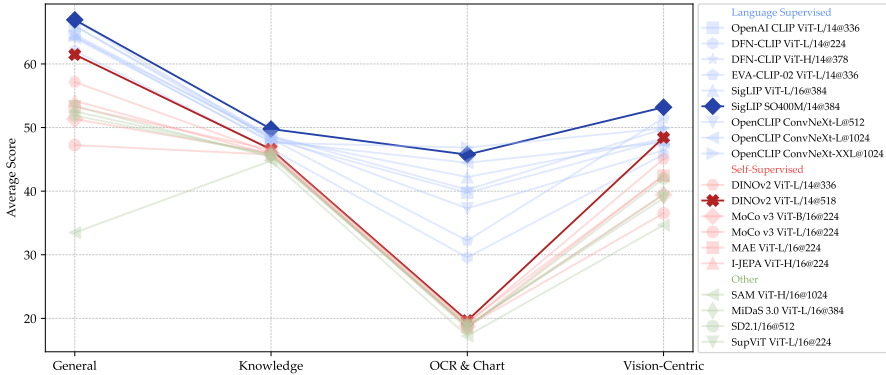
What do representations actually capture?



Language-supervised encoders lead overall, especially on OCR & Chart

Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs, Tong et al.

What do representations actually capture?



But SSL encoders are competitive on vision-centric tasks

Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs, Tong et al.

Two ways to learn visual representations

Self-Supervised

Learn from **images alone**

DINO, DINOv2, MAE, I-JEPA

No text, no labels

→ **Visual**

Competitive on vision-centric

Language-Supervised

Learn from **image-text pairs**

CLIP, SigLIP, OpenCLIP

Web-scale captions

→ **Semantic**

Leads on multimodal tasks

Two ways to learn visual representations

Self-Supervised

Learn from **images alone**

DINO, DINOv2, MAE, I-JEPA

No text, no labels

→ **Visual**

Competitive on vision-centric

Language-Supervised

Learn from **image-text pairs**

CLIP, SigLIP, OpenCLIP

Web-scale captions

→ **Semantic**

Leads on multimodal tasks

But wait — the data is completely different!

The confound

SSL training data

ImageNet: 1.3M, LVD-142M

Always <1B images

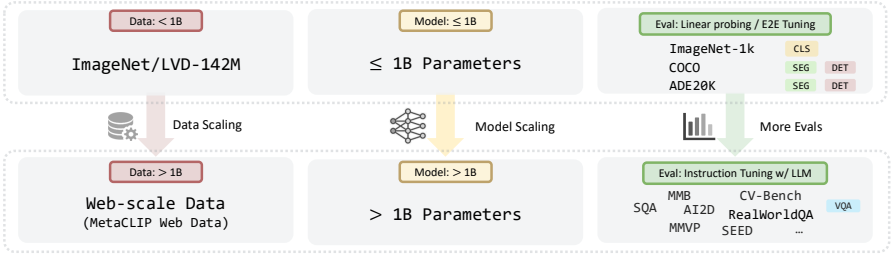
CLIP training data

Web-scale: **billions** of images

10× data gap.

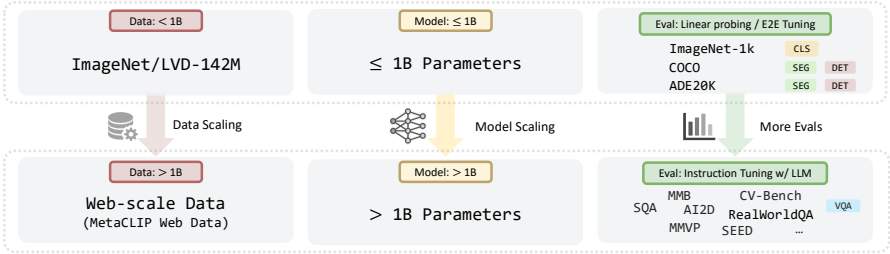
Not a fair comparison.

Fair comparison: same data, different supervision



1. Data: train both SSL and CLIP on the same web data

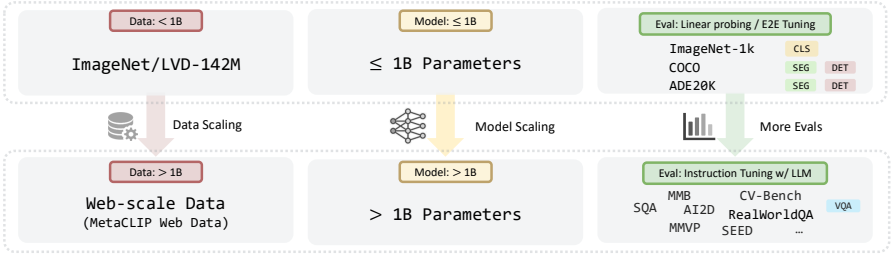
Fair comparison: same data, different supervision



2. Model: scale both to the same sizes (1B–7B)

Scaling Language-Free Visual Representation Learning, Fan et al.

Fair comparison: same data, different supervision



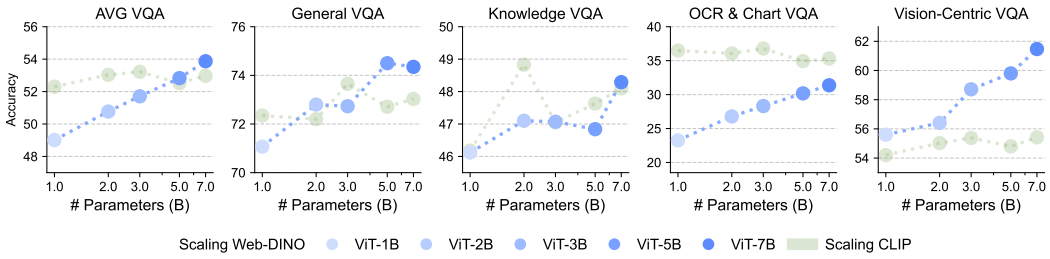
3. Evaluation: evaluate with the same MLLM instruction tuning

Model scaling

Fix data, increase **model size** from 1B to 7B

How does each supervision type scale with capacity?

Model scaling



SSL consistently improves with model size

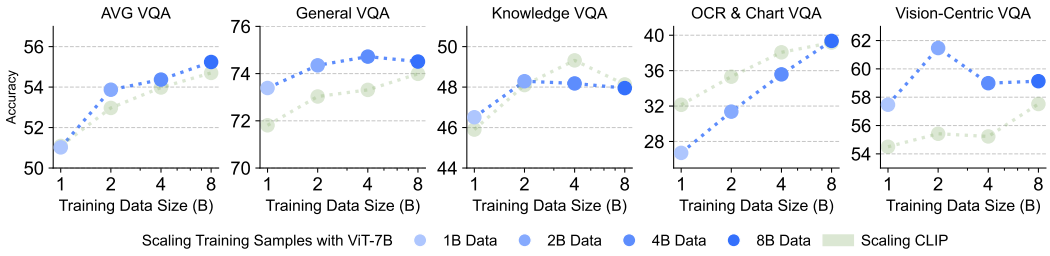
Scaling Language-Free Visual Representation Learning, Fan et al.

Data scaling

Fix model, increase **data** from 1B to 8B samples

How does each supervision type scale with data?

Data scaling



SSL improves consistently with more data

Scaling Language-Free Visual Representation Learning, Fan et al.

Why does scaling improve OCR & Chart?

SSL has **no text supervision** — yet scaling improves OCR & Chart

Why?

Web images naturally contain text

Web data is not ImageNet — images contain labels, signs, charts, documents

Light filter (50.3%): images containing any text.

Heavy filter (1.3%): charts, tables, documents only.

Data composition is the key lever

Method	% Data	AVG	Vision	OCR
CLIP 2B	100%	53.0	55.0	36.1
Web-DINO 2B	100%	50.8	56.4	26.8
+ Light filter	50.3%	53.4	55.6	33.2
+ Heavy filter	**1.3%**	**53.7**	56.2	**40.4**

With only

1.3%

of data

SSL surpasses CLIP on OCR & Chart

Scaling Language-Free Visual Representation Learning, Fan et al.

Two ways to learn visual representations

Self-Supervised

Learn from **images alone**

DINO, DINOv2, MAE, I-JEPA

No text, no labels

→ **More visual**

Language-Supervised

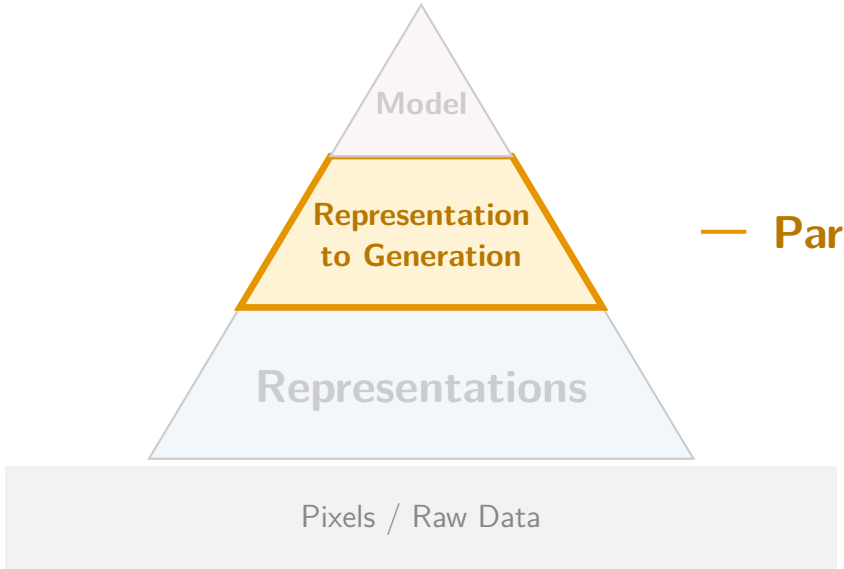
Learn from **image-text pairs**

CLIP, SigLIP, OpenCLIP

Web-scale captions

→ **More semantic**

But **data** is the major player — the elephant in the room

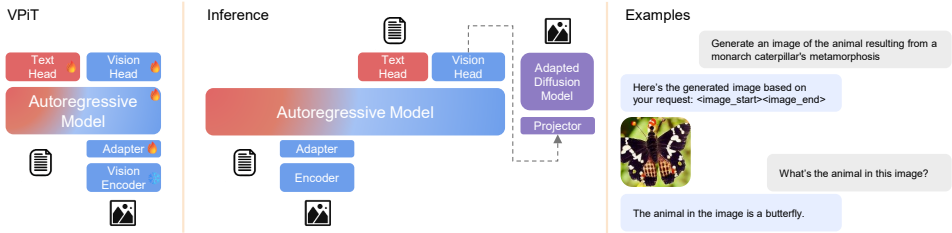

Model
Representation
to Generation
Representations
Pixels / Raw Data

— Part II

We **understand** images through representations.

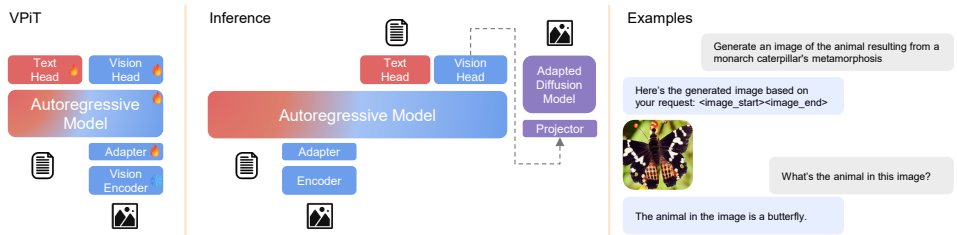
Can we **generate** representations?

The setup: MetaMorph



MetaMorph: Multimodal Understanding and Generation via Instruction Tuning, Tong et al.

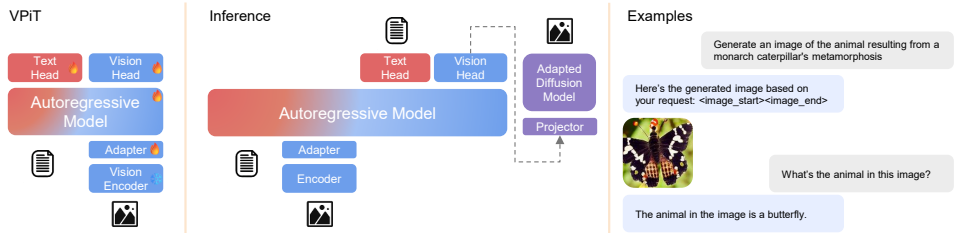
The setup: MetaMorph



Train an autoregressive model to predict both text and visual tokens

MetaMorph: Multimodal Understanding and Generation via Instruction Tuning, Tong et al.

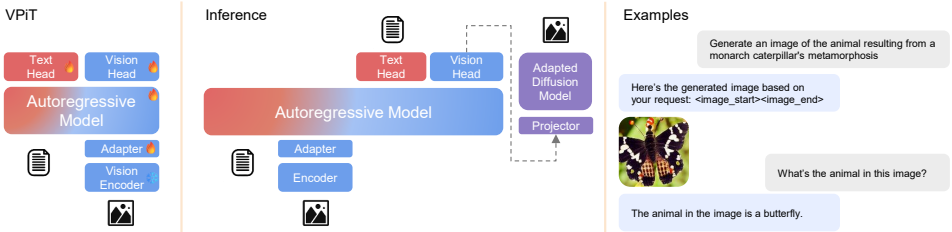
The setup: MetaMorph



Text: next-token prediction with cross-entropy
Vision: next-token prediction with regression

MetaMorph: Multimodal Understanding and Generation via Instruction Tuning, Tong et al.

The setup: MetaMorph



To visualize, train a separate diffusion decoder on the predicted visual tokens

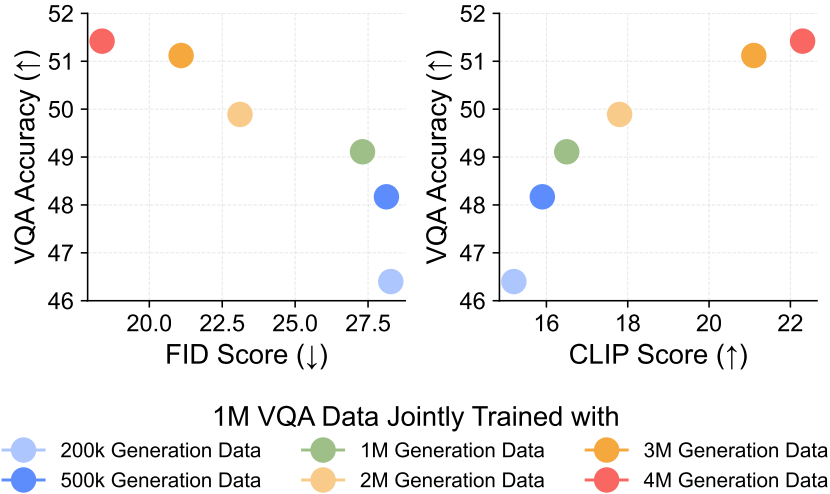
Experiment design

To understand the interplay:

Exp A: Fix understanding data, increase generation data

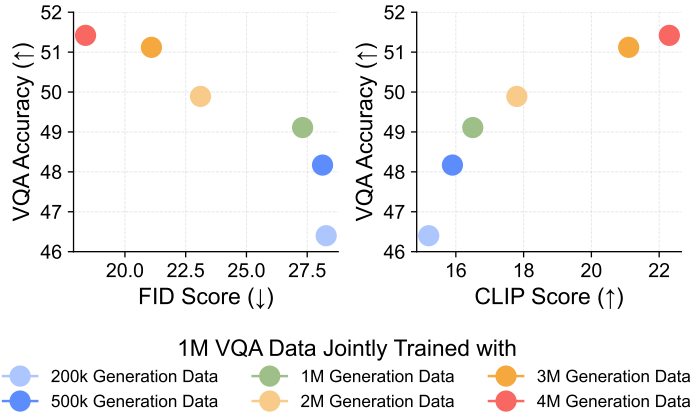
Exp B: Fix generation data, increase understanding data

More generation data



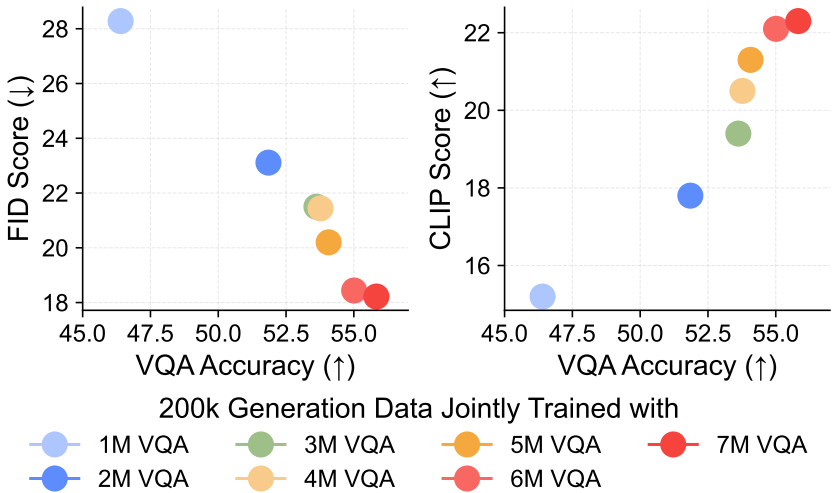
MetaMorph: Multimodal Understanding and Generation via Instruction Tuning, Tong et al.

More generation data



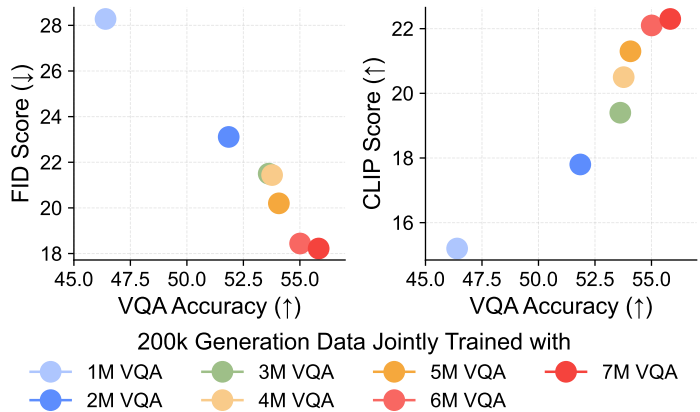
More generation data improves both generation and understanding

More understanding data



MetaMorph: Multimodal Understanding and Generation via Instruction Tuning, Tong et al.

More understanding data



More understanding data also improves generation

Understanding and generation are **mutually beneficial**

Implicit reasoning in generation

Prompt	Step-by-Step Logic Chain (For Reference)	Solution Examples	SD3.5-8B	Janus	MetaMorph
"The national flag of the country where Yellowstone National Park is located"	Yellowstone National Park's Location → America → American Flag				
"The flower celebrated in spring festivals in the country where sushi originated"	Sushi's Origin → Japan → Flower in Spring Festivals in Japan → Cherry Blossom (Sakura)				
"The large mammal that shares its name with a constellation often visible in the night sky and associated with the northern part of the world"	Constellation Associated with the Northern Sky → Ursa Major → Ursa (Latin for 'Bear') → Large Mammal Named 'Bear'				
"A musical instrument, this instrument is often played by the scientist who formulated the theory of special relativity"	Scientist Who Formulated Special Relativity → Albert Einstein → Instrument Often Played by Einstein → Violin				
"The animal associated with having (2+7) lives"	2+7 → 9 → Animal Believed to Have 9 lives → Cat				

MetaMorph: Multimodal Understanding and Generation via Instruction Tuning, Tong et al.

Implicit reasoning in generation

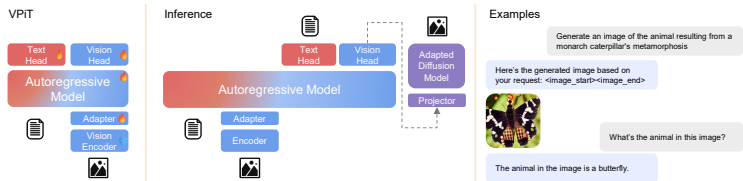
Prompt	Step-by-Step Logic Chain (For Reference)	Solution Examples	SD3.5-8B	Janus	MetaMorph
"The national flag of the country where Yellowstone National Park is located"	Yellowstone National Park's Location → America → American Flag				
"The flower celebrated in spring festivals in the country where sushi originated"	Sushi's Origin → Japan → Flower in Spring Festivals in Japan → Cherry Blossom (Sakura)				
"The large mammal that shares its name with a constellation often visible in the night sky and associated with the northern part of the world"	Constellation Associated with the Northern Sky → Ursa Major → Ursa (Latin for 'Bear') → Large Mammal Named 'Bear'				
"A musical instrument, this instrument is often played by the scientist who formulated the theory of special relativity"	Scientist Who Formulated Special Relativity → Albert Einstein → Instrument Often Played by Einstein → Violin				
"The animal associated with having (2+7) lives"	2+7 → 9 → Animal Believed to Have 9 lives → Cat				

Understanding flows into generation through representations

MetaMorph: Multimodal Understanding and Generation via Instruction Tuning, Tong et al.

MetaMorph predicts representations with **regression loss** — deterministic, no sampling

And to visualize, it still needs a **separate diffusion model** to decode to pixels



MetaMorph: Multimodal Understanding and Generation via Instruction Tuning, Tong et al.

The diffusion decoder:

Nondeterministic — stochastic sampling can hallucinate details

Relies on VAE embeddings — not the same space as the representation

Adds complexity — a whole separate model to train and maintain

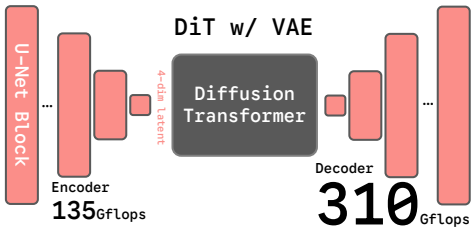
Can we generate directly in representation space?

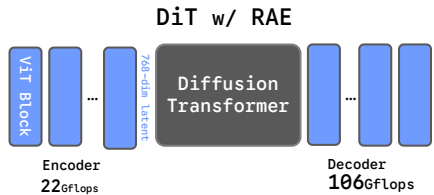
Can we generate directly in representation space?

Step 1: Can we **decode** representations back to pixels?

Step 2: Can we **sample** new representations?

Representation Autoencoder (RAE)





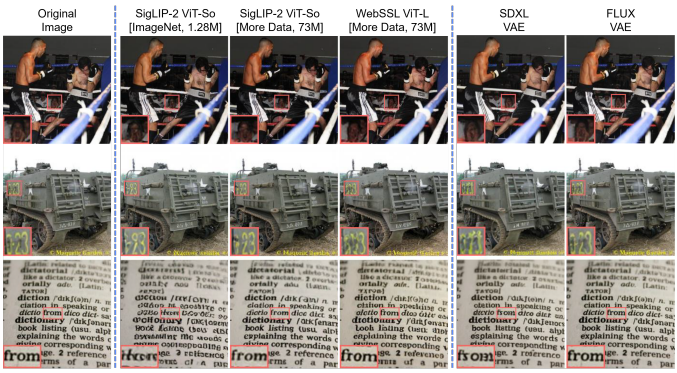
Diffusion Transformers with Representation Autoencoders, Zheng et al.

Step 1: Decoding representations back to pixels

Freeze the pretrained encoder, train a lightweight ViT decoder

Loss: ℓ_1 + LPIPS + GAN

Step 1: Decoding representations back to pixels



	rFID ↓
RAE (DINOv2)	**0.49**
RAE (MAE)	**0.16**
SD-VAE	0.62

Diffusion Transformers with Representation Autoencoders, Zheng et al.

We can encode and decode representations

What does it take to **diffuse** in this high-dimensional space?

What does it take to diffuse in this high-dimensional space?

Standard flow matching on RAE latents fails

768-dim tokens vs. 4-dim VAE tokens — a very different regime

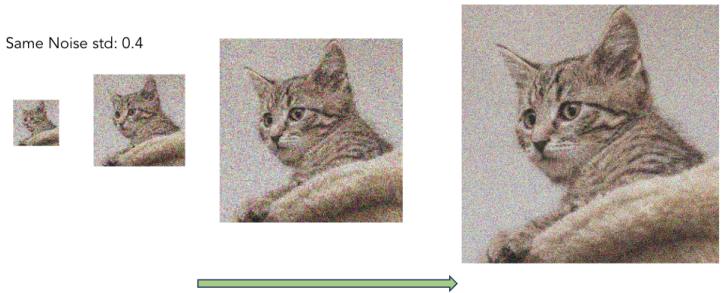
Standard flow matching on RAE latents

FID on ImageNet 256×256 (80 epochs, no guidance)

	RAE (DINOv2-B)	SD-VAE
DiT-S	**215.8**	51.7
DiT-XL	23.1	7.1

Standard recipe does not work — DiT-S completely fails, DiT-XL far behind VAE

Why? The noise schedule



Same noise (σ=0.4) corrupts low-dim representations much more than high-dim

$$t_m = \frac{\alpha\, t_n}{1 + (\alpha - 1)\, t_n}, \quad \alpha = \sqrt{\frac{m}{n}}$$

Shift the schedule based on the effective data dimension m

Diffusion Transformers with Representation Autoencoders, Zheng et al.

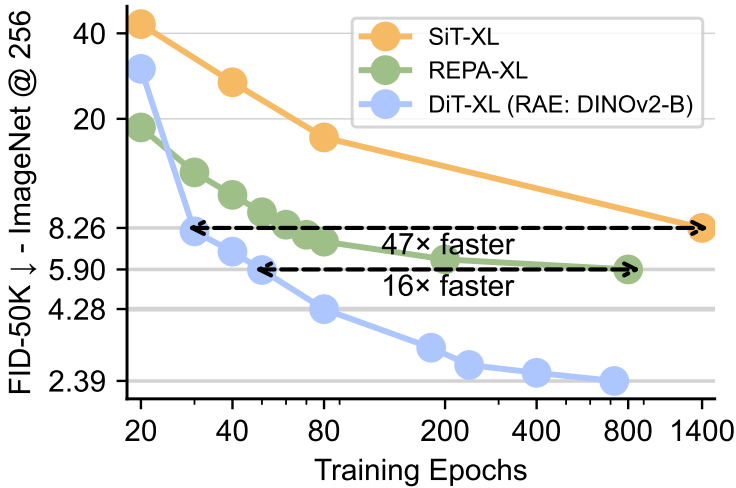
Dimension-dependent noise schedule: ImageNet

FID on ImageNet (DiT-XL, 80 epochs)

Setting	gFID ↓
w/o schedule shift	23.1
w/ schedule shift	**4.8**

Shifting the noise schedule for high dimensionality is critical

RAE: convergence on ImageNet

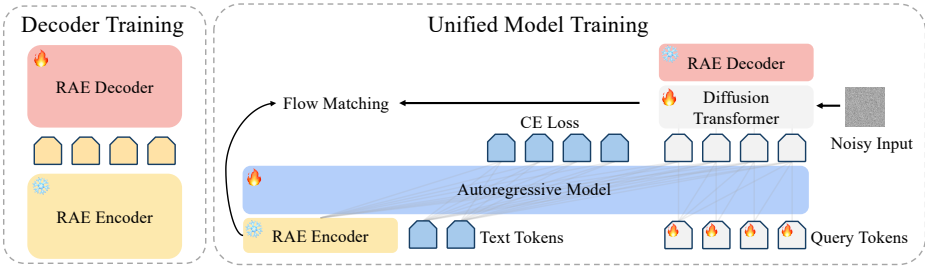


Diffusion Transformers with Representation Autoencoders, Zheng et al.

Diffusion in **high-dimensional representation space**
works —
with the right **noise schedule**

Does it scale to text-to-image?

Scaling RAE to text-to-image



Frozen RAE encoder + learned decoder, paired with an LLM and a Diffusion Transformer

Scaling Text-to-Image Diffusion Transformers with Representation Autoencoders, Tong et al.

Dimension-dependent noise schedule: text-to-image

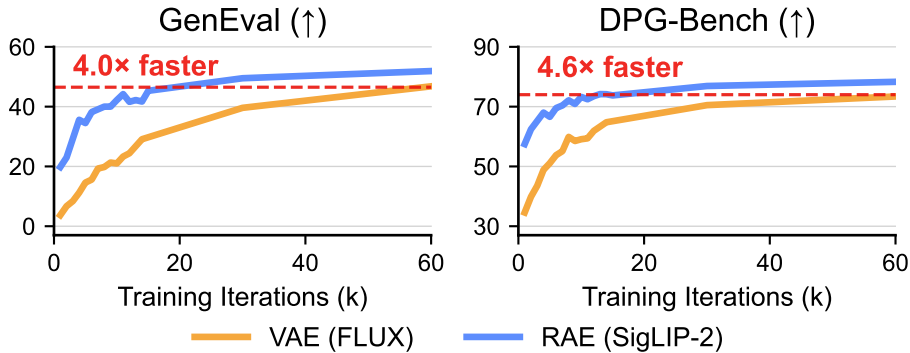
Text-to-image (Qwen 1.5B + DiT 2.4B, 70M data seen)

Setting	GenEval ↑	DPG-Bench ↑
w/o schedule shift	23.6	54.8
w/ schedule shift	**49.6**	**76.8**

Equally critical at text-to-image scale

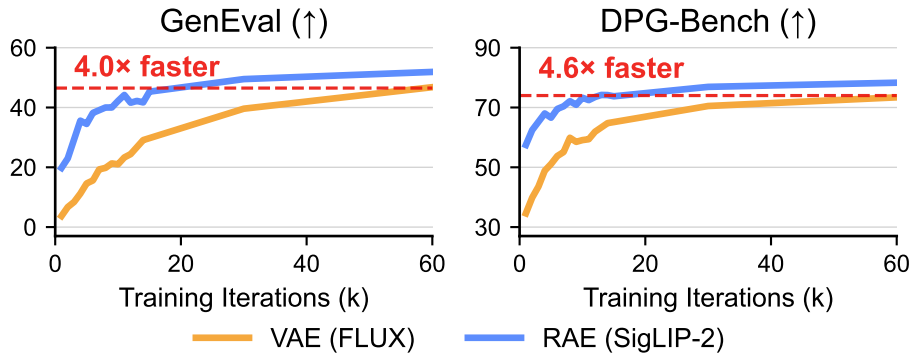
Scaling Text-to-Image Diffusion Transformers with Representation Autoencoders, Tong et al.

RAE: diffusing in representation space



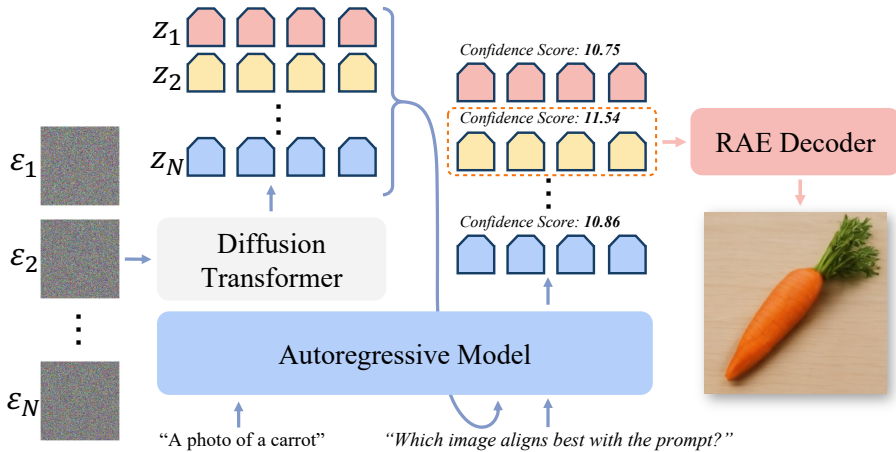
Scaling Text-to-Image Diffusion Transformers with Representation Autoencoders, Tong et al.

RAE: diffusing in representation space



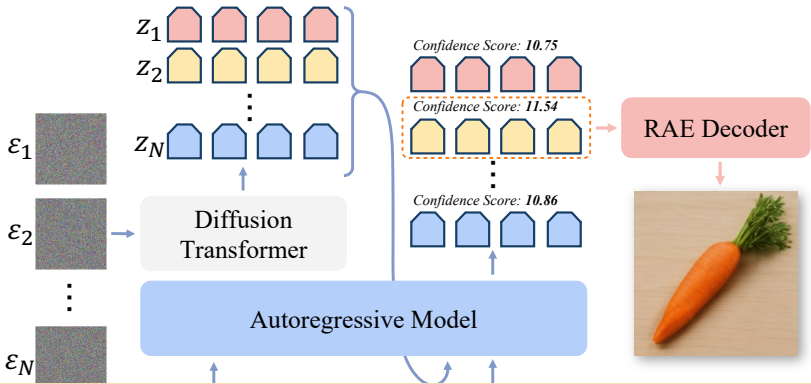
RAE converges significantly faster than VAE

Test-time scaling in representation space



Scaling Text-to-Image Diffusion Transformers with Representation Autoencoders, Tong et al.

Test-time scaling in representation space



Generate and evaluate multiple candidates directly in representation space — no decoding to pixels needed

Scaling Text-to-Image Diffusion Transformers with Representation Autoencoders, Tong et al.

Test-time scaling in representation space

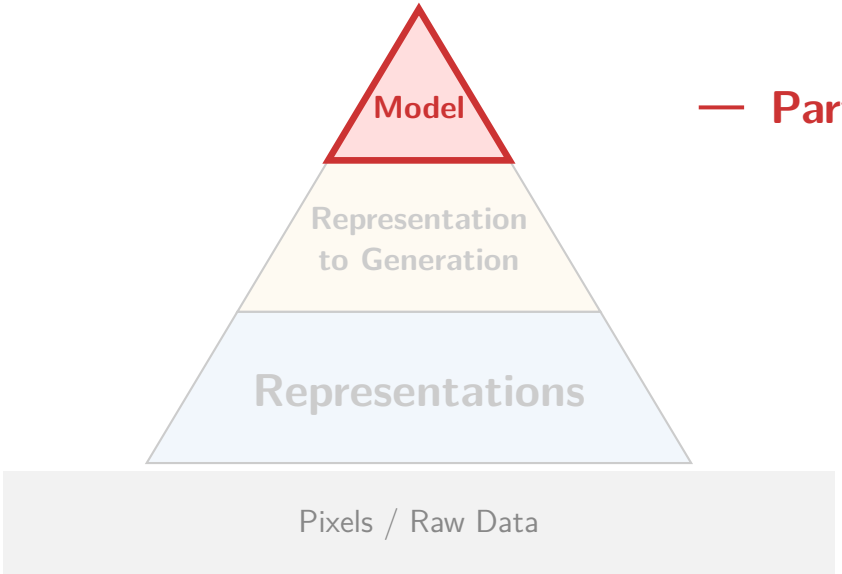
GenEval scores with best-of-N selection in latent space

Best-of-N	Prompt Confidence	Answer Logits
1.5B LLM + 5.5B DiT	(baseline GenEval = 53.2)	
4/8	56.7	59.6
4/32	60.0	64.3
7.0B LLM + 5.5B DiT	(baseline GenEval = 55.5)	
4/8	58.3	62.5
4/32	60.1	**67.8**

Test-time scaling works entirely in representation space

Scaling Text-to-Image Diffusion Transformers with Representation Autoencoders, Tong et al.

Understanding and generation operate in a
shared representation space


Model
Representation
to Generation
Representations
Pixels / Raw Data

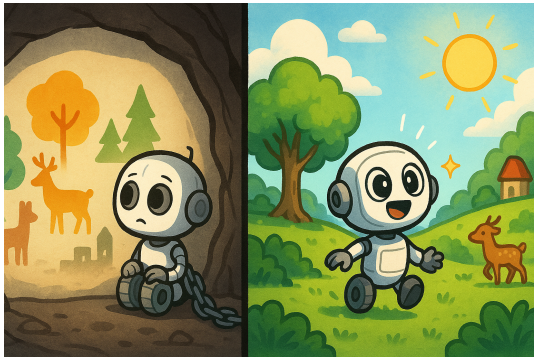
— Part III

We have representations.

We have understanding and generation.

Now let's put it all together: **one model**.

We are in the era of language models.



". . . prisoners chained in a cave,

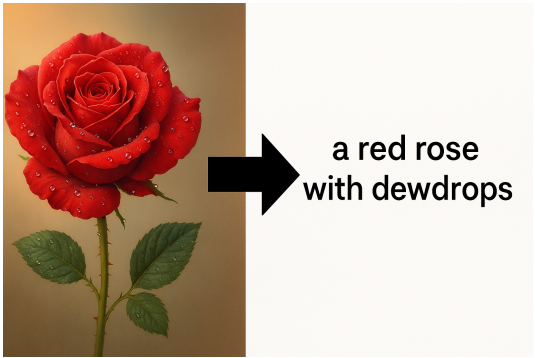
watching shadows on a wall, which they mistake for reality."

— Plato, The Republic, Book VII

Are today's language models
prisoners in the cave?

They read about the world but never *see* it.

Text is all they know — shadows on the wall.



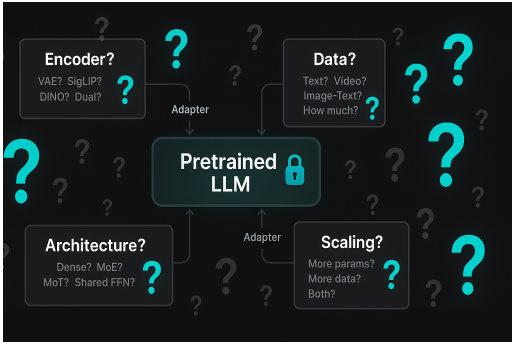
Text is a lossy compression of reality

High-quality text is finite and approaching exhaustion

The visual world is vast and largely untapped

Can we go beyond language?

The landscape is messy



Most multimodal models start from a **pretrained LLM**.

This creates a bias: you *must* preserve language capability.

- Design choices shaped by not breaking what's there
- Conclusions confounded — learned vs. inherited?

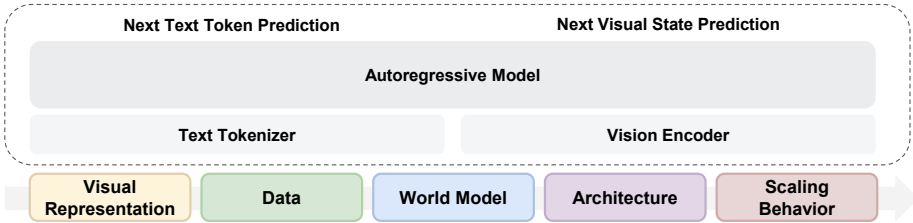
Our approach:

Train from scratch

Change one thing at a time

Measure everything

Everything is a sequence

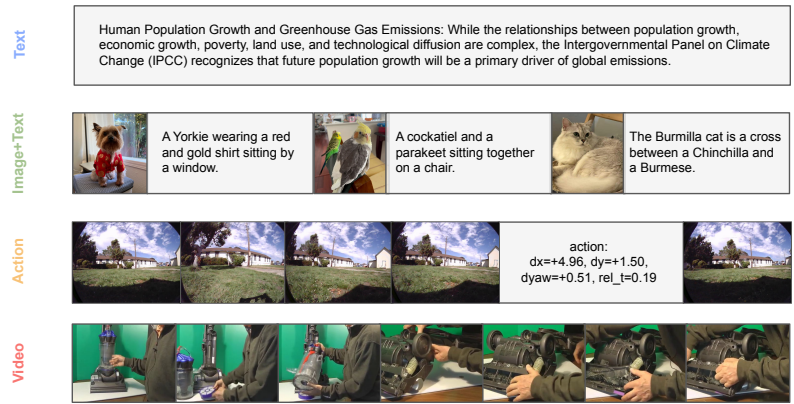


T next-token prediction + I flow matching

Any combination of modalities, any order.

Beyond Language Modeling: An Exploration of Multimodal Pretraining, Tong et al.

All data as sequential tasks



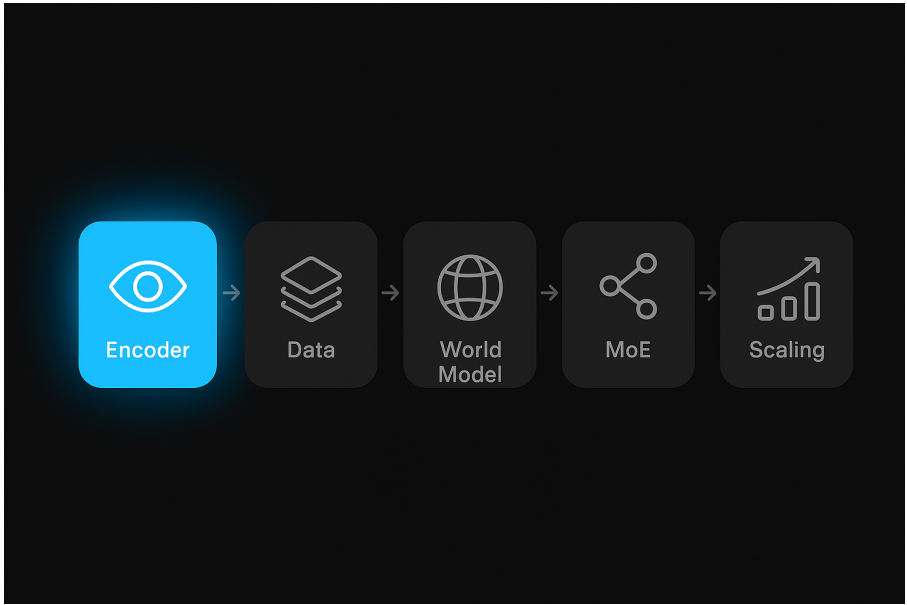
Text	Video	Image-Text	Action Video
T→T	I→I→I	I→T or T→I	I+T→I

Beyond Language Modeling: An Exploration of Multimodal Pretraining, Tong et al.

Five design axes, studied one at a time

1. Encoder
2. Data
3. World Modeling
4. Architecture
5. Scaling

Beyond Language Modeling: An Exploration of Multimodal Pretraining, Tong et al.



Beyond Language Modeling: An Exploration of Multimodal Pretraining, Tong et al.

Do you need two encoders?

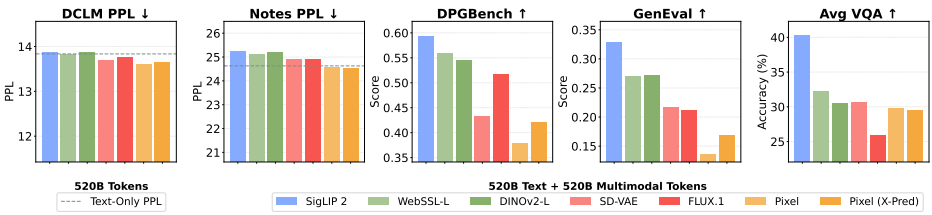
From Part II: RAE outperforms VAE for generation.

Does this hold when training **from scratch**, co-training with language?

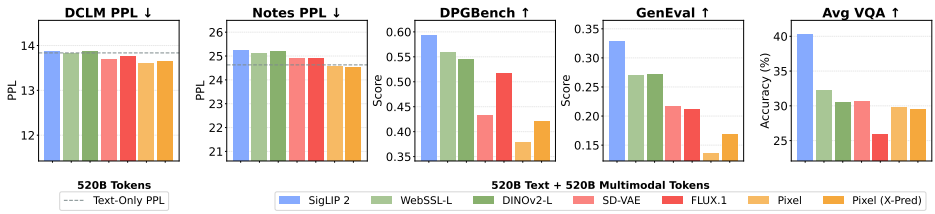
Compare 6 visual representations — VAEs, semantic encoders, raw pixels — all else fixed.

Beyond Language Modeling: An Exploration of Multimodal Pretraining, Tong et al.

Do you need two encoders?

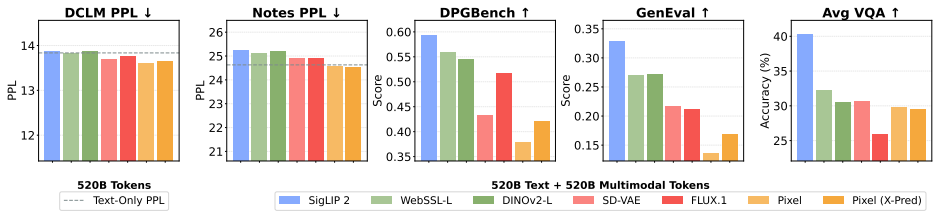


Do you need two encoders?



Semantic encoders (RAE) outperform VAEs even on generation

Do you need two encoders?

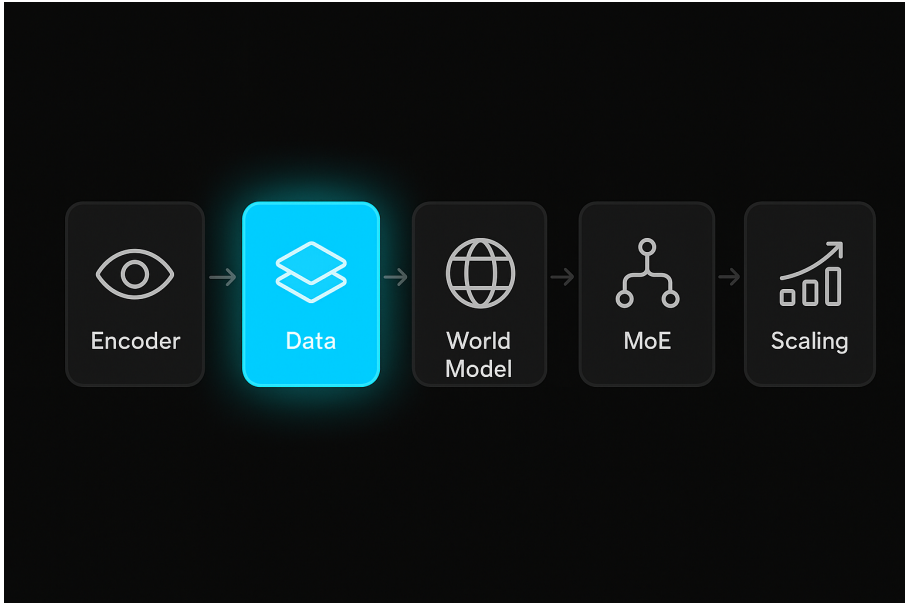


A single representation encoder shows strong potential for both understanding and generation

We have a good encoder. Now:

What data should we feed it?

Text, video, image-text pairs, actions —
how do they interact?



Beyond Language Modeling: An Exploration of Multimodal Pretraining, Tong et al.

"Adding vision will degrade my language model."

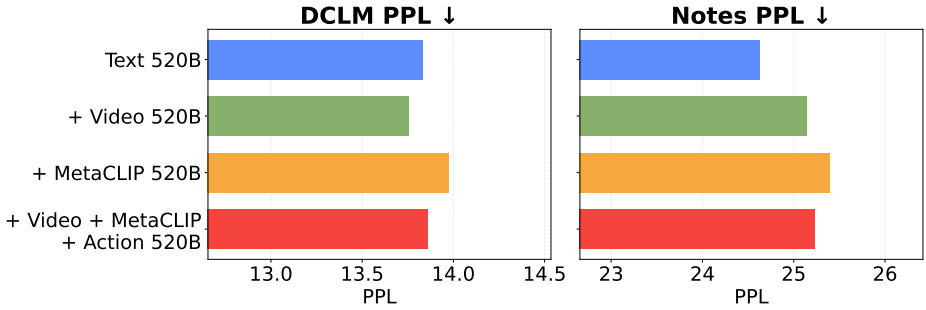
The #1 concern about multimodal pretraining. Let's test it.

Experiment: data mixtures

Three data mixtures, all at $\sim$1T tokens, compared to text-only:

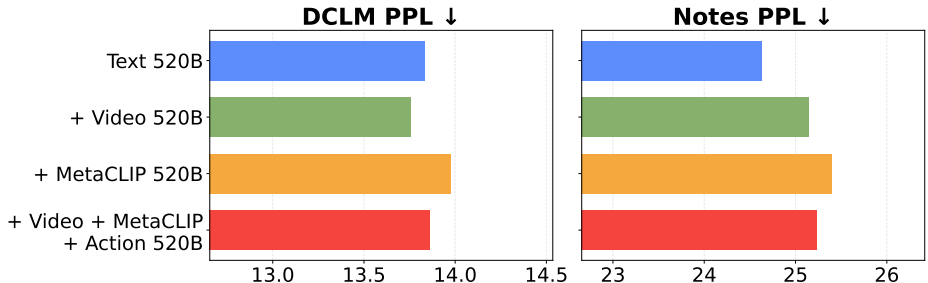
1. Text + Video (raw video, no captions)
2. Text + MetaCLIP (image-text pairs)
3. Text + Video + MetaCLIP + Action (everything)

Text perplexity across data mixtures



Beyond Language Modeling: An Exploration of Multimodal Pretraining, Tong et al.

Text perplexity across data mixtures

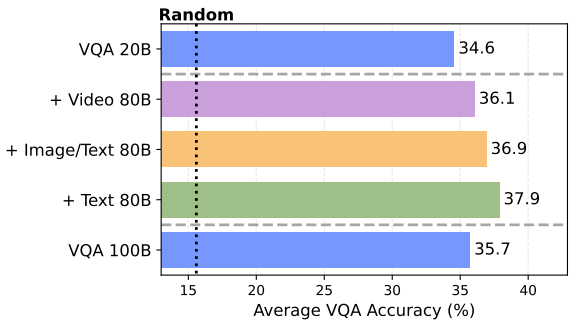


Text+Video matches text-only perplexity on in-distribution data

Text+MetaCLIP slightly worse — the tax comes from caption distribution, not vision

Vision minimally impacts language. Does it actually **help**?

Cross-modal synergy



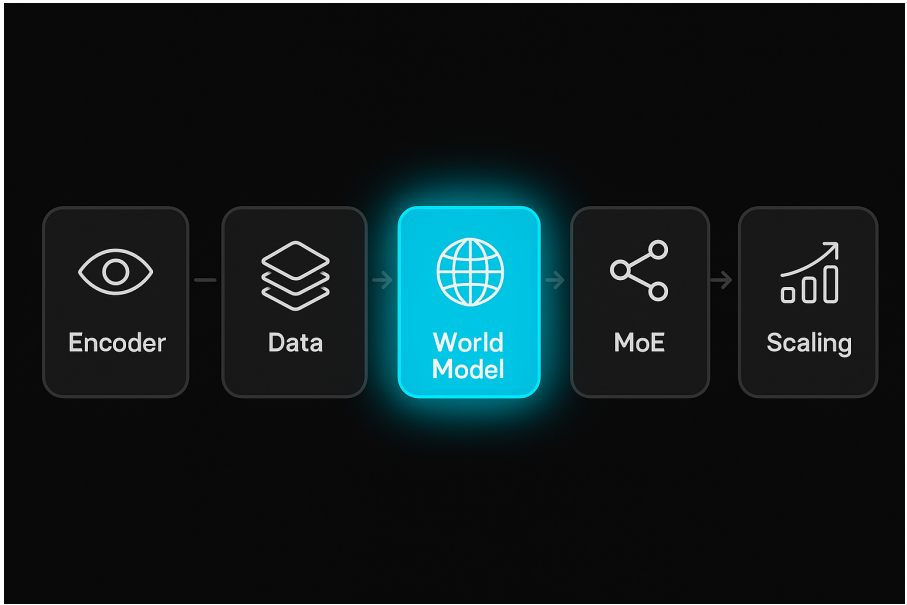
20B VQA + 80B diverse outperforms 100B VQA-only

5× less domain data!

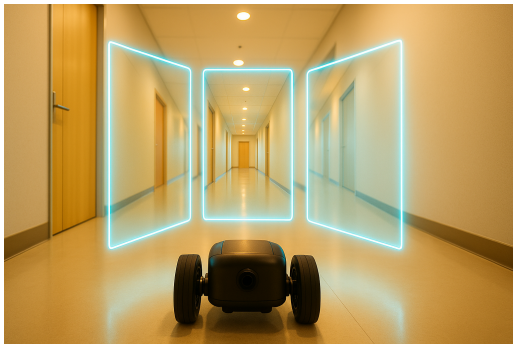
Diverse pretraining is more data-efficient than task-specific scaling.

Beyond Language Modeling: An Exploration of Multimodal Pretraining, Tong et al.

If a model sees enough of how the world looks and moves, does it learn how the world *works*?



Beyond Language Modeling: An Exploration of Multimodal Pretraining, Tong et al.

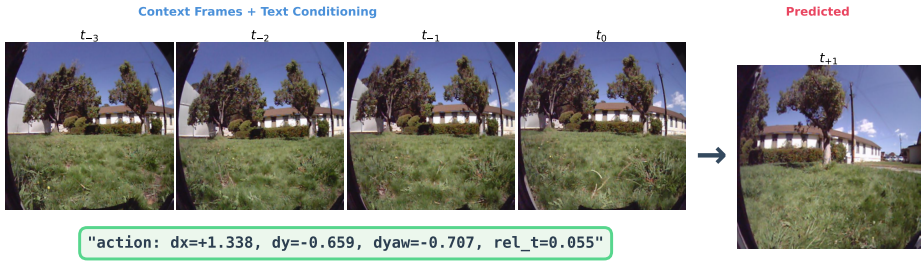


If diverse pretraining outperforms domain-specific scaling...

Does world modeling emerge for free?

Beyond Language Modeling: An Exploration of Multimodal Pretraining, Tong et al.

Navigation as a sequential task



$\mathrm{I} + \mathrm{T} \rightarrow \mathrm{I}$ Actions = **text tokens**. No special adapters. No architecture changes.

Beyond Language Modeling: An Exploration of Multimodal Pretraining, Tong et al.

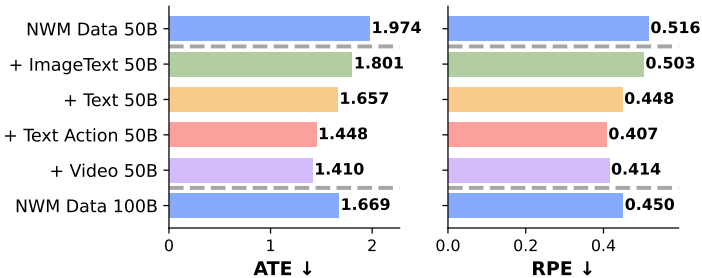
What data helps world modeling?

Does world modeling come from domain-specific navigation data, or from broader multimodal pretraining?

Experiment: fix total budget at 100B tokens, compare:

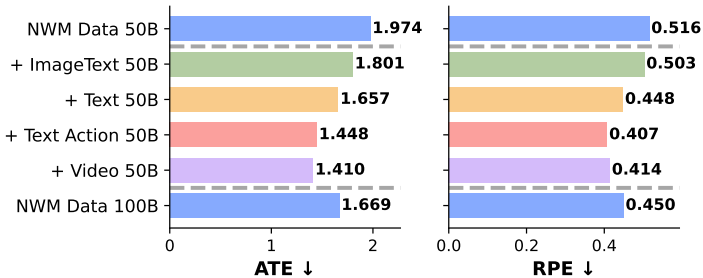
- 100B NWM-only (domain-specific baseline)
- 50B NWM + 50B of various multimodal data

What data helps world modeling?



Beyond Language Modeling: An Exploration of Multimodal Pretraining, Tong et al.

What data helps world modeling?



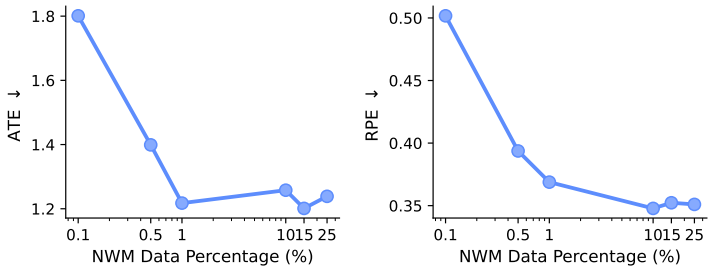
Video helps more than doubling NWM data —
world modeling relies on multimodal pretraining

How much domain data do you need?

General video helps. But how much domain-specific NWM data is actually needed?

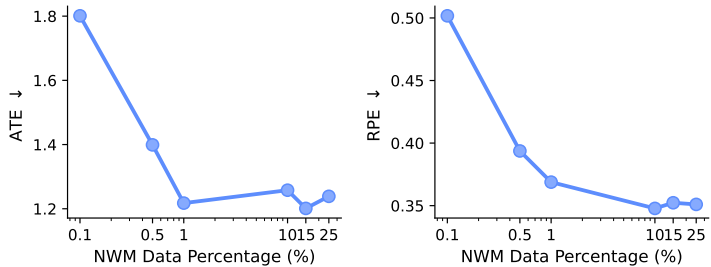
Experiment: fix total budget at 200B tokens,
vary NWM fraction from 0.1% to 50%.

How much domain data do you need?



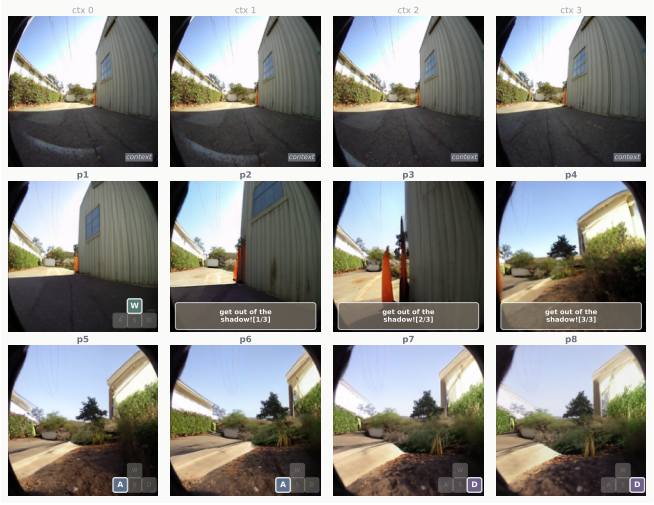
Beyond Language Modeling: An Exploration of Multimodal Pretraining, Tong et al.

How much domain data do you need?



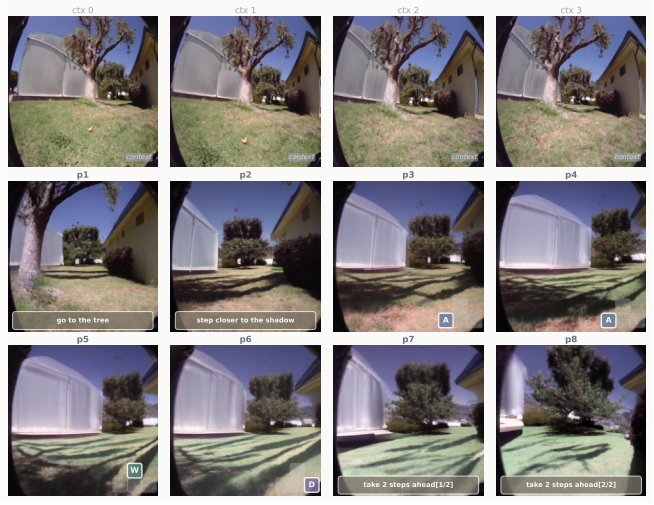
Saturates at 1% — world modeling requires minimal domain data given diverse pretraining

Zero-shot navigation rollouts



Beyond Language Modeling: An Exploration of Multimodal Pretraining, Tong et al.

Language-driven navigation

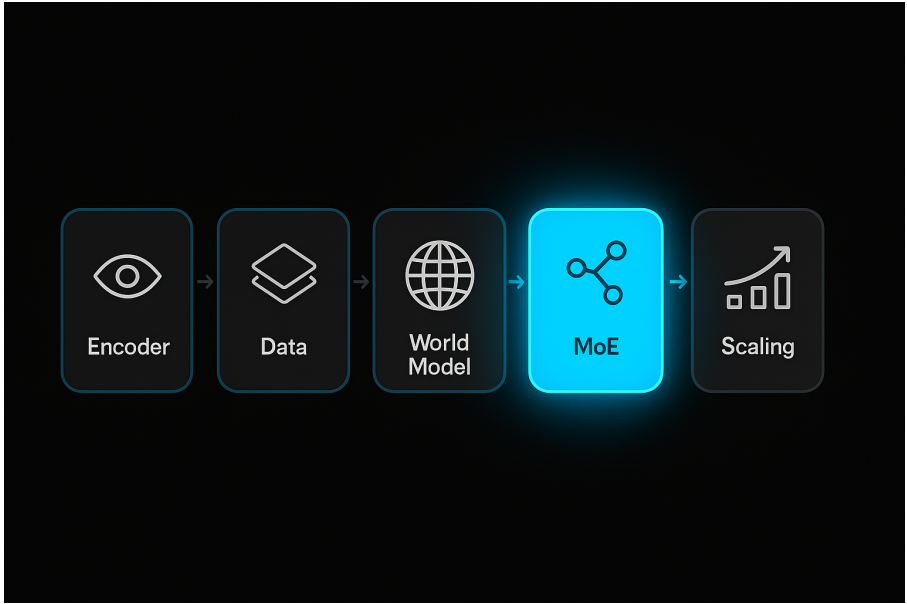


Beyond Language Modeling: An Exploration of Multimodal Pretraining, Tong et al.

So the framework is flexible —
one encoder, diverse data, emergent world modeling.

But there's still a practical problem:

How do you allocate model capacity
across modalities?



Beyond Language Modeling: An Exploration of Multimodal Pretraining, Tong et al.

The capacity problem

Vision and language need different parameters. How to allocate?

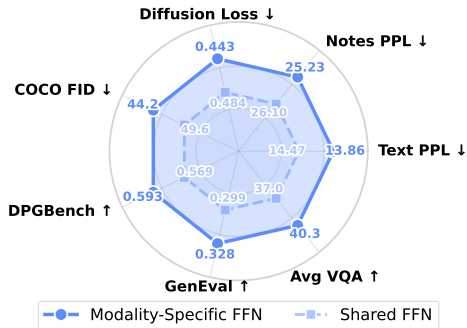
- **Shared FFN** — same weights for everything
- **Modality-specific FFN** — fixed 50/50 split
- **MoE** — learn the split from data

Even simple separation helps

Shared → modality-specific FFN:

- Better perplexity
- Better generation
- Better VQA
- No extra inference cost

But 50/50 is arbitrary — can MoE learn a better allocation?



MoE: how fine-grained should experts be?

Few big experts vs. **Many small experts**

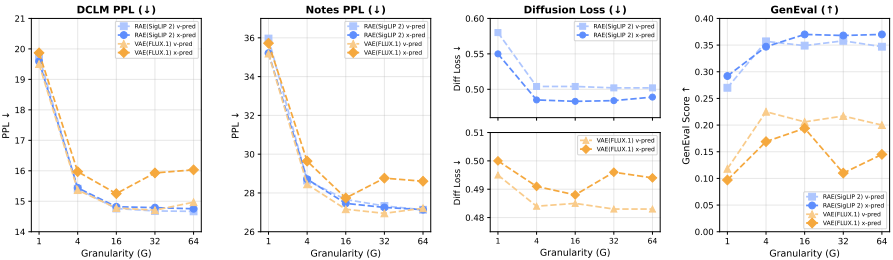
e.g. pick 1 of 16 large experts vs. pick 16 of 256 small experts

Same total parameters. Same active compute.

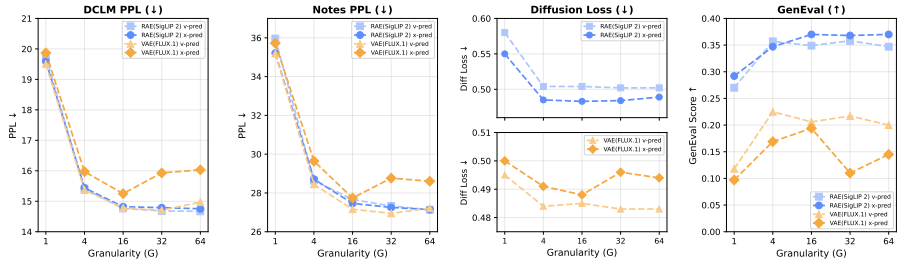
Only how **finely** the router can allocate capacity changes.

Beyond Language Modeling: An Exploration of Multimodal Pretraining, Tong et al.

MoE: how fine-grained should experts be?



MoE: how fine-grained should experts be?



Fine-grained experts improve both vision and language

Beyond Language Modeling: An Exploration of Multimodal Pretraining, Tong et al.

MoE: what about total expert count?

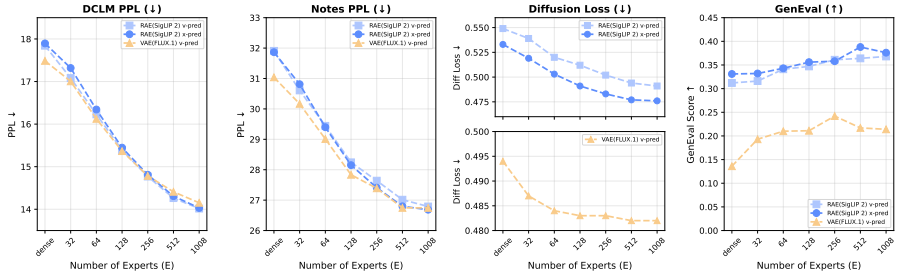
MoE decouples active compute from total capacity.

Experiment: fix active experts at 16, grow total pool:

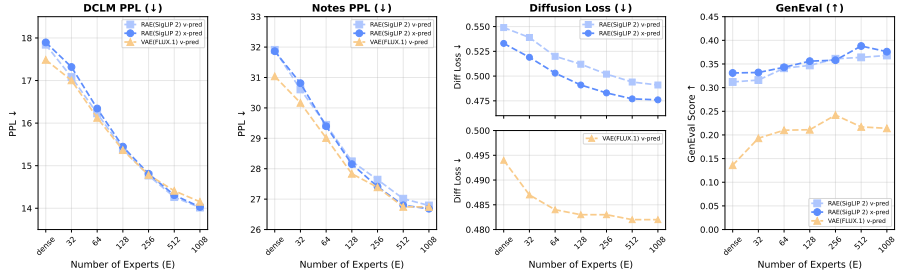
$32 \rightarrow 64 \rightarrow 128 \rightarrow 256 \rightarrow 1008$ experts

Active ratio drops from 50% to 1.6% — same inference cost.

MoE: what about total expert count?



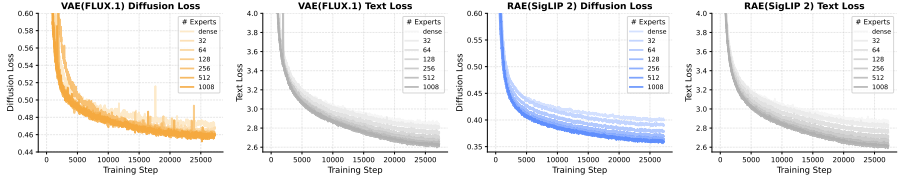
MoE: what about total expert count?



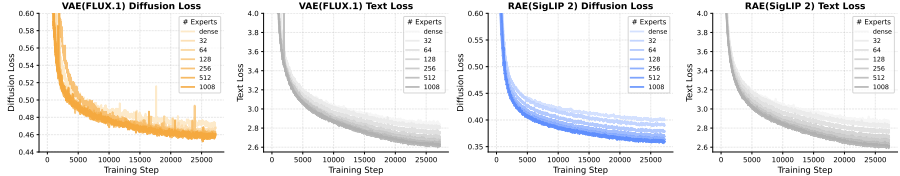
Same inference cost — more total experts consistently improves both modalities with RAE

Beyond Language Modeling: An Exploration of Multimodal Pretraining, Tong et al.

Why does the encoder choice matter for MoE?



Why does the encoder choice matter for MoE?



RAE keeps improving with more experts — VAE saturates

Rich representations can leverage the extra capacity that MoE provides

Beyond Language Modeling: An Exploration of Multimodal Pretraining, Tong et al.

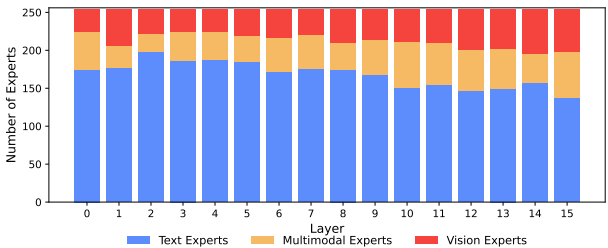
Finer experts, more sparsity —

MoE consistently improves multimodal training

at no extra inference cost

Especially with RAE

What does MoE learn inside?

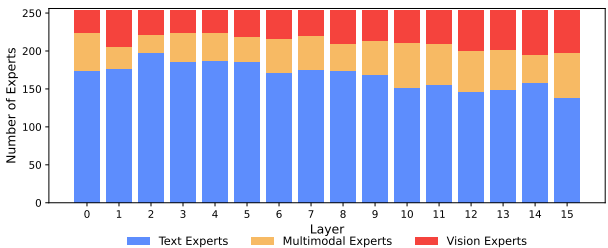


Each bar = one layer.

Colors = expert specialization.

Beyond Language Modeling: An Exploration of Multimodal Pretraining, Tong et al.

What does MoE learn inside?



Early layers are text-heavy

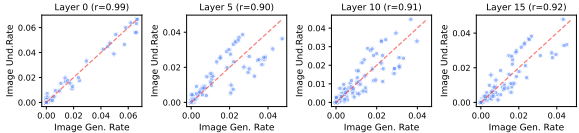
Deep layers grow vision + multimodal experts

Separate first, then integrate

Fully learned from data.

Beyond Language Modeling: An Exploration of Multimodal Pretraining, Tong et al.

Do generation and understanding use different experts?



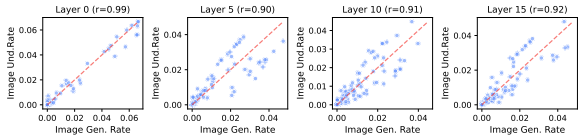
Each dot = one expert.

x-axis: activation during captioning.

y-axis: activation during denoising.

Beyond Language Modeling: An Exploration of Multimodal Pretraining, Tong et al.

Do generation and understanding use different experts?



$r \geq 0.90$

The **same experts** fire for both captioning and denoising.

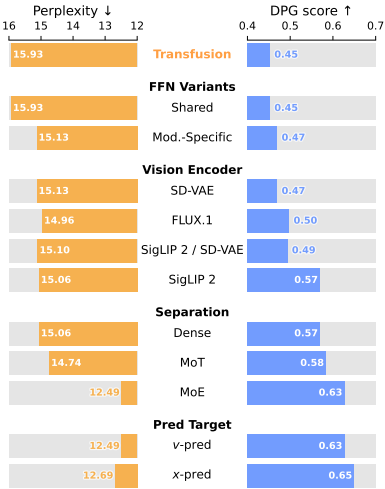
The model converges to a unified visual representation — without any constraint.

Beyond Language Modeling: An Exploration of Multimodal Pretraining, Tong et al.

Without any constraint,
the model converges to a unified visual representation
for both understanding and generation.

No constraint was imposed — this structure was fully learned from data.

Stacking design choices: step 1



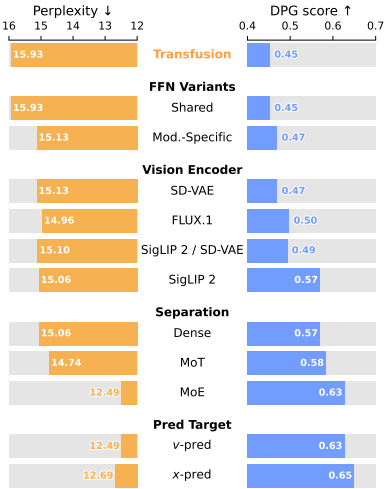
FFN structure

Shared FFN →
Modality-specific FFN

Both PPL and DPG improve

Even the simplest separation helps
— and costs nothing at inference.

Stacking design choices: step 2



1. Modality-specific FFN

Vision encoder

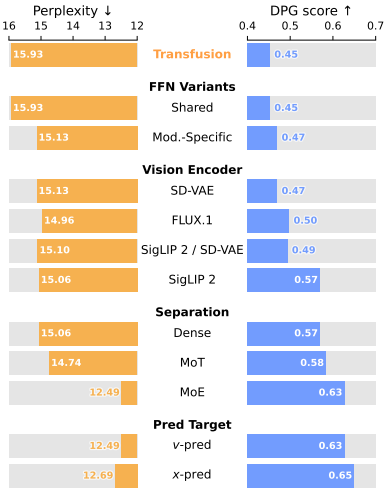
VAE → RAE (SigLIP 2)

Largest single improvement

Semantic encoder leaps ahead of VAEs and dual-encoder designs.

Beyond Language Modeling: An Exploration of Multimodal Pretraining, Tong et al.

Stacking design choices: step 3



1. Modality-specific FFN
2. RAE (SigLIP 2)

Capacity separation

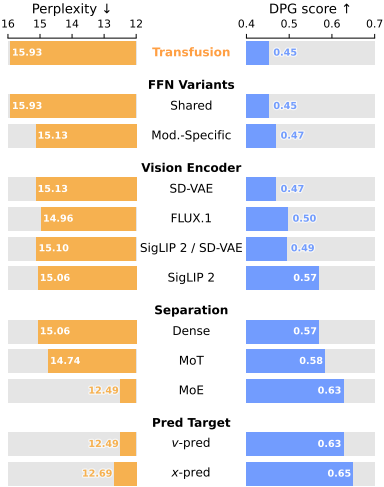
Dense → MoE

Another jump on both axes

MoE outperforms both dense and MoT at the same compute.

Beyond Language Modeling: An Exploration of Multimodal Pretraining, Tong et al.

Stacking design choices: step 4



1. Modality-specific FFN
2. RAE (SigLIP 2)
3. MoE

Prediction target

v-pred $\rightarrow$ x-pred

Generation improves further

These choices stack —

improvements are complementary.

Beyond Language Modeling: An Exploration of Multimodal Pretraining, Tong et al.

MoE gives each modality its own capacity.

But does this actually scale?

What does compute-optimal multimodal training look like?

Beyond Language Modeling: An Exploration of Multimodal Pretraining, Tong et al.

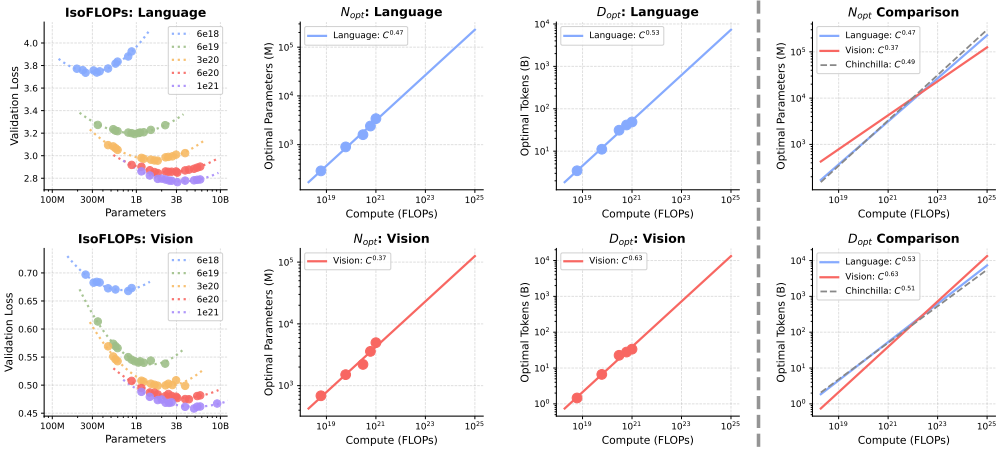
Chinchilla, but for multimodal

IsoFLOP: sweep model size and tokens at fixed compute.

Same methodology as Chinchilla, applied to joint vision + language.

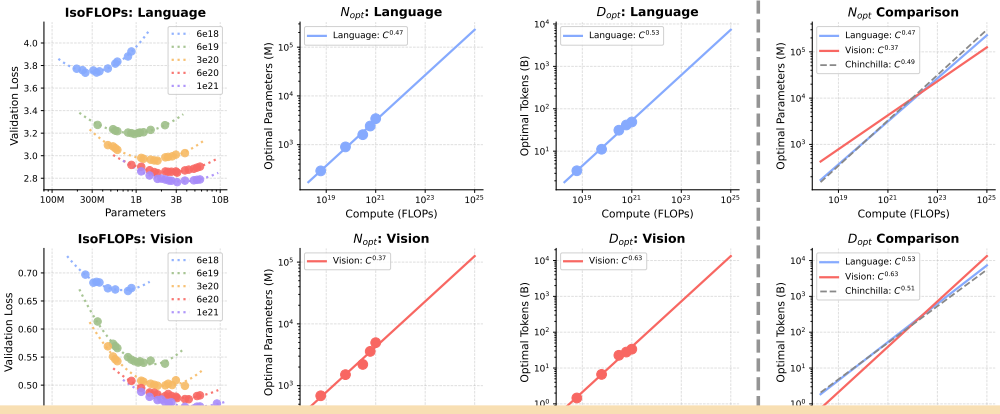
What does compute-optimal look like with two modalities?

Dense scaling laws



Beyond Language Modeling: An Exploration of Multimodal Pretraining, Tong et al.

Dense scaling laws



Vision (b=0.63) is significantly more data-hungry than language (b=0.53)

Beyond Language Modeling: An Exploration of Multimodal Pretraining, Tong et al.

A fundamental asymmetry

Language

$D_{\text{opt}} \propto C^{0.53}$

Parameter-hungry

Vision

$D_{\text{opt}} \propto C^{0.63}$

Data-hungry

At **1T parameters**, the gap compounds:

Extrapolating to 1T parameters, vision would need **51**$\times$ more data than language.

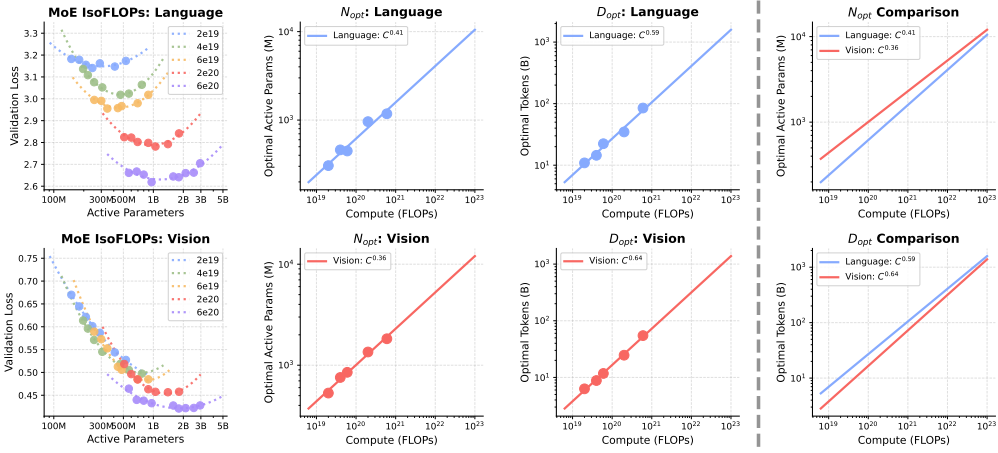
No single dense configuration can satisfy both.

Beyond Language Modeling: An Exploration of Multimodal Pretraining, Tong et al.

Dense models face a fundamental tradeoff between modalities.

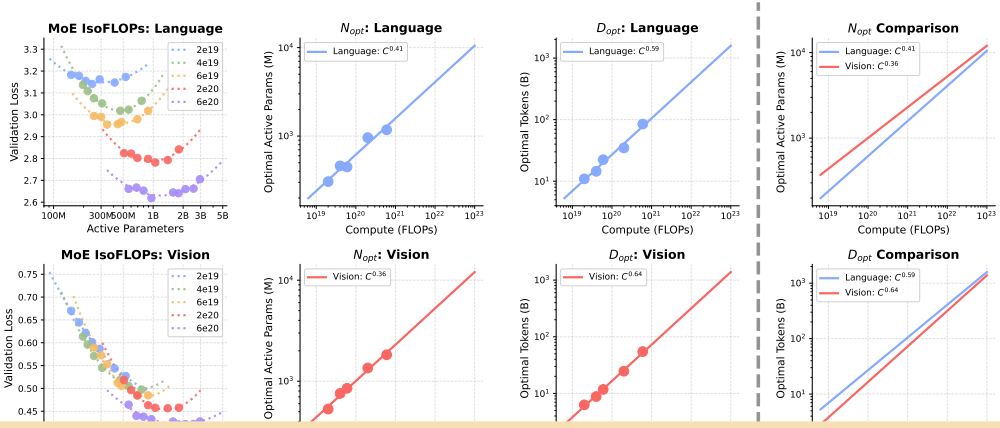
Can MoE resolve this?

MoE scaling laws



Beyond Language Modeling: An Exploration of Multimodal Pretraining, Tong et al.

MoE scaling laws



Exponent gap halved: 0.10 (dense) → 0.05 (MoE)

In summary

1. A single representation has high potential to serve both understanding and generation.

1. A single representation has high potential to serve both understanding and generation.

2. Modalities coexist with minimal cost and can be synergistic.

1. A single representation has high potential to serve both understanding and generation.

2. Modalities coexist with minimal cost and can be synergistic.

3. Action-conditioned world modeling can emerge from diverse pretraining, especially pure video.

1. A single representation has high potential to serve both understanding and generation.

2. Modalities coexist with minimal cost and can be synergistic.

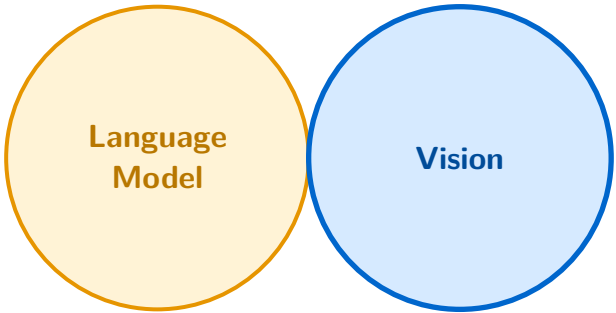
3. Action-conditioned world modeling can emerge from diverse pretraining, especially pure video.

4. Architecture matters — MoE narrows the scaling asymmetry and naturally learns modality separation and unification.

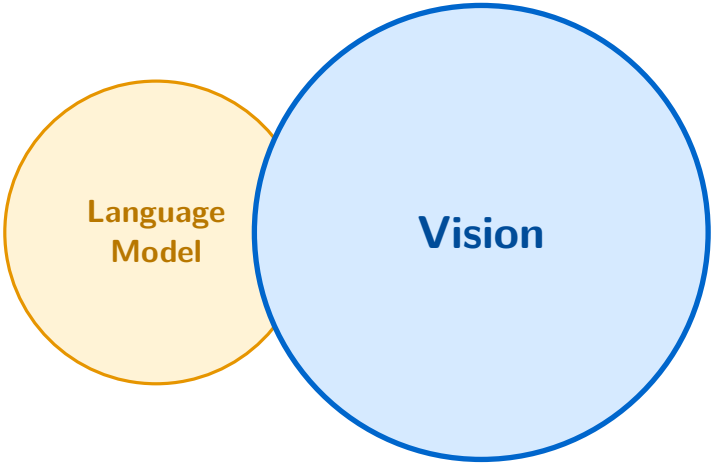
The future

The next generation of foundation models won't just process language.

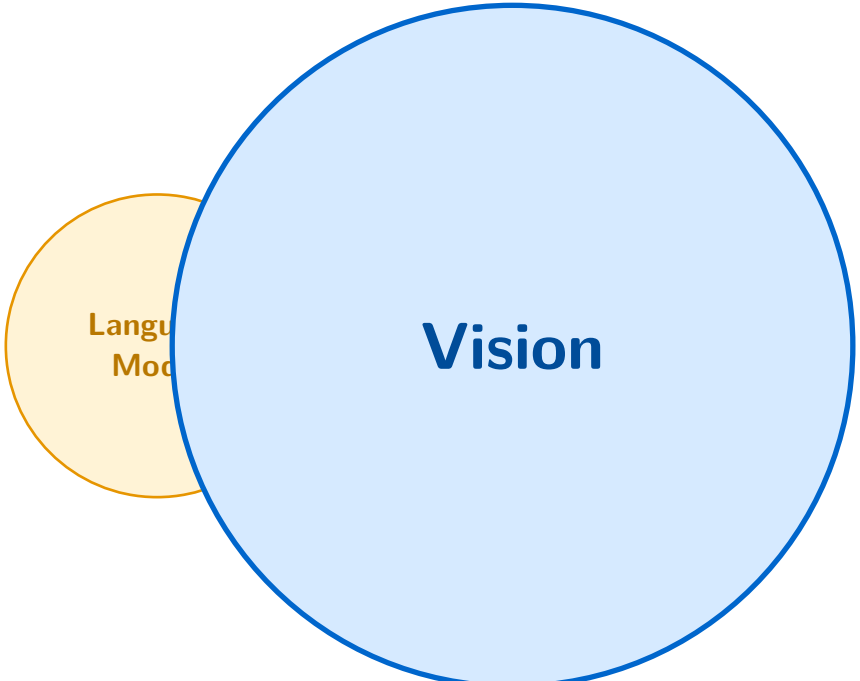
They'll understand and simulate the world.



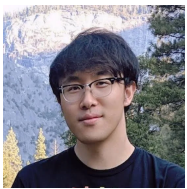
The future



The future


Langu
Mod
Vision

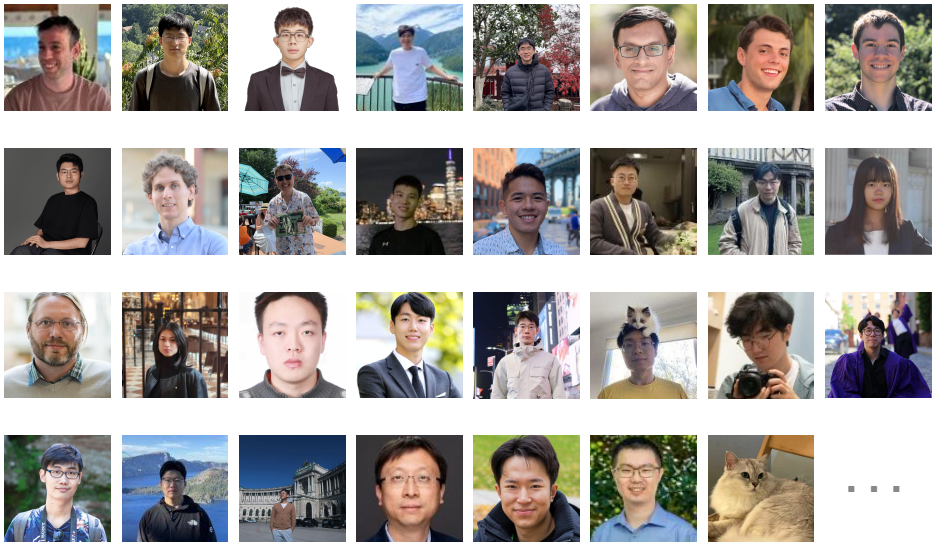
Advisors



Saining Xie

Yann LeCun

Mentors & Collaborators



Thank you

Shengbang Tong
New York University