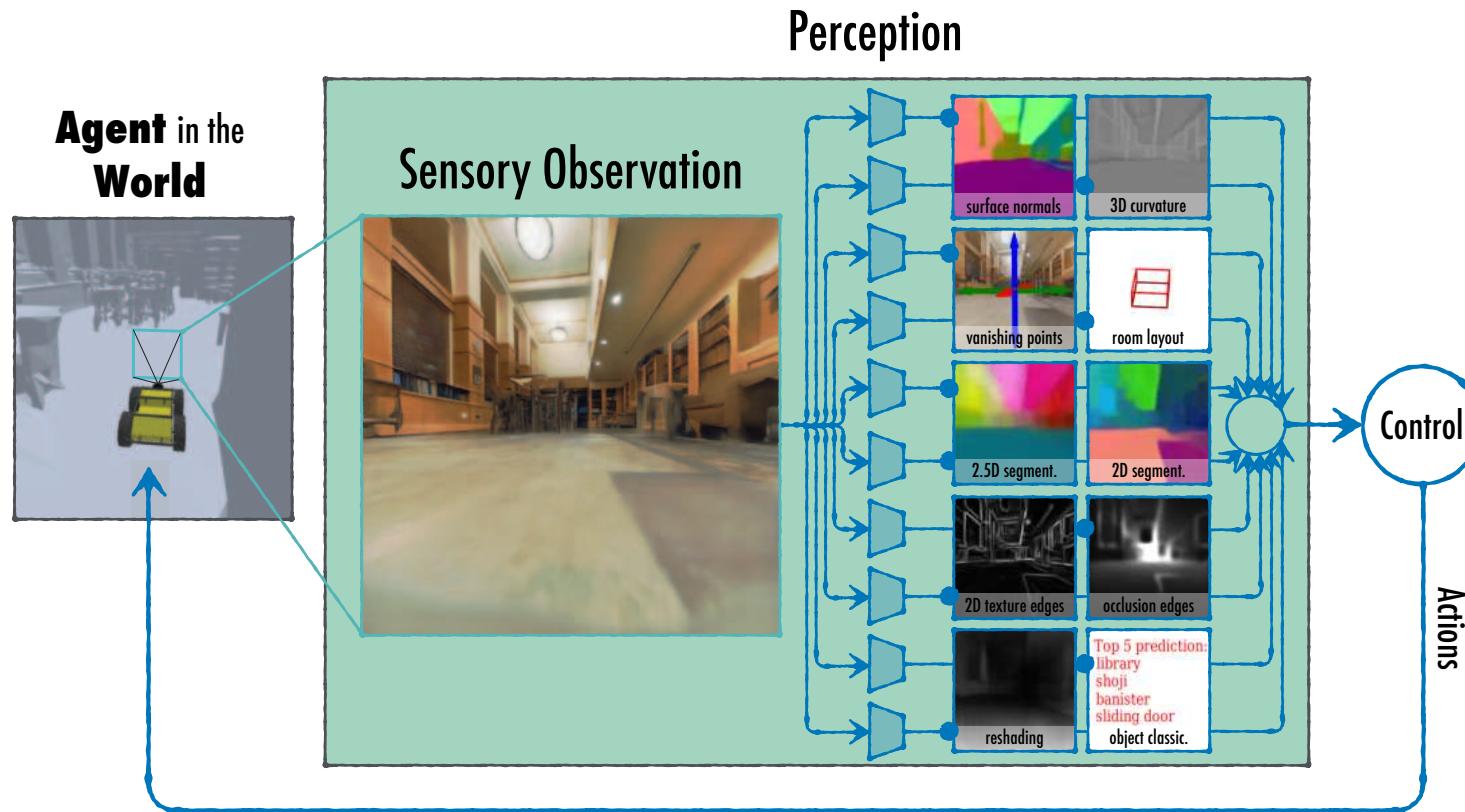


Logistics

Week Num.	Date	Item
1	20.02	- lecture 1
2a	25.02	- lecture 2
2b	27.02	- lecture 3
3a	04.03	- lecture 4
3b	06.03	- lecture 5
4a	11.03	- lecture 6 (+ Q&A)
	11.03	- Transformers notebook assignment due
4b	13.03	- lecture 7
5a	18.03	- lecture 8
5b	20.03	- lecture 9
6a	25.03	- lecture 10
6b	27.03	- lecture 11 (+ Q&A)
	01.04	- Active agents notebook assignment due
7a	01.04	- lecture 12
7b	03.04	- lecture 13
8a	08.04	- lecture 14
8b	10.04	- lecture 15 (+ Matchmaking session)
	13.04	- Project proposals due
	15.04	- all subsequent sessions from 15.04 onwards are for Q&A
	18.04	- Project proposals due, when revision is needed.
	22.04	- MidSem break - No classes
	25.04	- MidSem break - No classes
	29.04	- Foundation Models assignment due
	01.05	- lecture 16
	09.05	- Project progress report due
	13.05	- Robustness assignment due (extra credit)
	20.05	- Moodle homework due
	26.05	- Final project presentation video due
	27.05	- Final project presentation Part I
	29.05	- Final project presentation Part II
	30.05	- Project report due

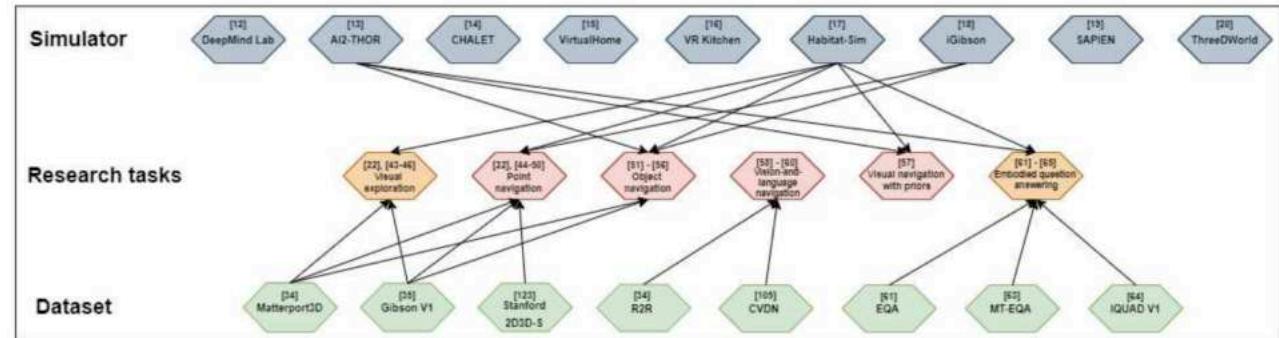
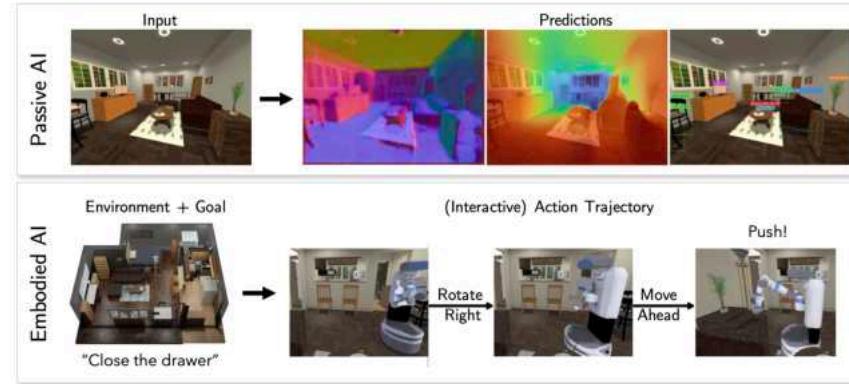
Recap

Vision In-the-loop

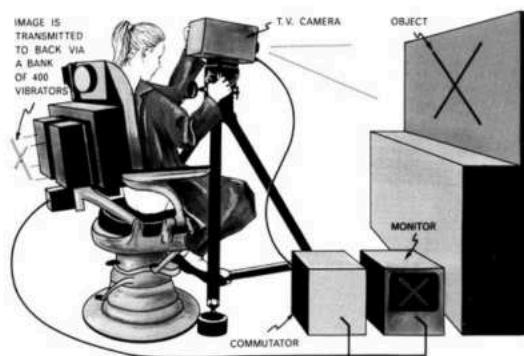


Common Tasks (~so far)

- (1) visual navigation
- (2) rearrangement
- (3) embodied vision-and-language

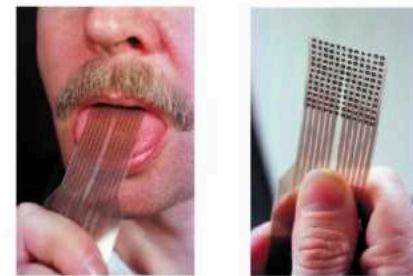


Sensorimotor Contingency



Bach-y-Rita et al., Vision substitution by tactile image projection, *Nature* (1969)

Tongue Display Unit



Sampaio, E., S. Maris., and P. Bach-y-Rita. 2001. Brain plasticity: 'Visual' acuity of blind persons via the tongue. *Brain Research* 908(July 13):204.



David Ha 2022. Eriemann & Kohler 1931. Stratton 1897.
Paul Bach-y-rita (1934-2006) (the father of sensory substitution).

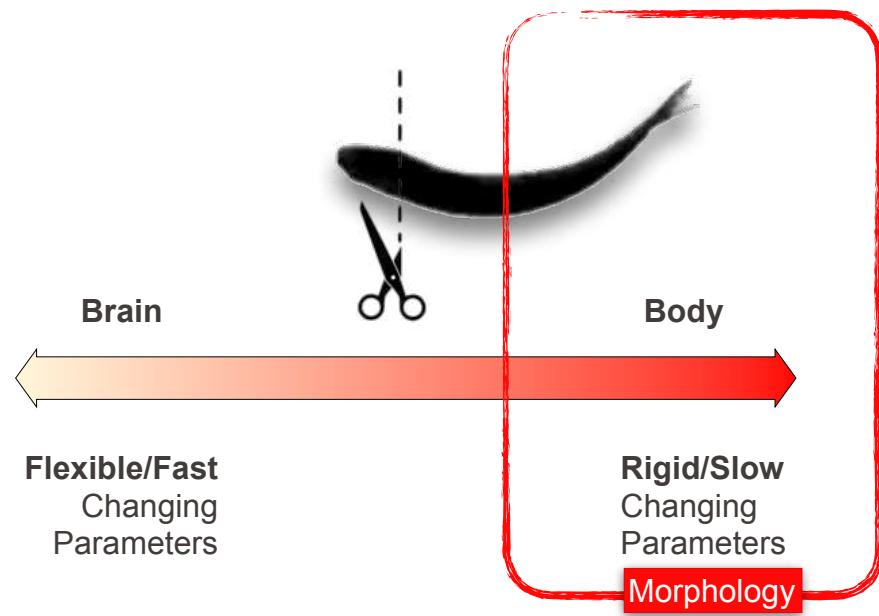
Fish Swimming



Dead Fish Swimming



- Is the fish intelligent?
- Where is the intelligence?



Automatic Design of Morphologies

Genotype: directed graph. Phenotype: hierarchy of 3D parts.

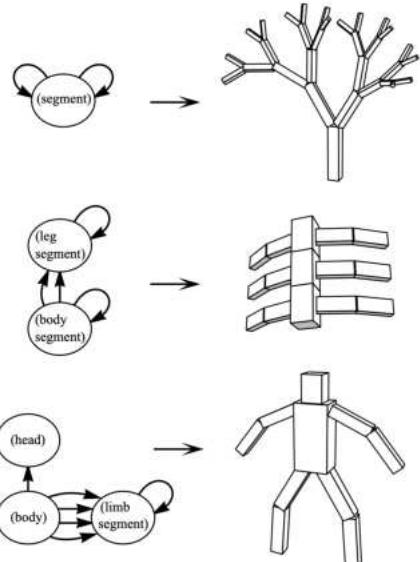
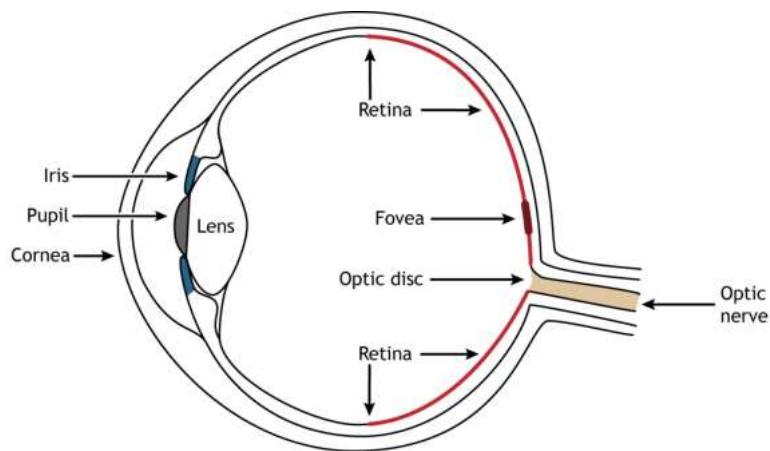


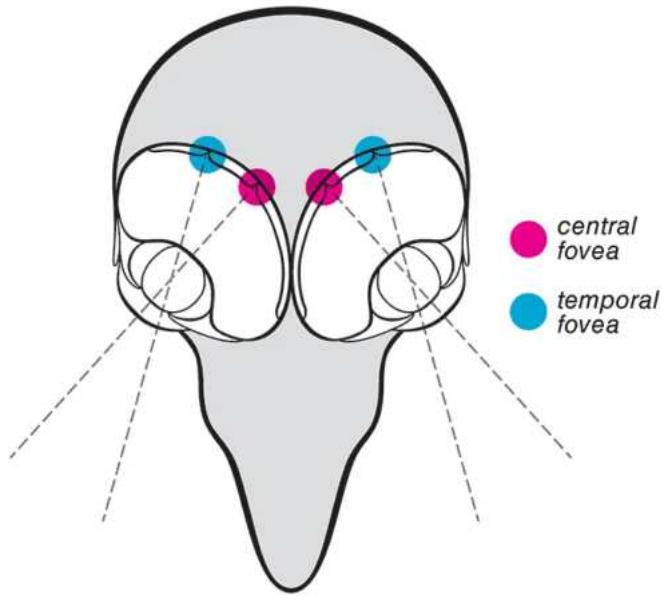
Figure 1: Designed examples of genotype graphs and corresponding creature morphologies.



- Evolving Virtual Creatures, Sims 1994

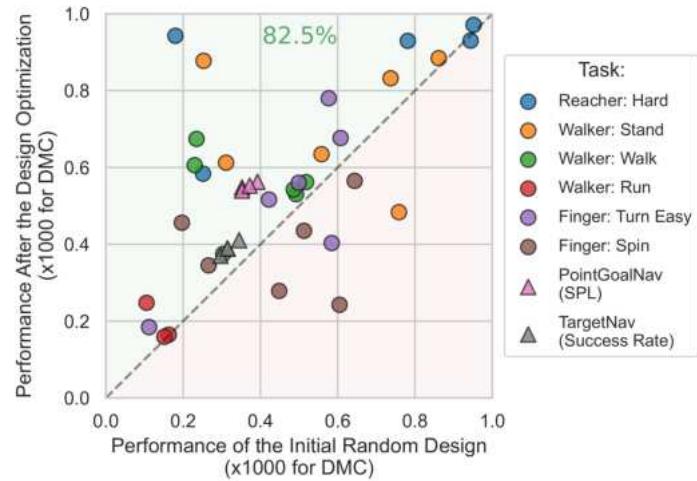
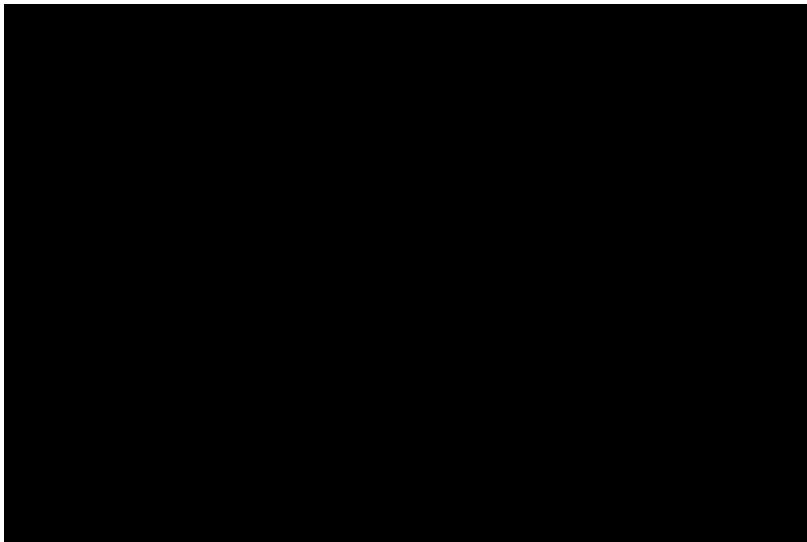
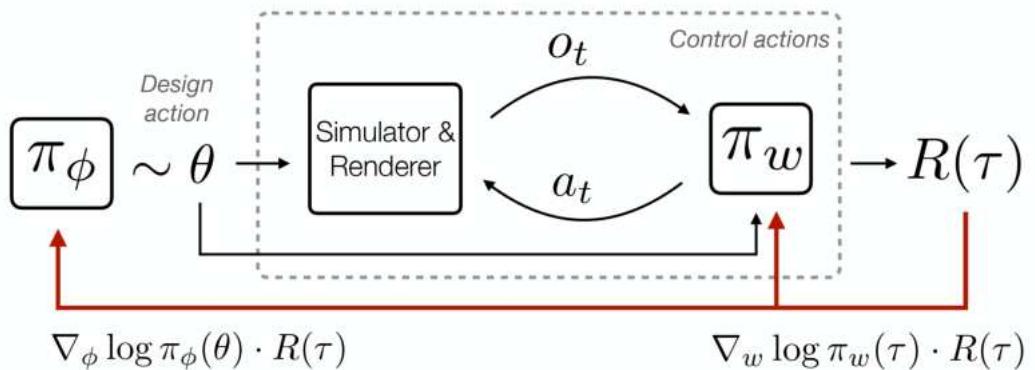


Human

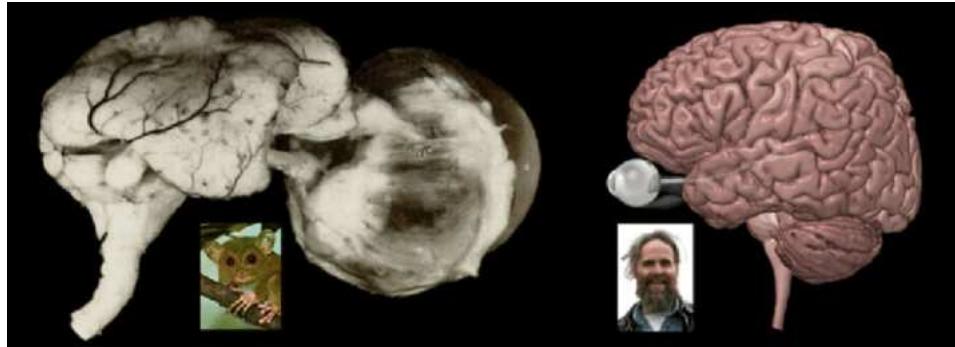


Hawks, falcons, etc.

Designing the morphology



Multimodality



Human



Cat



Rat



Fly



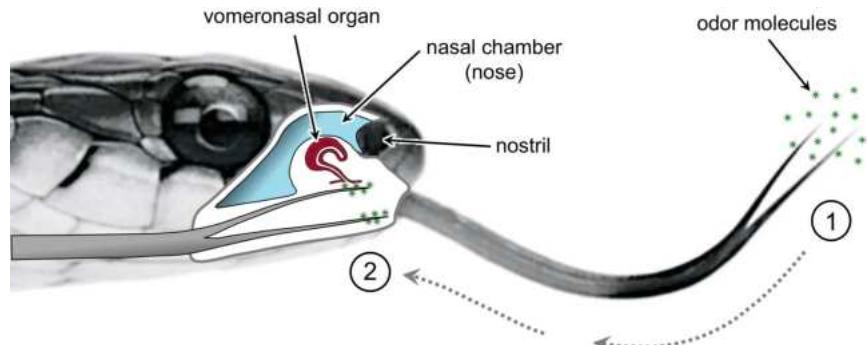
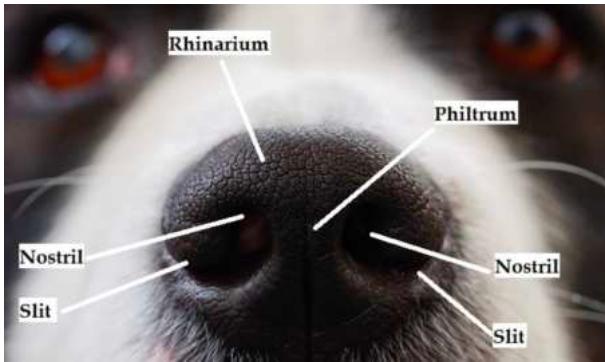
Mosquito



- <https://visual-morphology.epfl.ch/>



Other modalities: olfactory

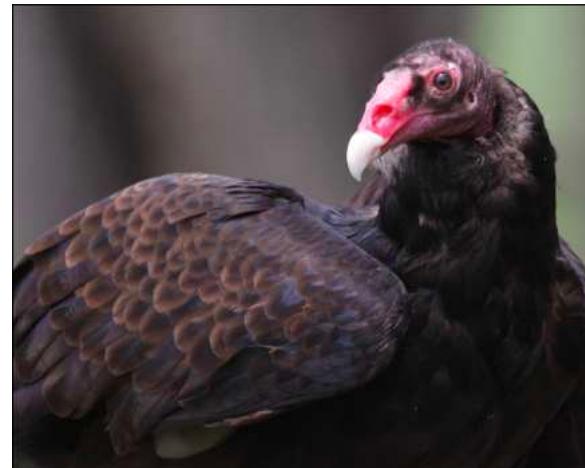


Other modalities: olfactory



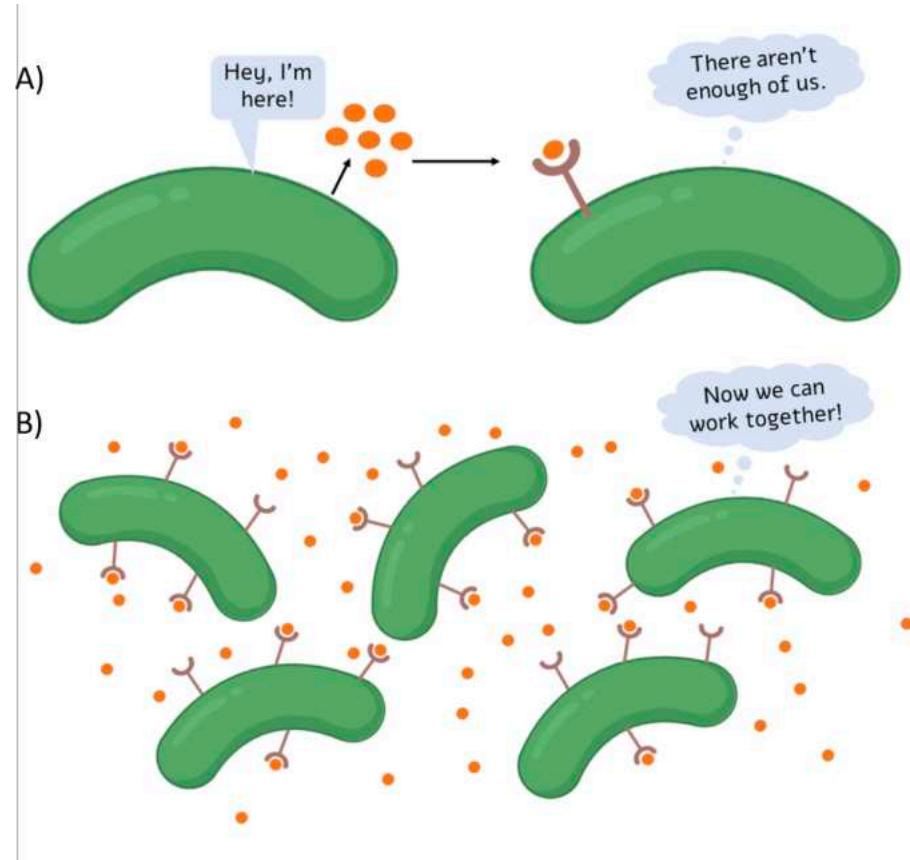
Other modalities: olfactory

- Almost never unimodal/vision-only.
- “In the 1930s, Union Oil Company discovered that adding ethyl mercaptan, a strong-smelling chemical that smells like rotting cabbage or eggs, to natural gas pipelines **could help them find leaks by monitoring vulture activity.**”



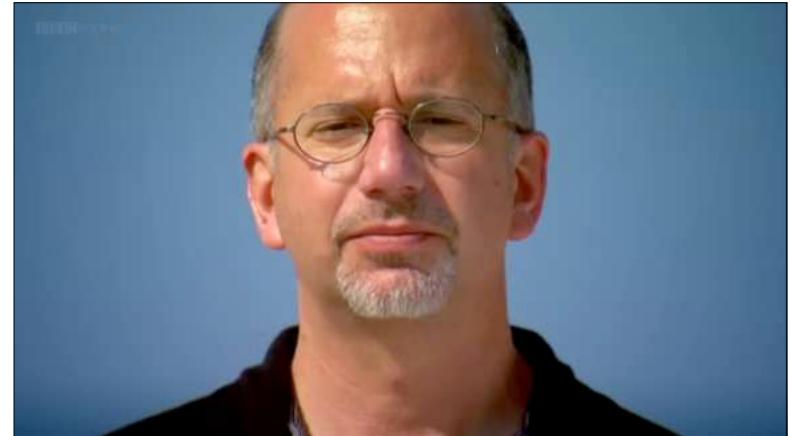
Other modalities: olfactory

- Chemical sensing





Vision: “va”
Audio: “ba”



Vision: “ba”
Audio: “ba”

-



-

"McGurk Effect". Hearing lips and seeing voices, McGurk & MacDonald, Nature 1976 .

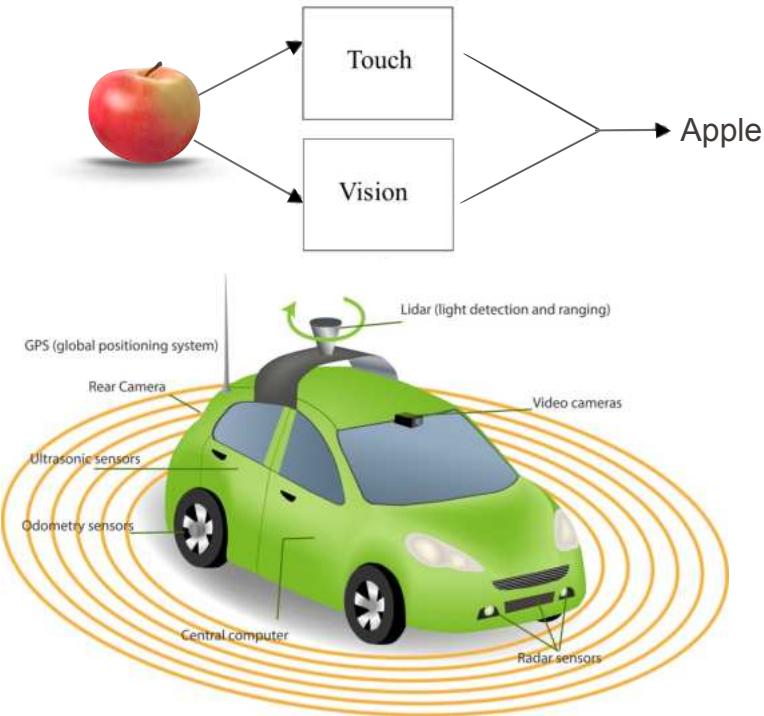
- Ian Waterman
 - feedback
 - multimodality



Swain, K. (2017). The phenomenology of touch. *The Lancet Neurology*.

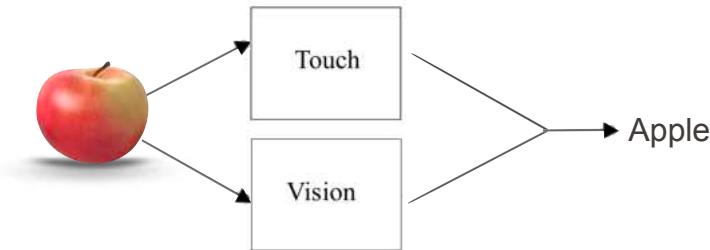
Roles of Multimodality in Learning

For sensory fusion / better inference



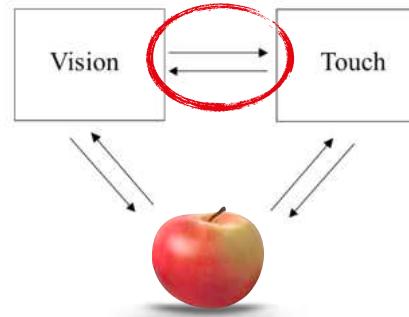
Roles of Multimodality in Learning

For sensory fusion / better inference

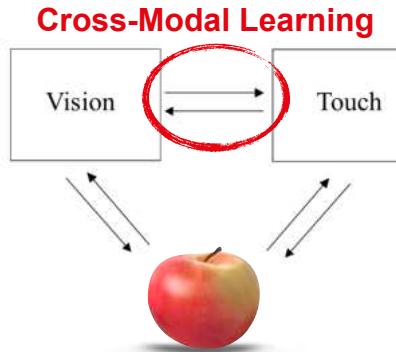


For self-supervision

Cross-Modal Learning

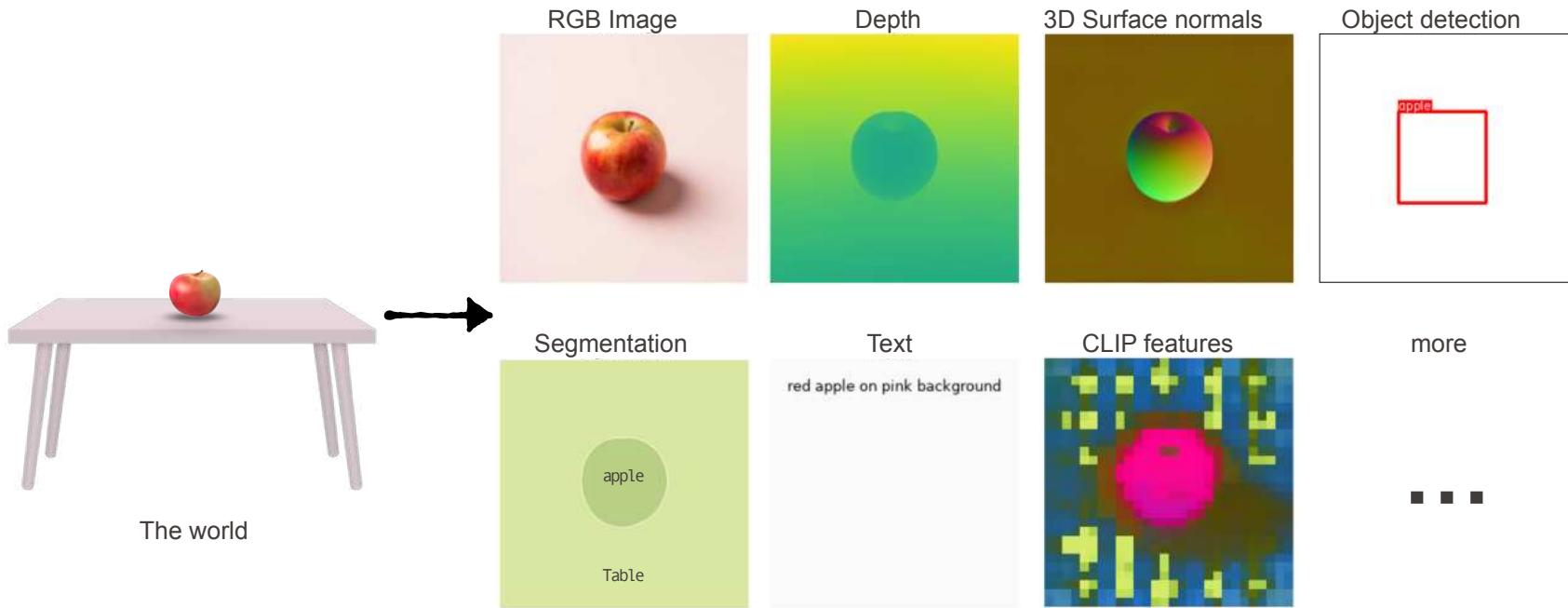


- “Six Lessons from Babies”, Smith&Gasser’05:
 - *Babies’ experience of the world is profoundly **multimodal**.*
 - *Infants spend hours watching their own actions.*
 - *Multiple overlapping and time-locked sensory systems enable the developing system to **educate (“supervise”)** itself.*

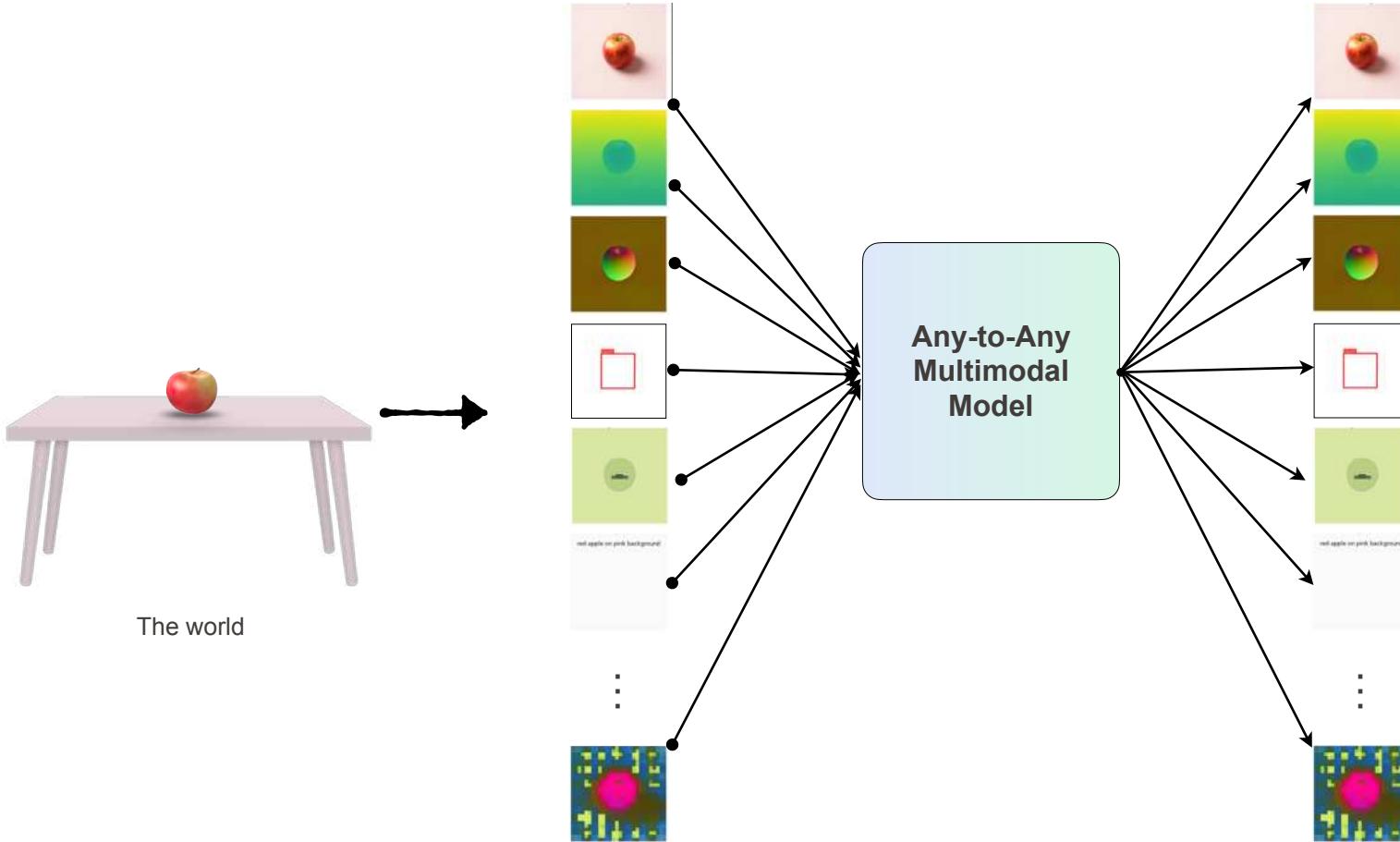


Smith, L., & Gasser, M. The development of embodied cognition: Six lessons from babies. Artificial life 2005.

Core function: Predict anything from anything



Core function: Predict anything from anything



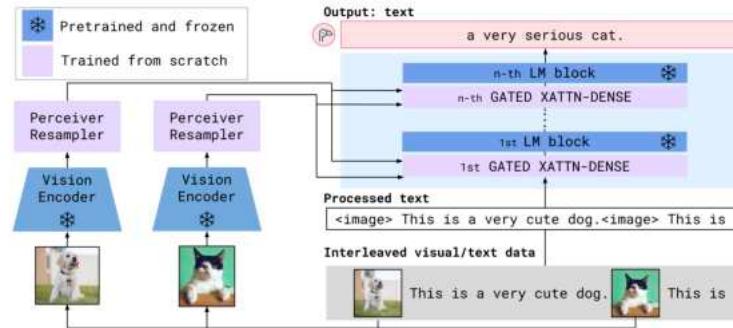
VLM (Vision-Language Model)

~ (RGB-Text chatbot)

Input Prompt					Completion
	This is a chinchilla. They are mainly found in Chile.		This is a shiba. They are very popular in Japan.		This is a flamingo. They are found in the Caribbean and South America.
	What is the title of this painting? Answer: The Hallucinogenic Toreador.		Where is this painting displayed? Answer: Louvres Museum, Paris.		What is the name of the city where this was painted? Answer:
	Output: "Underground"		Output: "Congress"		Output: "Soulomes"
	2+1=3		5+6=11		3x6=18
	Output: A propaganda poster depicting a cat dressed as French emperor Napoleon holding a piece of cheese.		Output: A pink room with a flamingo pool float.		Output: A portrait of Salvador Dali with a robot head.
	Les sanglots longs des violons de l'automne blescent mon cœur d'une langueur monotone.		Pour qui sont ces serpents qui sifflent sur vos têtes?		Je suis un cœur qui bat pour vous.
	pandas: 3		dogs: 2		giraffes: 4
	I like reading		, my favourite play is Hamlet. I also like		, my favorite book is
					What happens to the man after hitting the ball? Answer: he falls down.

VLM (Vision-Language Model)

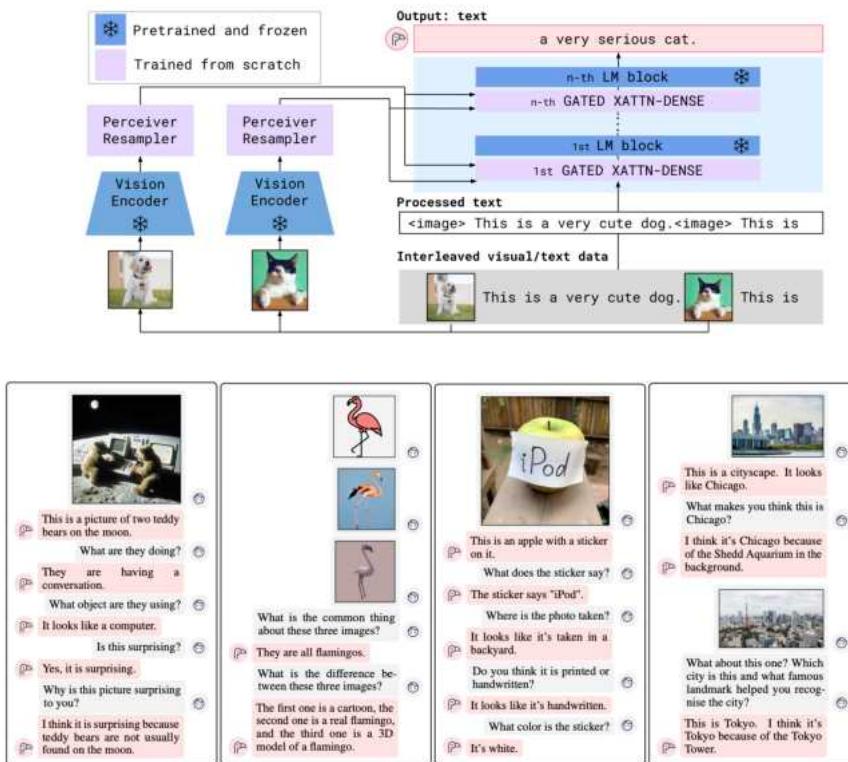
~ (RGB-Text chatbot)



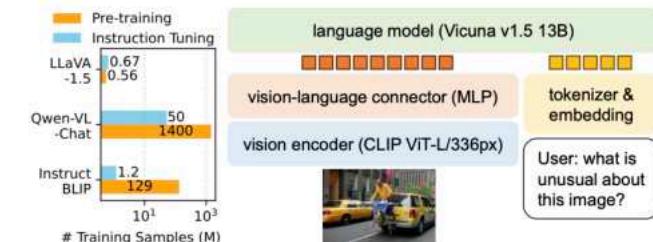
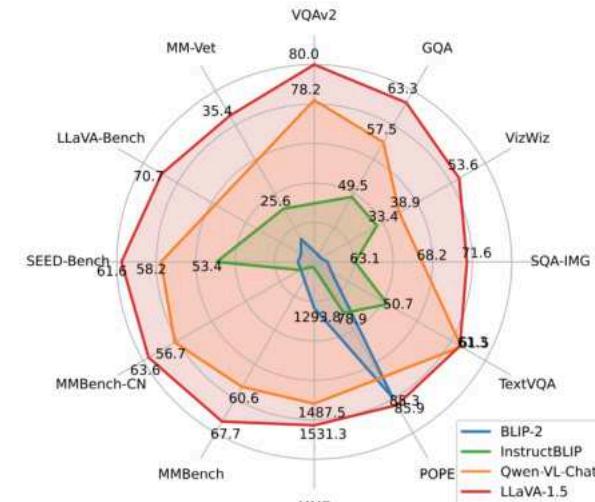
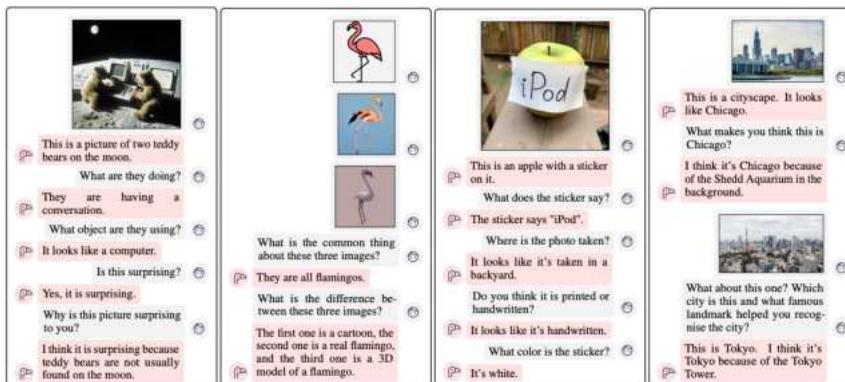
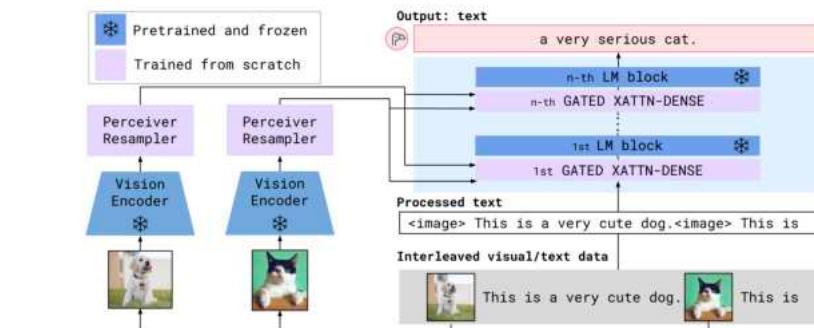
Input Prompt				Completion
This is a chinchilla. They are mainly found in Chile.	This is a shiba. They are very popular in Japan.	This is		a flamingo. They are found in the Caribbean and South America.
What is the title of this painting? Answer: The	Hallucinogenic Toreador.	Where is this painting displayed? Answer: Louvres Museum, Paris.	What is the name of the city where this was painted? Answer:	→ Arles.
Output: "Underground"	Output: "Congress"	Output:		→ "Soulomes"
2+1=3	5+6	5+6=11	3×6	→ 3x6=18
Output: A propaganda poster depicting a cat dressed as French emperor Napoleon holding a piece of cheese.	Output: A pink room with a flamingo pool float.	Output:		→ A portrait of Salvador Dali with a robot head.
Les sanglots longs des violons de l'automne blesSENT mon cœur d'une langueur monotone.	Pour qui sont ces serpents qui sifflent sur vos têtes?	Pour qui sont ces serpents qui sifflent sur vos têtes?		→ Je suis un cœur qui bat pour vous.
pandas: 3	dogs: 2			→ giraffes: 4
I like reading	, my favourite play is Hamlet. I also like	, my favorite book is		→ Dreams from my Father.
What happens to the man after hitting the ball? Answer:				→ he falls down.

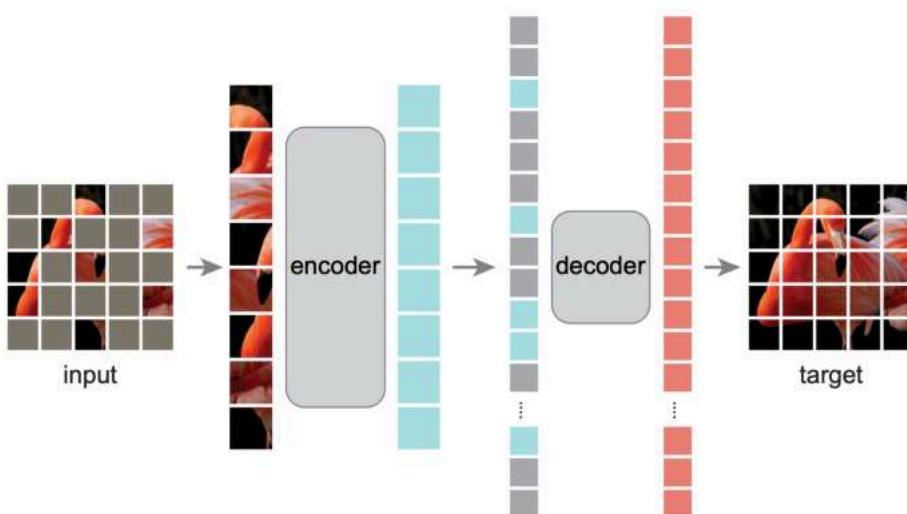
VLM (Vision-Language Model)

~ (RGB-Text chatbot)

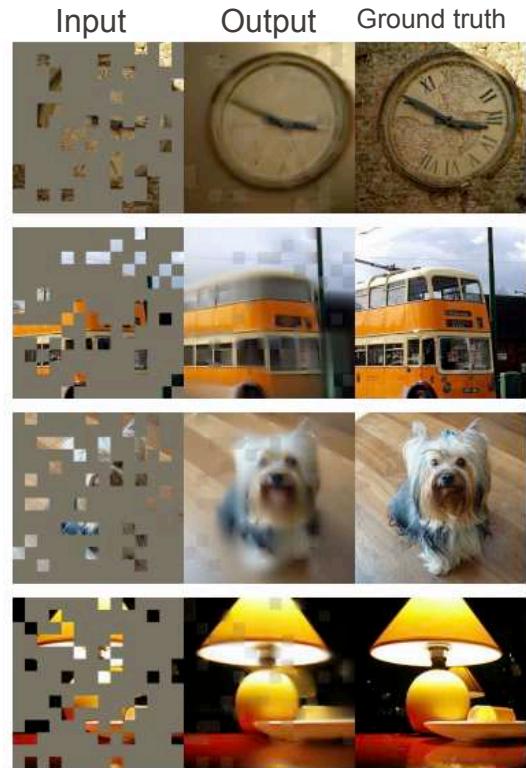


Input Prompt				Completion
	This is a chinchilla. They are mainly found in Chile.		This is a shiba. They are very popular in Japan.	a flamingo. They are found in the Caribbean and South America.
	What is the title of this painting? Answer: The Hallucinogenic Toreador.		Where is this painting displayed? Answer: Louvres Museum, Paris.	What is the name of the city where this was painted? Answer:
	Output: "Underground"		Output: "Congress"	Output: Arles.
	2+1=3		5+6=11	
	Output: A propaganda poster depicting a cat dressed as French emperor Napoleon holding a piece of cheese.		Output: A pink room with a flamingo pool float.	Output: "Soulomes"
	Les sanglots longs des violons de l'automne blesseront mon cœur d'une langueur monotone.		Pour qui sont ces serpents qui sifflent sur vos têtes?	A portrait of Salvador Dalí with a robot head.
	pandas: 3		dogs: 2	Output: giraffes: 4
	I like reading		, my favourite play is Hamlet. I also like	, my favorite book is
				What happens to the man after hitting the ball? Answer: he falls down.



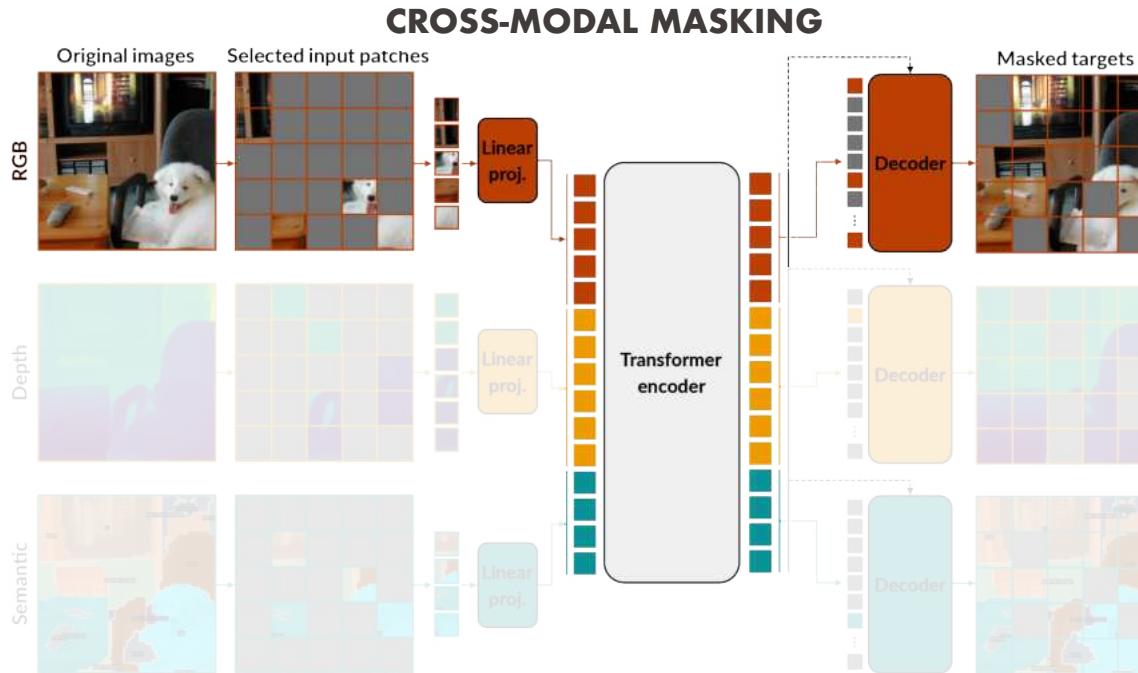


[Masked Autoencoders Are Scalable Vision Learners. He et al. 2021]



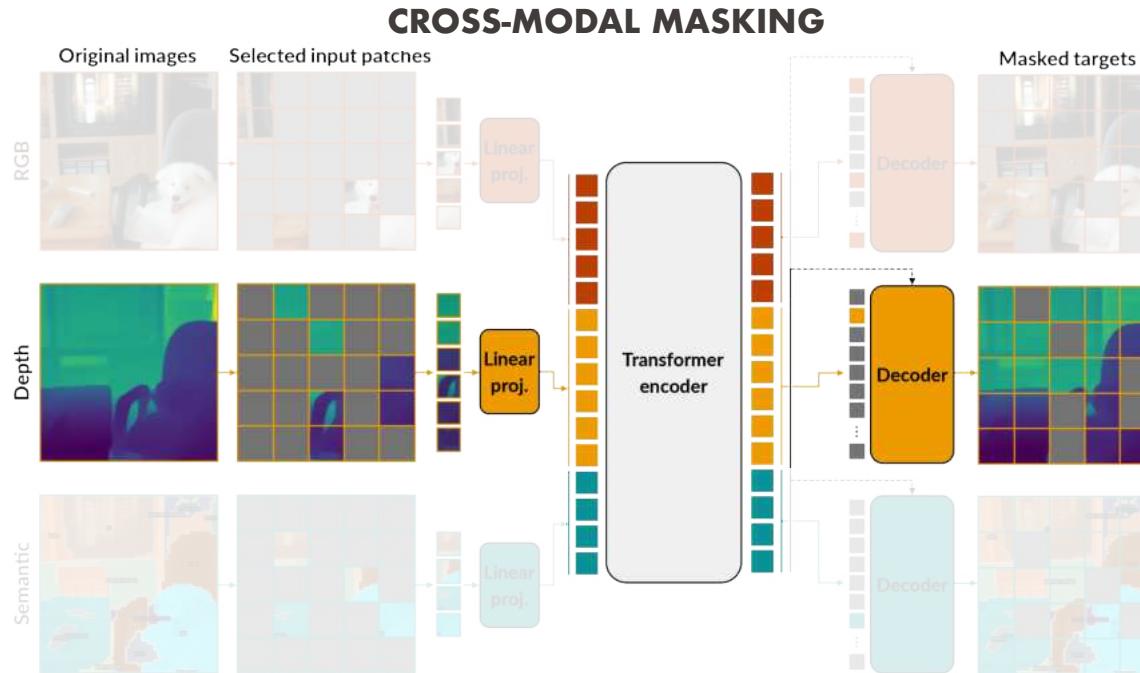
Cross-Modal Masked Modeling

MultiMAE: Multi-Modal Multi-Task Masked Autoencoders



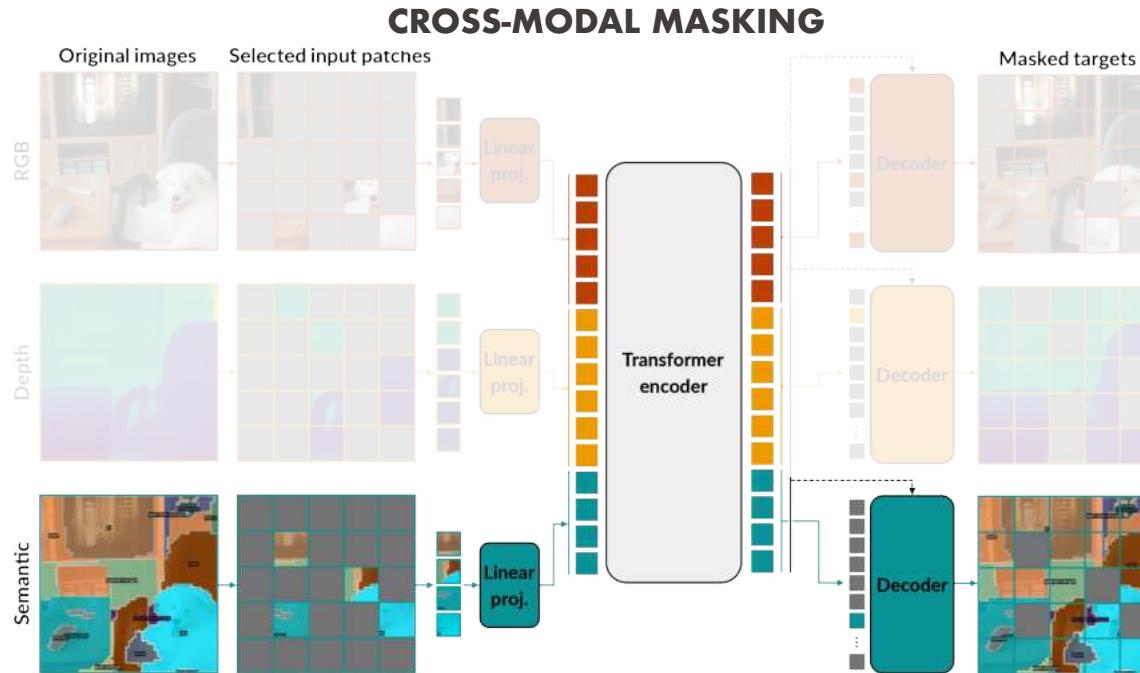
Cross-Modal Masked Modeling

MultiMAE: Multi-Modal Multi-Task Masked Autoencoders



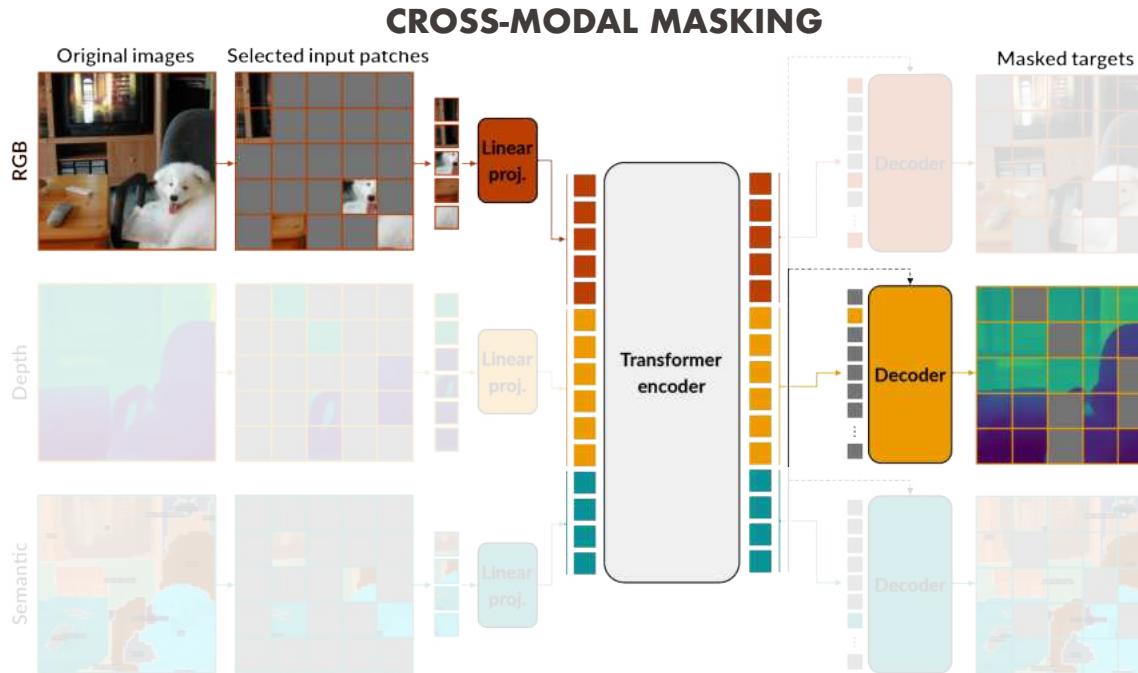
Cross-Modal Masked Modeling

MultiMAE: Multi-Modal Multi-Task Masked Autoencoders



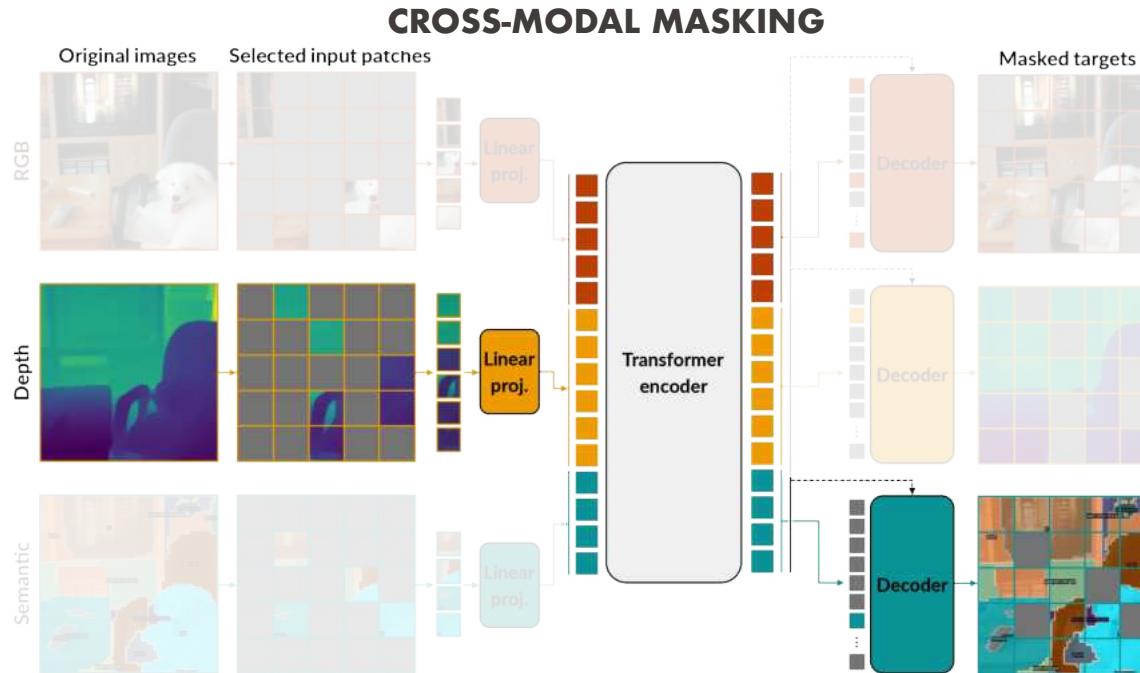
Cross-Modal Masked Modeling

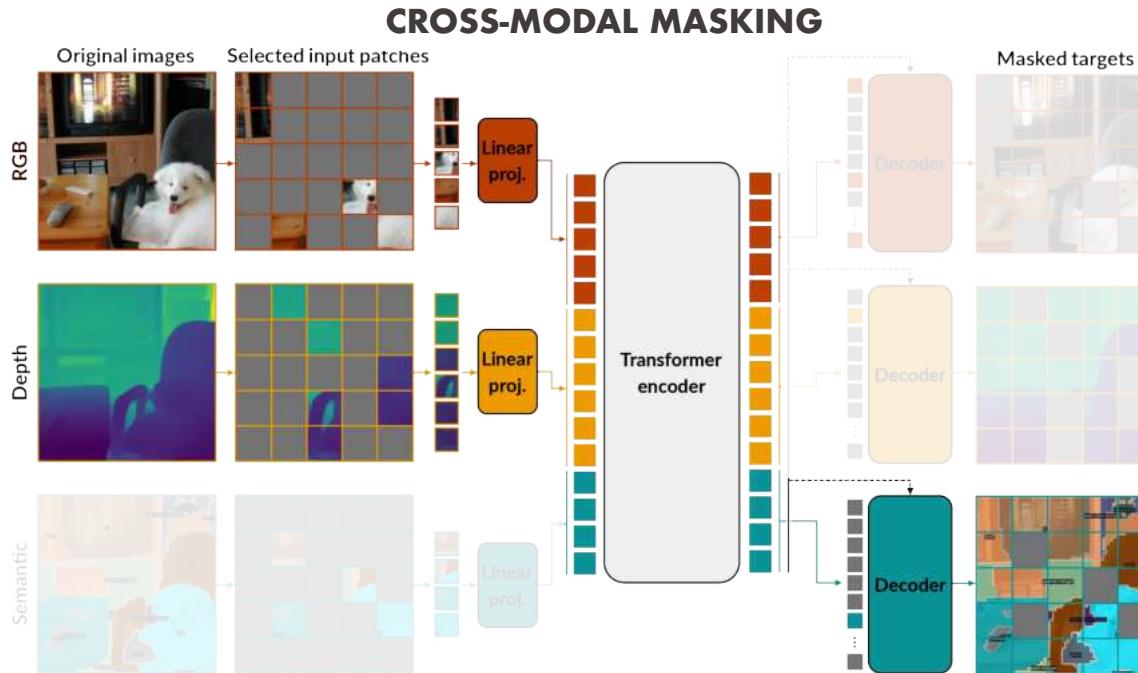
MultiMAE: Multi-Modal Multi-Task Masked Autoencoders

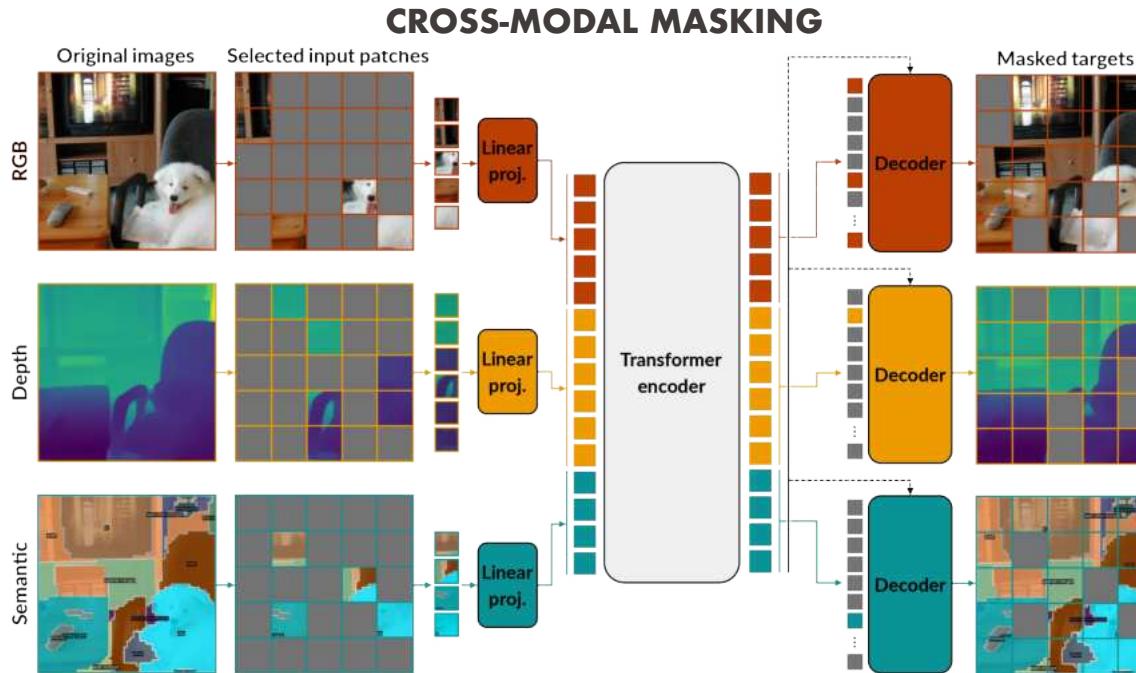


Cross-Modal Masked Modeling

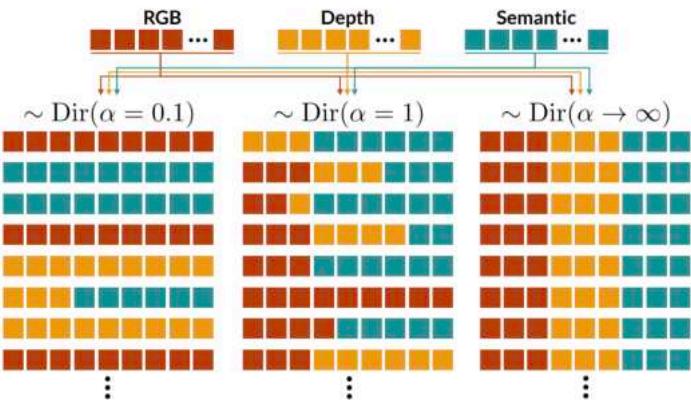
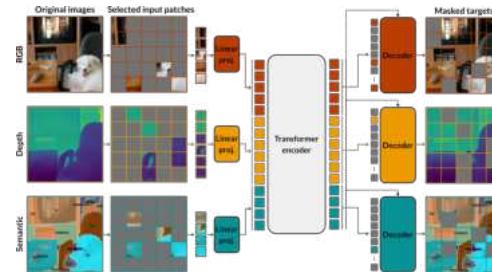
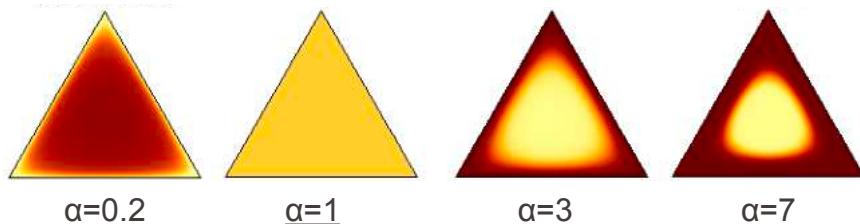
MultiMAE: Multi-Modal Multi-Task Masked Autoencoders



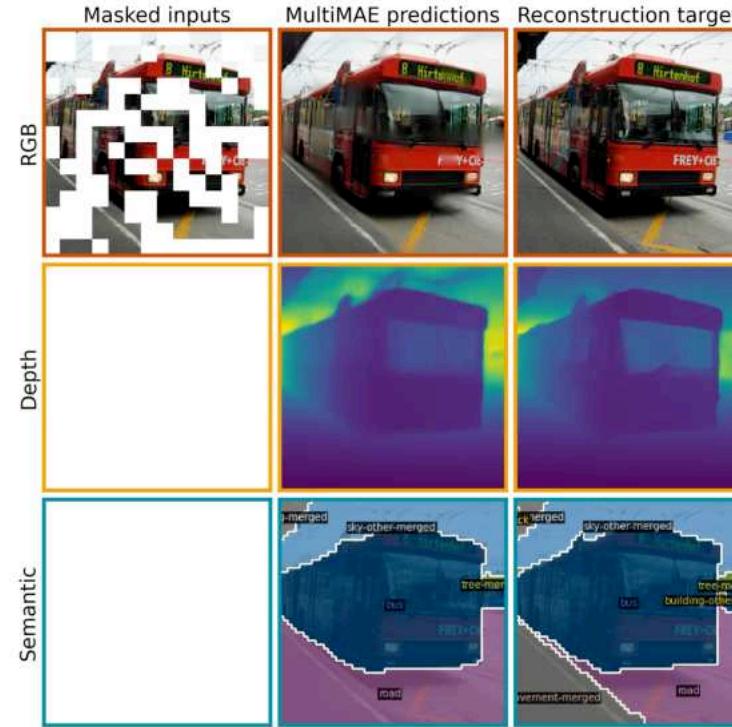
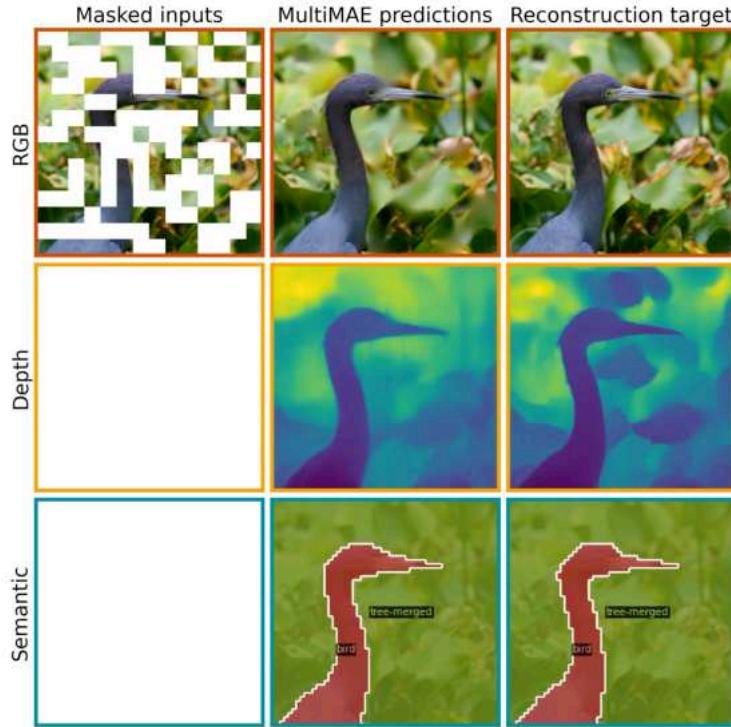


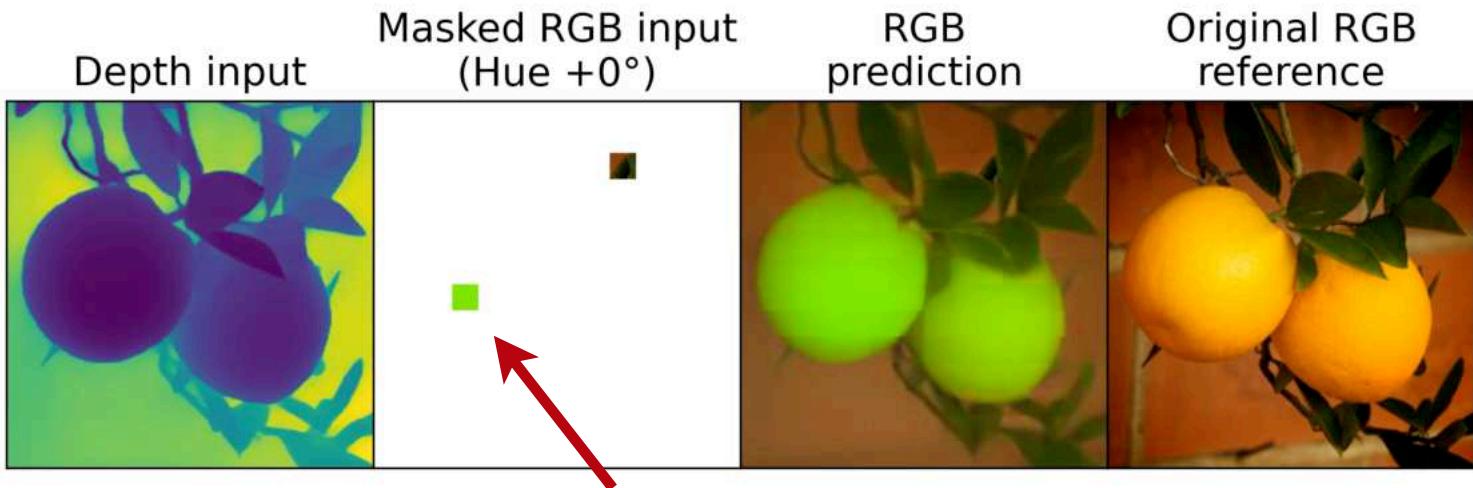


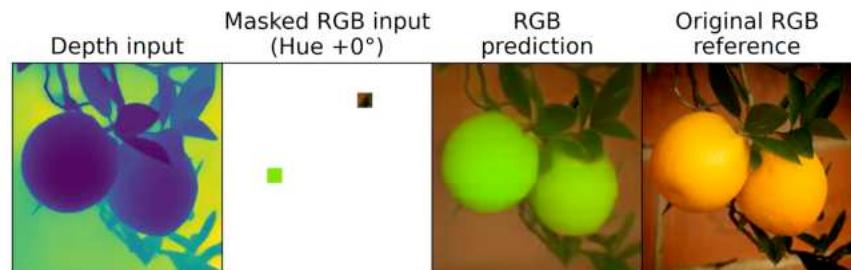
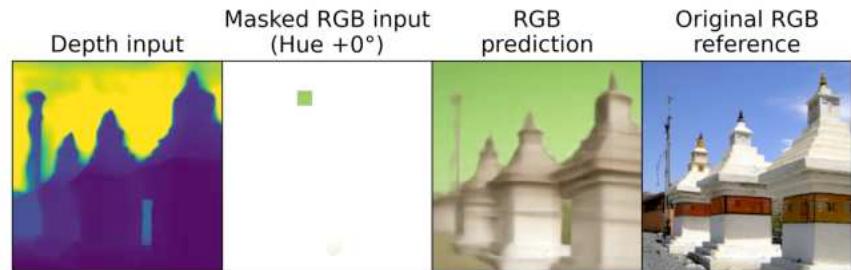
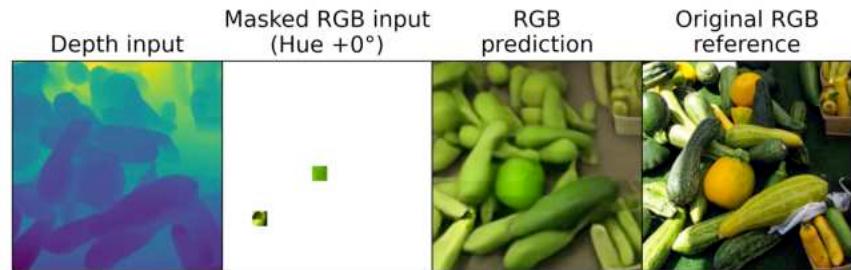
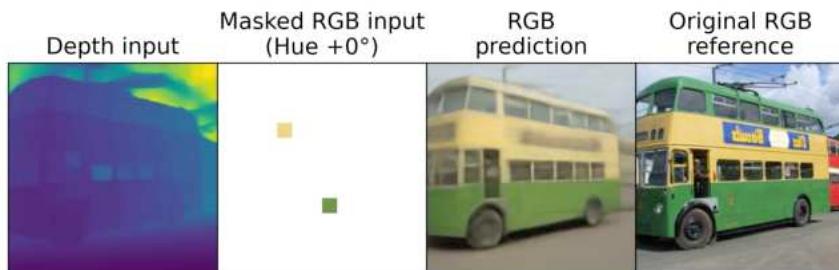
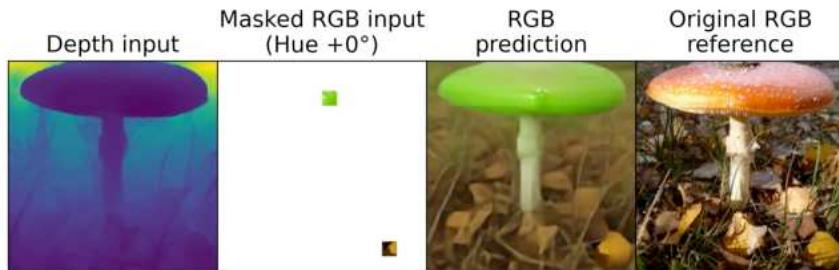
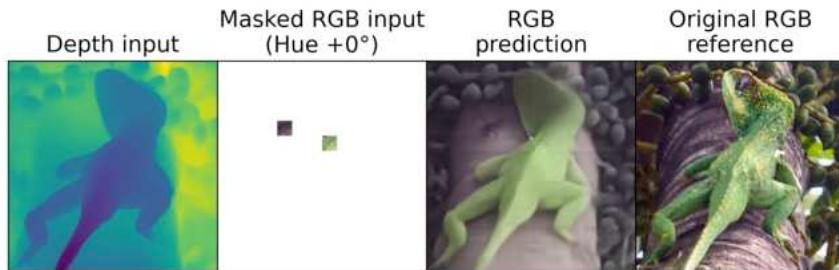
Sampling modalities



Modality-Agnosticity

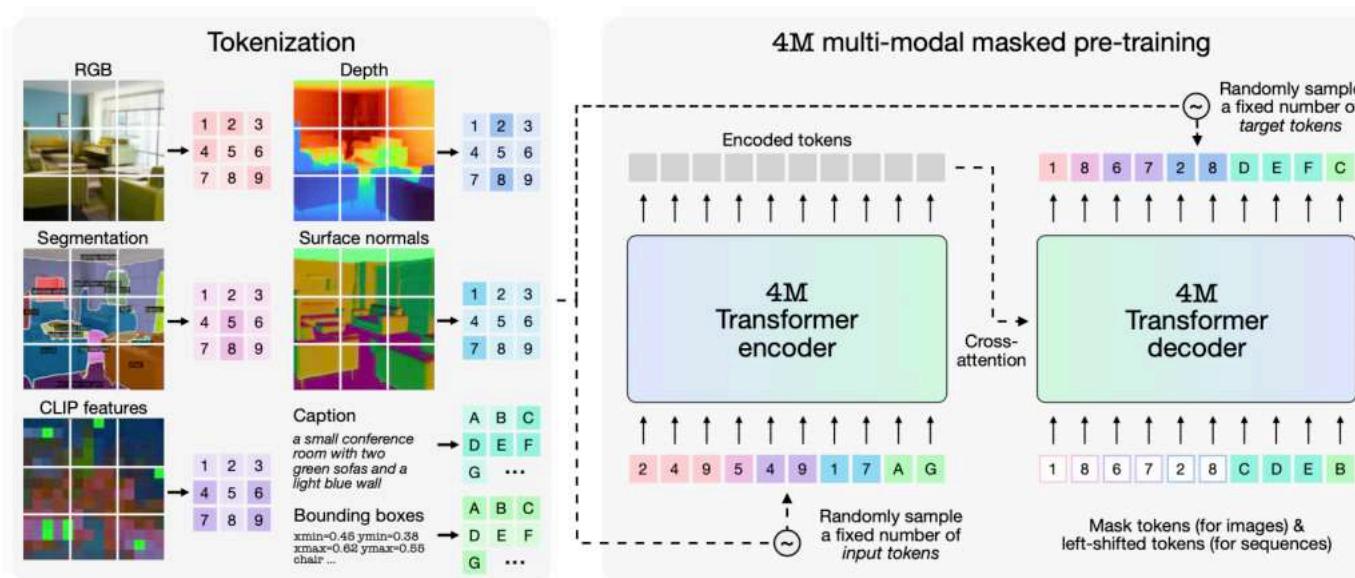






- <https://multimae.epfl.ch/>

4M: Massively Multimodal Masked Modeling



- **Re-designed architecture:** format compatibility, tokenization, randomized token subset training.
- **Scaled up:** tens of modalities. Data and model size to billions scale. Training length trillions of tokens.

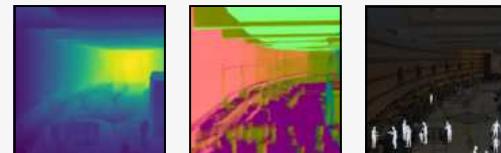
RGB modalities

RGB Color palette



Geometric modalities

Depth Surface normals 3D human poses



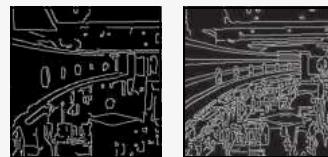
Semantic modalities

Bounding boxes Semantic segmentation SAM instances



Edge modalities

SAM edges Canny edges



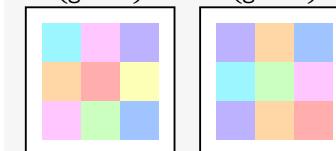
Feature map modalities

CLIP features (dense) DINOv2 features (dense) ImageBind features (dense)



Global feature modalities

DINOv2 features (global) ImageBind features (global)



Text modalities

Caption T5-XXL embeddings Web text

Getting ready for my flight!



Albany International Airport serves as the major air center for the Capital Region, Northeastern ...

Metadata modalities

Image metadata Semantic metadata Geometric metadata

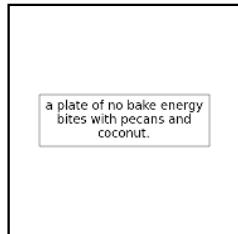
Orig. res.: 512x512
Colorfulness: 35%
Contrast: 45%
Brightness: 60%
Saturation: 40%
...

Humans: 7
Instances: 12
Objectness: 40%
Walkability: 40%
Clutter score: 75%
...

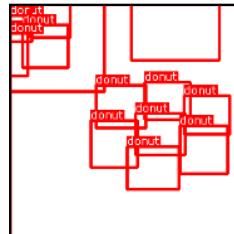
Geometric complexity: 55%
Occlusion score: 25%
...



Query image



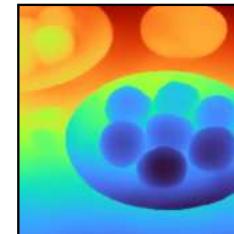
Caption



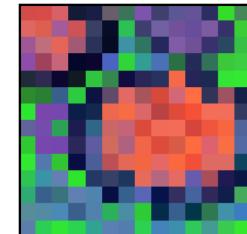
Bounding Boxes



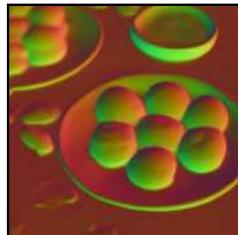
Semantic Seg.



Depth



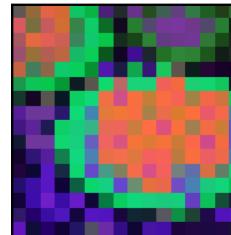
CLIP



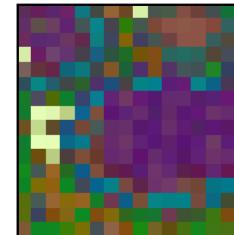
Surface Normals



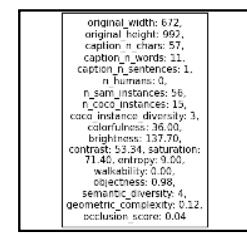
Human poses



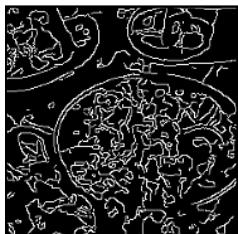
DINOv2



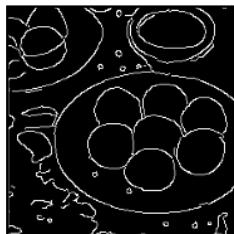
ImageBind



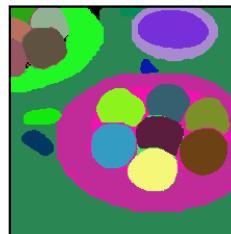
Metadata



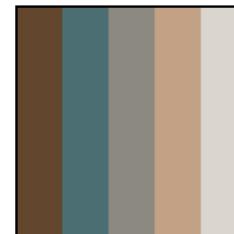
Texture Edges



SAM Edges



SAM instances



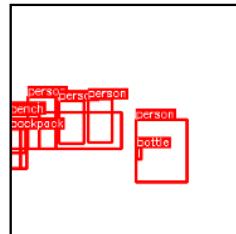
Color Palette



Query image



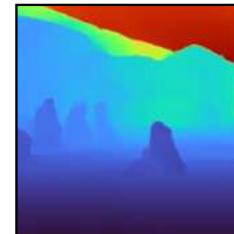
Caption



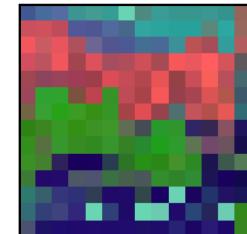
Bounding Boxes



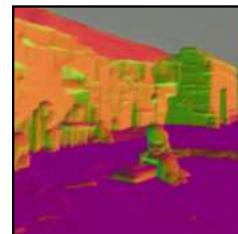
Semantic Seg.



Depth



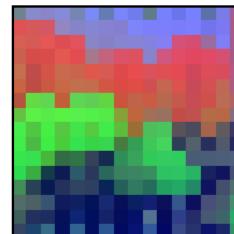
CLIP



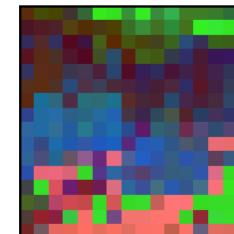
Surface Normals



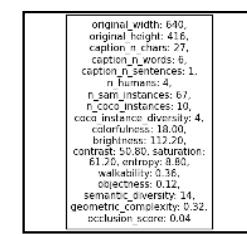
Human poses



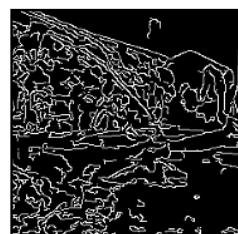
DINOv2



ImageBind



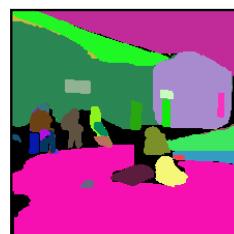
Metadata



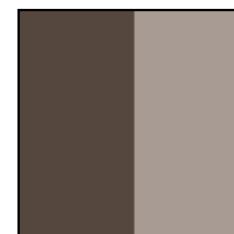
Texture Edges



SAM Edges



SAM instances



Color Palette

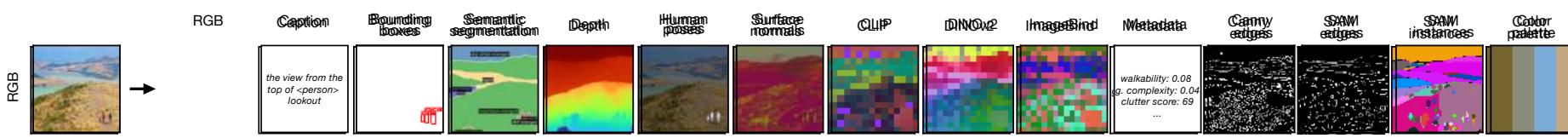
Any-to-Any generation

RGB

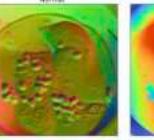
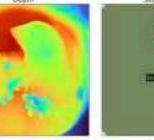
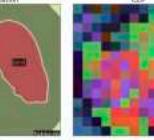
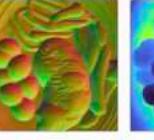
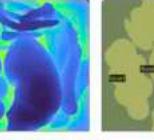
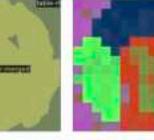
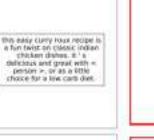
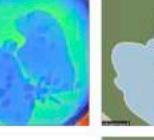
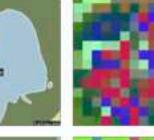
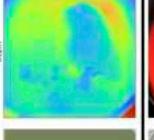
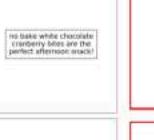
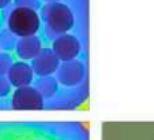
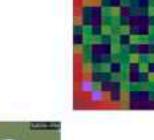
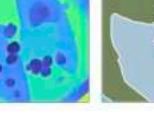


Any-to-Any
generation

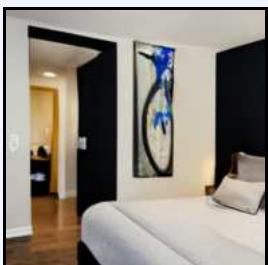
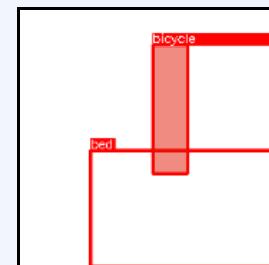
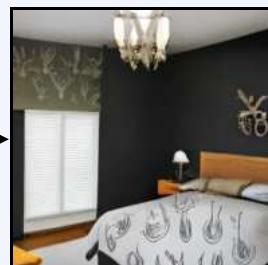
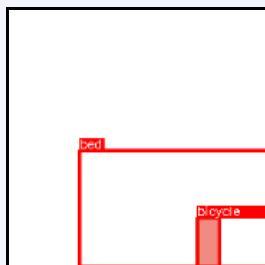
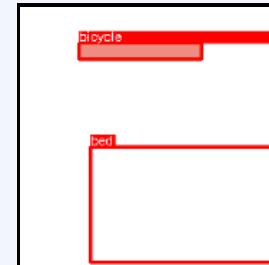
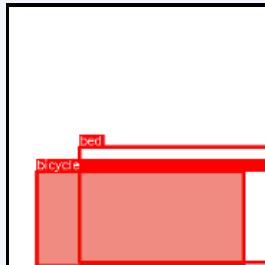
Any-to-Any generation



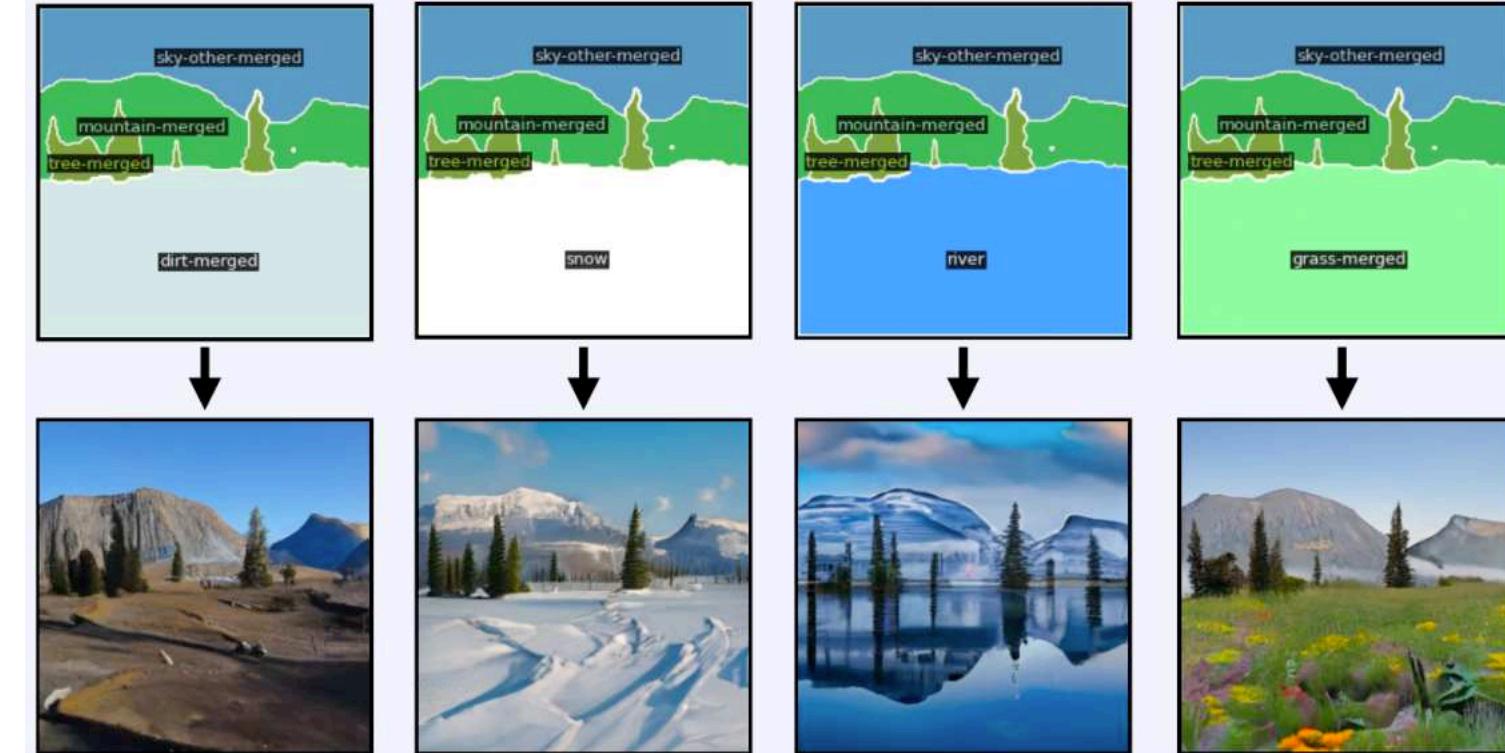
Any-to-Any generation

Input		Prediction						
RGB		RGB		Detection	Normal	Depth	Segmentation	CLIP
RGB								
Caption	chicken breast topped with the ingredients to make this dish make a pressure cooker							
Detection				onion and carrot slices in a saucepan				
Normal				this easy curry makh recipe is a fun twist on classic indian cooking. it's simple to make, delicious and great with n person > or as a little choice for a little extra time				
Depth				kadai chicken in a pot				
Segmentation				no bake white chocolate strawberry bites are the perfect afternoon snack				
CLIP				overhead shot of a white bowl filled with grain, beans and chicken wings on a wooden surface				

Probing the learned model



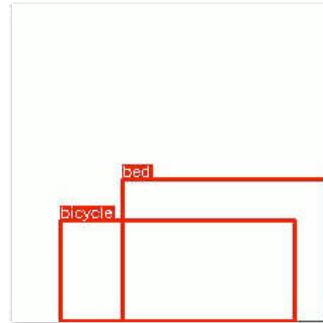
Probing the learned model



Probing the learned model

Probing the learned model

Bounding box input



Caption input

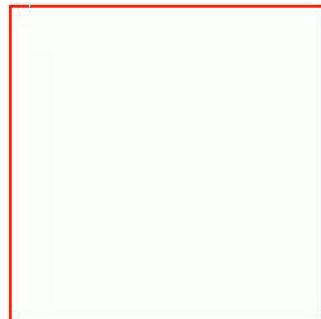
a photo of a
bedroom, studio
light



Frame-by-frame
Predictions

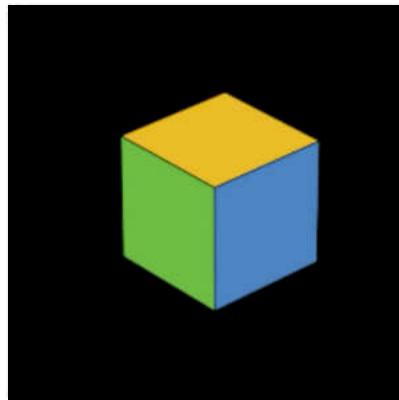
Probing the learned model

Bounding box input



Probing the learned model

Changing SAM polygon input



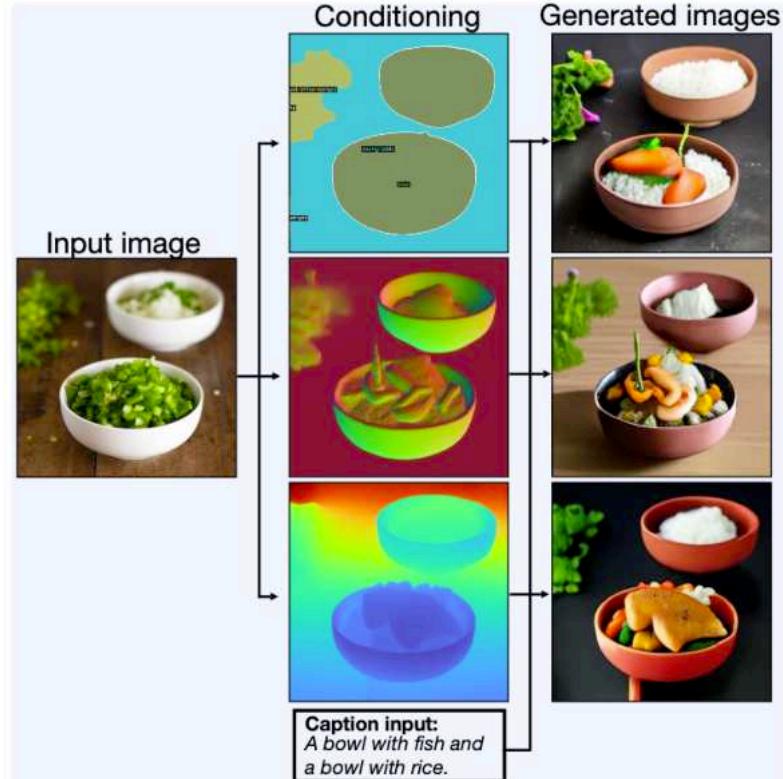
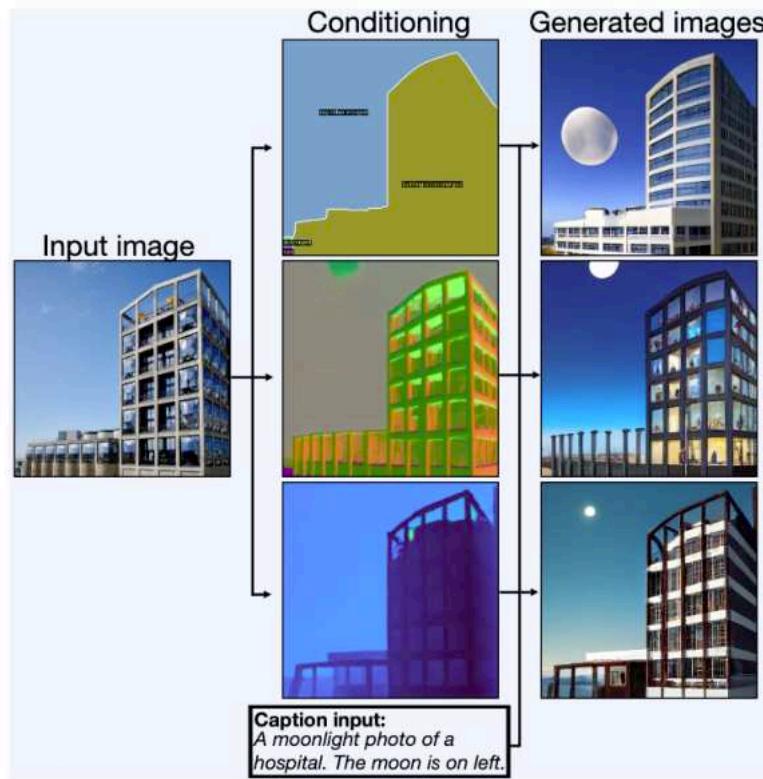
RGB prediction



RGB prediction (with polygon overlay)

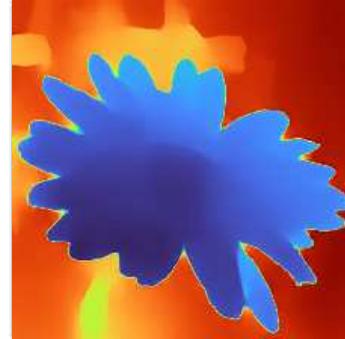


Fixed caption
a framed painting of
mountains inside a
bedroom

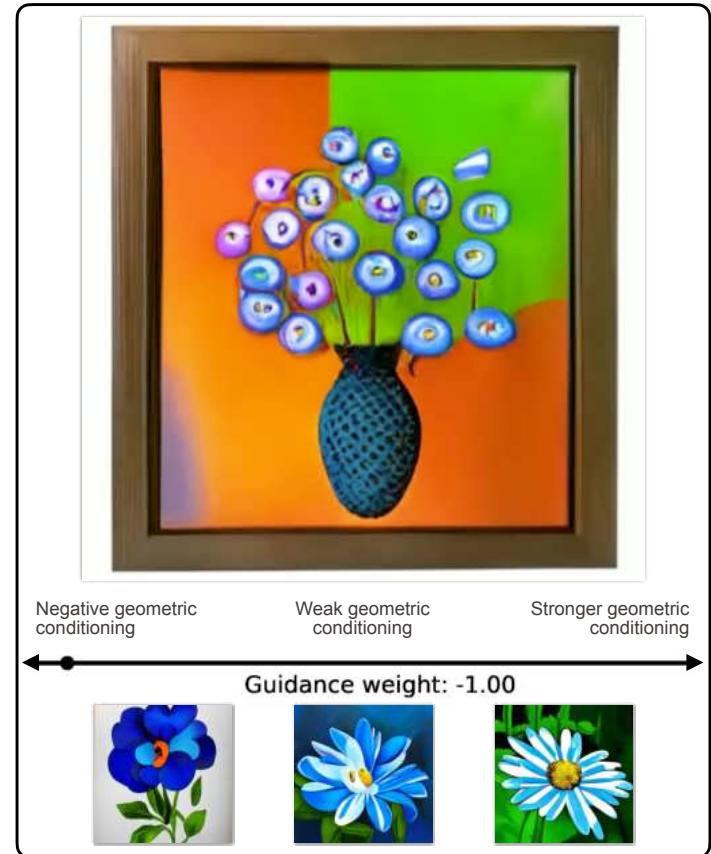


Caption input
an oil painting of
a blue flower

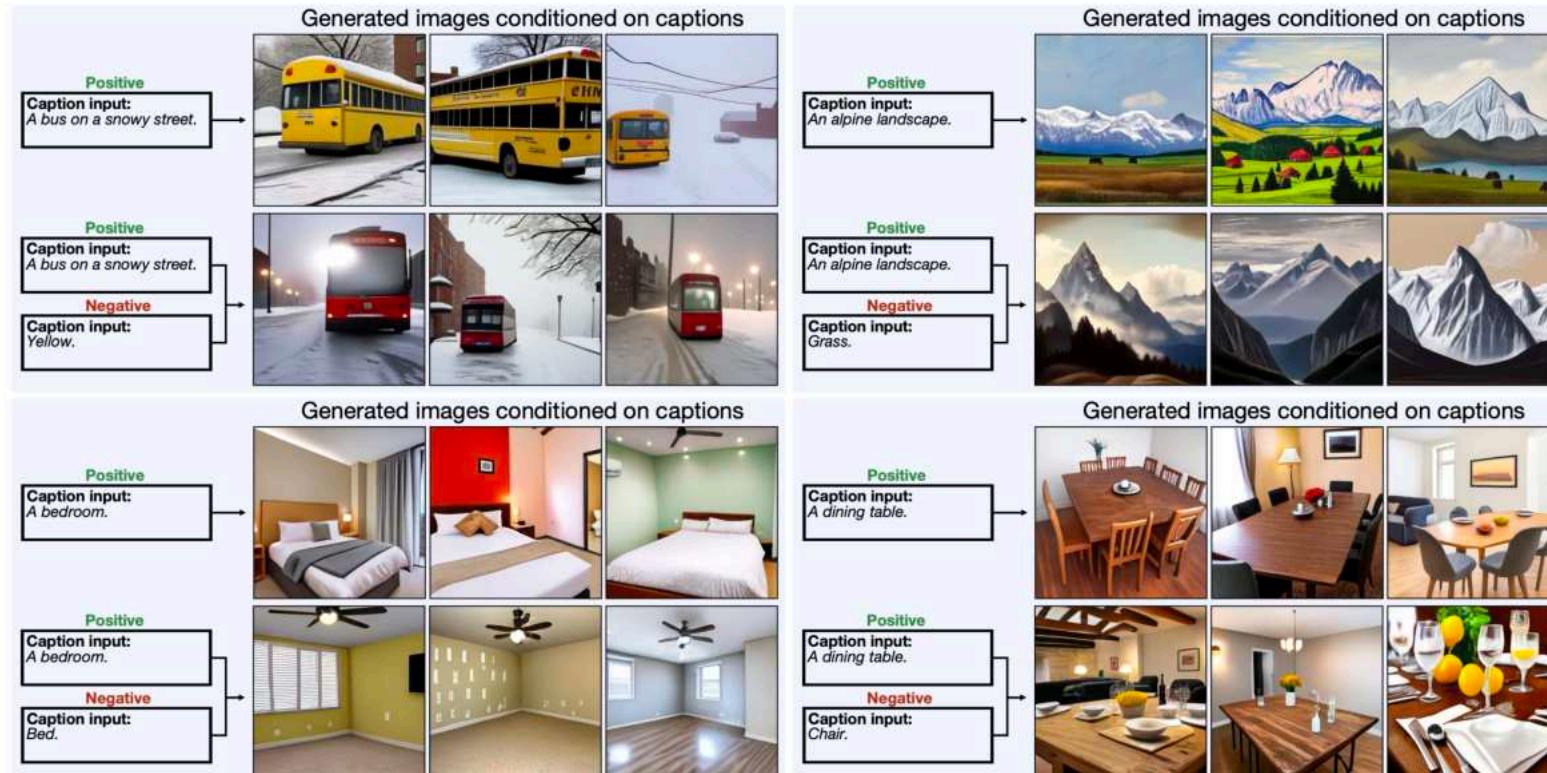
Fixed weight
→
🔒 2.0

Depth


Varying
weight
→

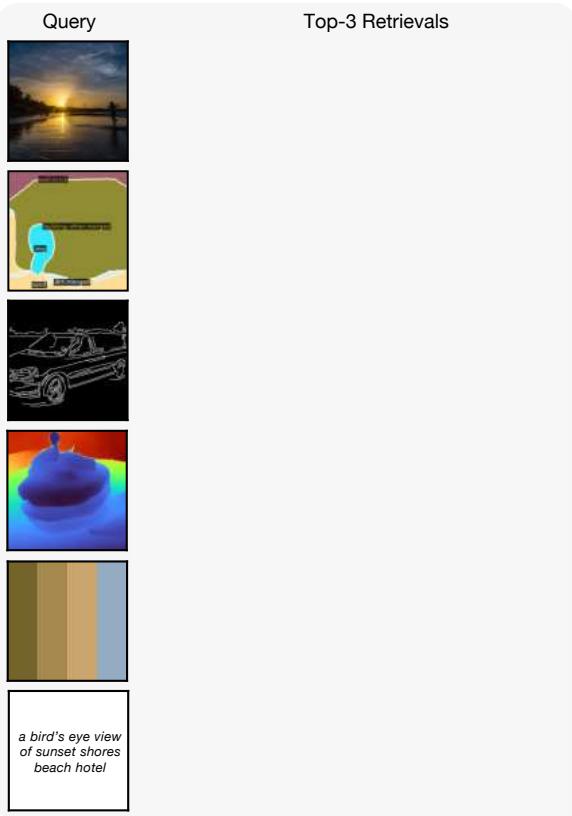


multi-modal guidance & weighting



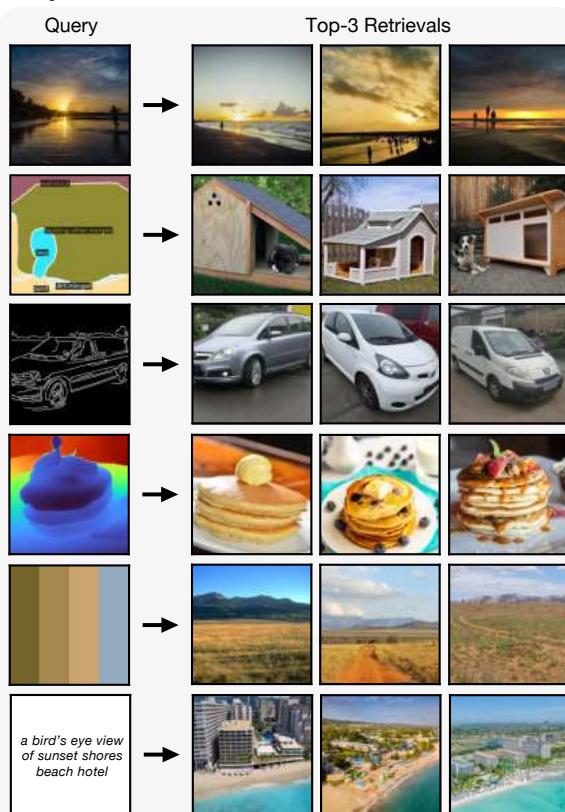
Multimodal retrieval

Any-to-RGB retrieval

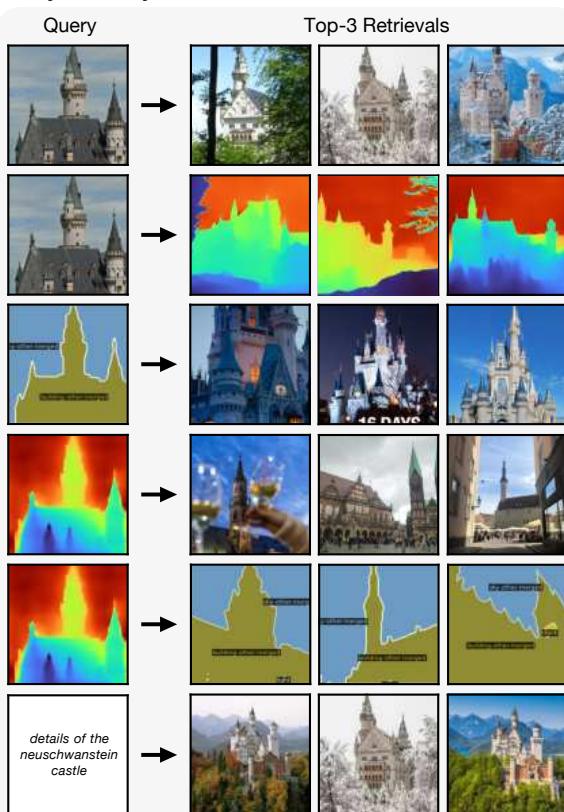


Multimodal retrieval

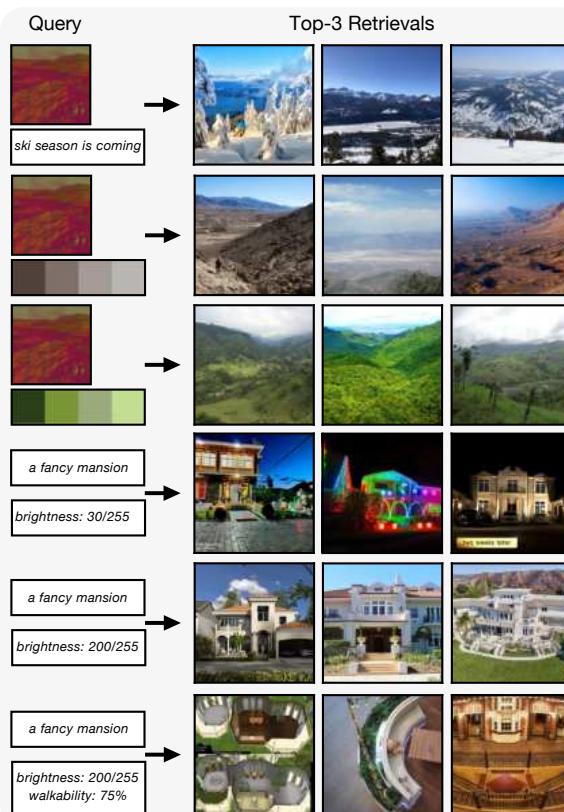
Any-to-RGB retrieval



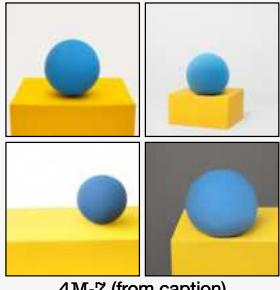
Any-to-any retrieval



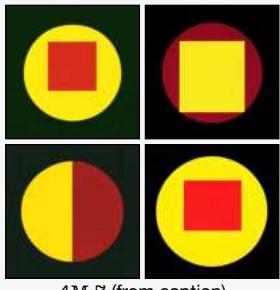
Multimodal retrieval



Caption input: a metallic blue sphere to the left of a yellow box made of felt



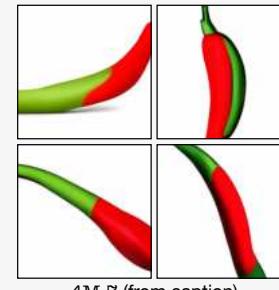
Caption input: a black background with a large yellow circle and a small red square



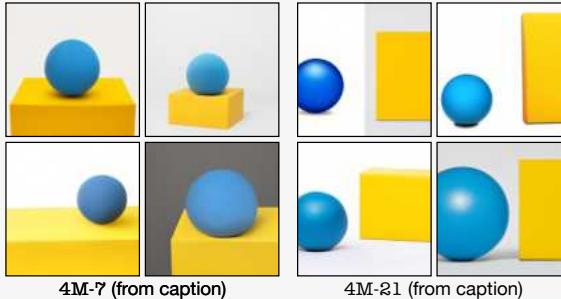
Caption input: a blue semi-truck and its trailer jumping over a row of motorcycles



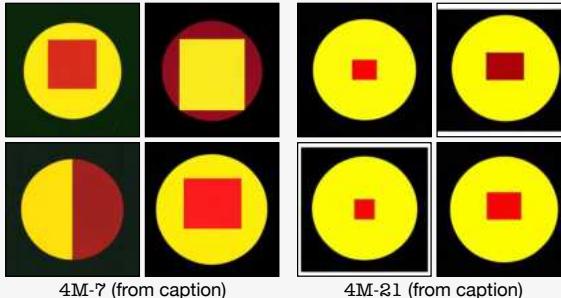
Caption input: a green pepper to the left of a red pepper



Caption input: a metallic blue sphere to the left of a yellow box made of felt



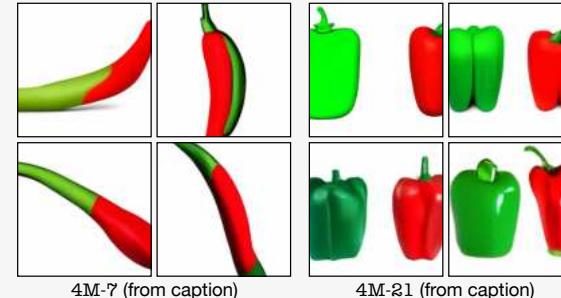
Caption input: a black background with a large yellow circle and a small red square



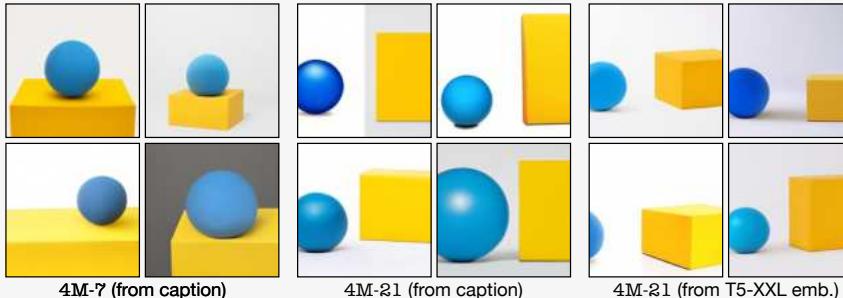
Caption input: a blue semi-truck and its trailer jumping over a row of motorcycles



Caption input: a green pepper to the left of a red pepper



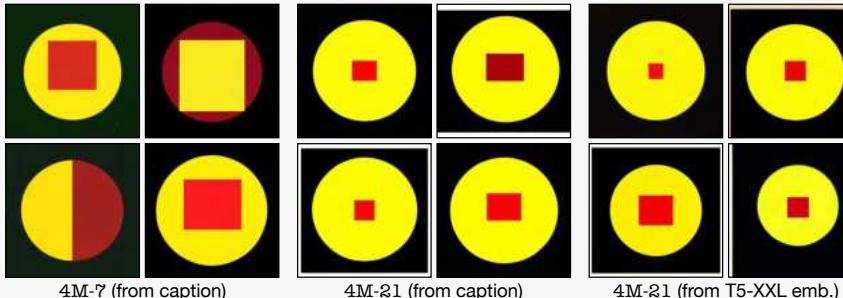
Caption input: a metallic blue sphere to the left of a yellow box made of felt



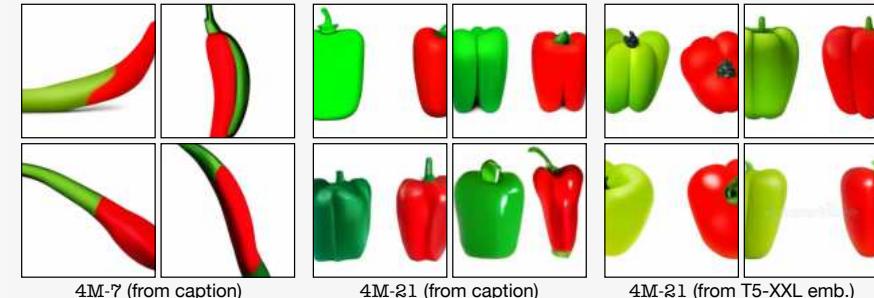
Caption input: a blue semi-truck and its trailer jumping over a row of motorcycles



Caption input: a black background with a large yellow circle and a small red square



Caption input: a green pepper to the left of a red pepper



The screenshot shows a web-based application interface for image processing. The interface is divided into three main sections: Inputs, Outputs, and Results.

Inputs: A section on the left containing a list of items: 'Blurry camera', 'Depth', 'Distorted illumination', 'Caption', 'Bouncing leaves', and 'None'. A 'Add' button is located above the list. A '2a' label and an 'Advanced mode' button are positioned above the list.

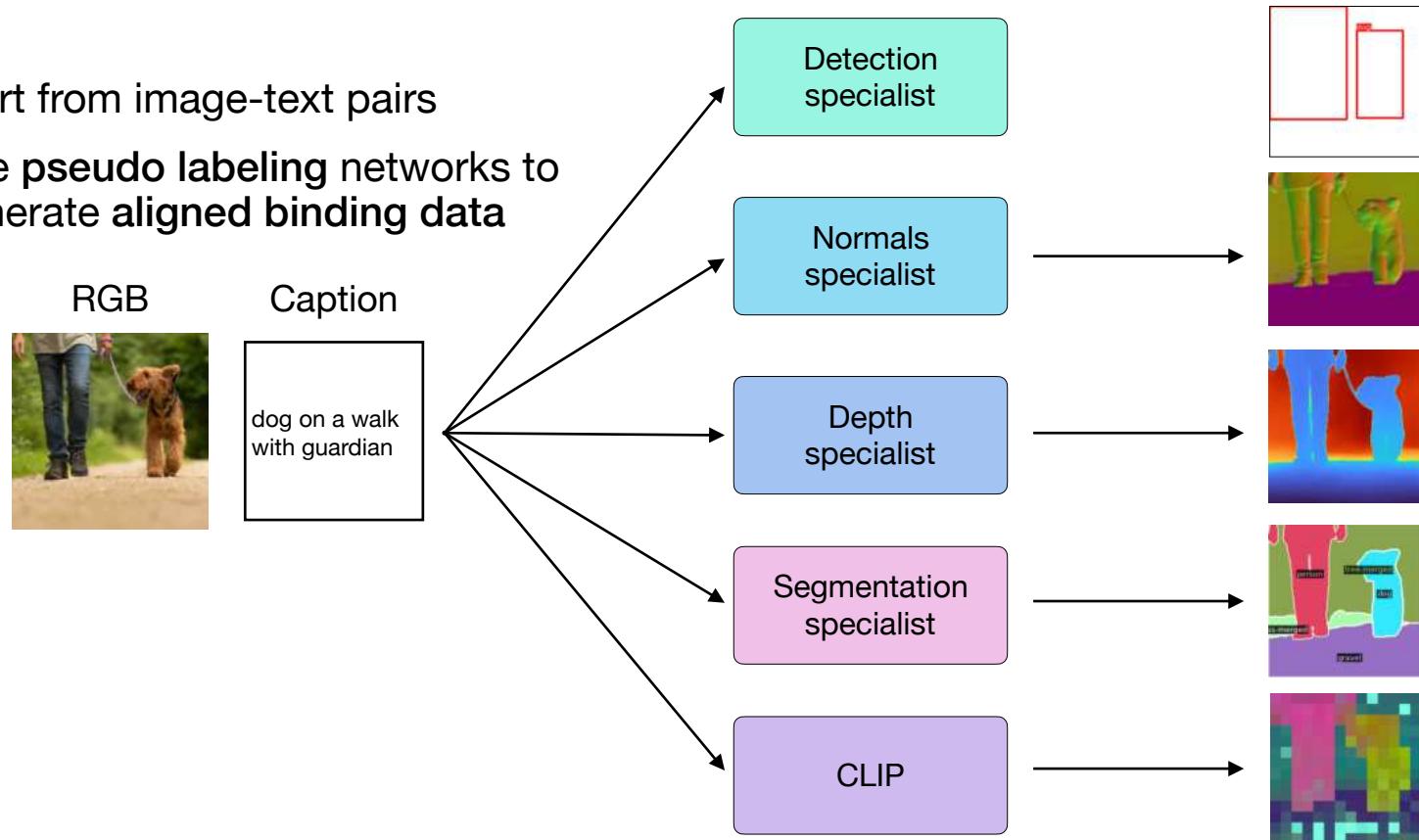
Outputs: A section on the right containing a list of items: 'Super Resolution', 'Model filters', and 'None'. A 'Clear' button is located above the list. A '3a' label and an 'Advanced mode' button are positioned above the list.

Results: A large, empty rectangular area in the center where processed images would be displayed. It contains a 'Download' button, a 'Clear' button, and a 'Copy to clipboard' link.

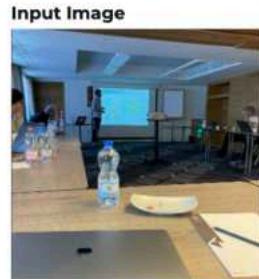
<https://4m.epfl.ch/>

Pseudo Labeling

- Start from image-text pairs
- Use **pseudo labeling** networks to generate aligned binding data



Pseudo Labeling



Upload your own image or click on one of the sample queries below. Click on the cube to use a random query image from previous uploads.

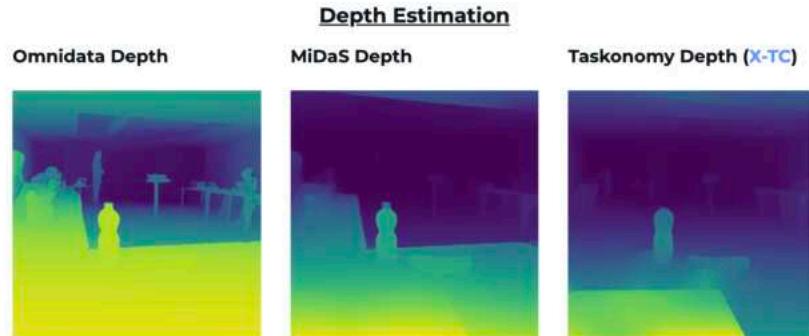


refresh to upload new image

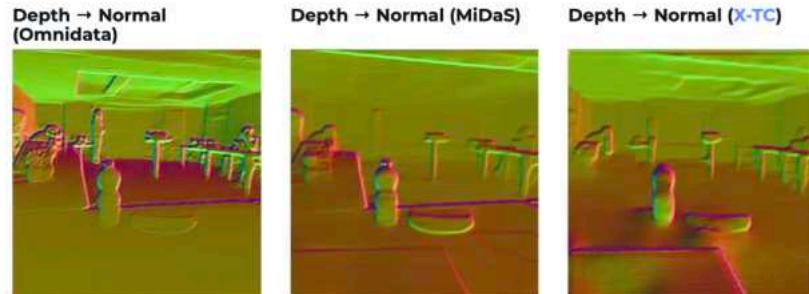
Verification expired. Check the checkbox again.

I'm not a robot

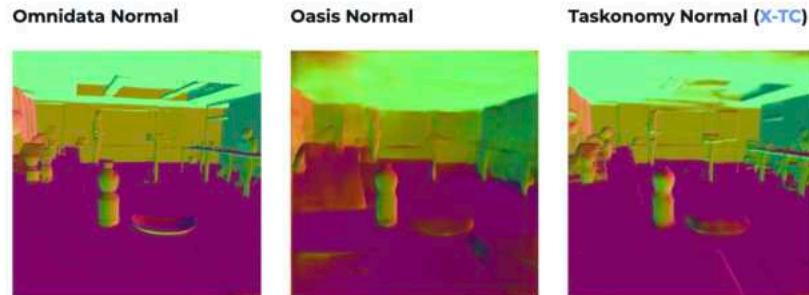
reCAPTCHA



Surface Normals Extracted from Predicted Depth



Surface Normal Estimation



<https://omnidata.vision/demo/>

Omnidata, ICCV'21.

3D Common Corruptions CVPR'22.

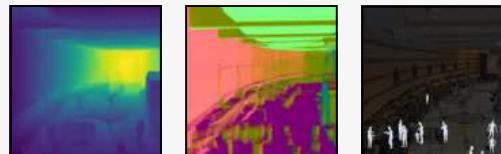
RGB modalities

RGB Color palette



Geometric modalities

Depth Surface normals 3D human poses



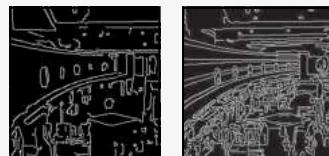
Semantic modalities

Bounding boxes Semantic segmentation SAM instances



Edge modalities

SAM edges Canny edges



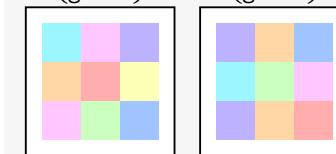
Feature map modalities

CLIP features (dense) DINOv2 features (dense) ImageBind features (dense)



Global feature modalities

DINOv2 features (global) ImageBind features (global)



Text modalities

Caption T5-XXL embeddings Web text

Getting ready for my flight!



Albany International Airport serves as the major air center for the Capital Region, Northeastern ...

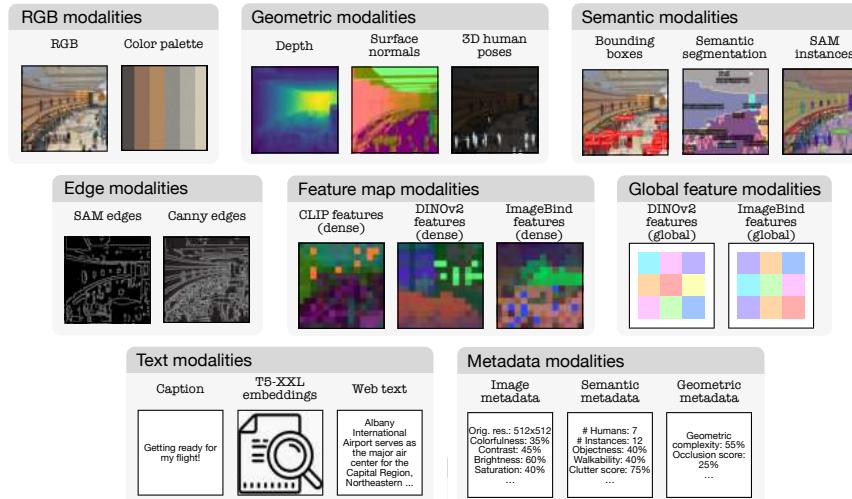
Metadata modalities

Image metadata Semantic metadata Geometric metadata

Orig. res.: 512x512
Colorfulness: 35%
Contrast: 45%
Brightness: 60%
Saturation: 40%
...

Humans: 7
Instances: 12
Objectness: 40%
Walkability: 40%
Clutter score: 75%
...
...

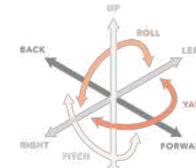
Geometric complexity: 55%
Occlusion score: 25%
...



Planned



Motion

IMU &
Motor control3D & space-level
fusion

Sketch



Video



Multi-view

Out-of-the-box evaluation

Method	Normals ↓	Depth ↓	Sem. seg. ↑	Inst. seg. ↑	IN1K kNN ↑	3D human KP ↓
Pseudo labelers	Omnidata [44]	22.5	0.68	✗	✗	✗
	M2F-B [19]	✗	✗	45.7	✗	✗
	SAM [47]	✗	✗	✗	32.9	✗
	DINOv2-B14 [65]	✗	✗	✗	✗	82.1 / 93.9
	ImageBind-H14 [33]	✗	✗	✗	✗	81.1 / 94.4
	4D-Humans [35]	✗	✗	✗	✗	81.3
	OASIS [18]	34.3	✗	✗	✗	✗
	MiDaS DPT [70]	✗	0.73	✗	✗	✗
	M2F-S [19]	✗	✗	44.6	✗	✗
	M2F-L [19]	✗	✗	48.0	✗	✗
HMR	HMR [43]	✗	✗	✗	✗	130.0
	UnifiedIO-B [59]	35.7	1.00	32.9	✗	✗
	UnifiedIO-L [59]	33.9	0.87	41.6	✗	✗
	UnifiedIO-XL [59]	31.0	0.82	44.3	✗	✗
	UnifiedIO 2-L [58]	37.1	0.96	38.9	✗	✗
	UnifiedIO 2-XL [58]	34.8	0.86	39.7	✗	✗
	UnifiedIO 2-XXL [58]	37.4	0.84	41.7	✗	✗
	4M-7 B [62]	21.9	0.71	43.3	✗	✗
	Ours B	21.7	0.71	42.5	15.9	73.1 / 89.7
	4M-7 L [62]	21.5	0.69	47.2	✗	✗
Ours	Ours L	21.1	0.69	46.4	31.2	77.0 / 91.9
	4M-7 XL [62]	20.6	0.69	48.1	✗	✗
	Ours XL	20.8	0.68	48.1	<u>32.0</u>	78.3 / 92.4
	Tokenizer bound*	4.0	0.06	90.5	91.2	80.2 / 93.0

- The multitask learning aspect works well-> one effective network for 100s of tasks.

Out-of-the-box evaluation

Method	Normals ↓	Depth ↓	Sem. seg. ↑	Inst. seg. ↑	IN1K kNN ↑	3D human KP ↓
Pseudo labelers						
Omnidata [44]	22.5	0.68	x	x	x	x
M2F-B [19]	x	x	45.7	x	x	x
SAM [47]	x	x	x	32.9	x	x
DINOv2-B14 [65]	x	x	x	x	82.1 / 93.9	x
ImageBind-H14 [33]	x	x	x	x	81.1 / 94.4	x
4D-Humans [35]	x	x	x	x	x	81.3
OASIS [18]	34.3	x	x	x	x	x
MiDaS DPT [70]	x	0.73	x	x	x	x
M2F-S [19]	x	x	44.6	x	x	x
M2F-L [19]	x	x	48.0	x	x	x
HMR [43]	x	x	x	x	x	130.0
UnifiedIO-B [59]	35.7	1.00	32.9	x	x	x
UnifiedIO-L [59]	33.9	0.87	41.6	x	x	x
UnifiedIO-XL [59]	31.0	0.82	44.3	x	x	x
UnifiedIO 2-L [58]	37.1	0.96	38.9	x	x	x
UnifiedIO 2-XL [58]	34.8	0.86	39.7	x	x	x
UnifiedIO 2-XXL [58]	37.4	0.84	41.7	x	x	x
4M-7 B [62]	21.9	0.71	43.3	x	x	x
Ours B	21.7	0.71	42.5	15.9	73.1 / 89.7	108.3
4M-7 L [62]	21.5	0.69	47.2	x	x	x
Ours L	21.1	0.69	46.4	31.2	77.0 / 91.9	97.4
4M-7 XL [62]	20.6	0.69	48.1	x	x	x
Ours XL	20.8	0.68	48.1	32.0	78.3 / 92.4	92.0
Tokenizer bound*	4.0	0.06	90.5	91.2	80.2 / 93.0	17.5

- The multitask learning aspect works well-> one effective network for 100s of tasks.
- Long-way to go in terms of transfer/emergence.

Multimodal transfer

Method	NYUv2-S		Hypersim		ARKitScenes	
	mIoU ↑		mIoU ↑		AP ^{3D} ↑	
4M-7 B	56.6	57.5	40.2	43.9	40.3	46.5
Ours B	58.7	59.7	38.6	46.4	42.4	48.1
4M-7 L	61.2	61.4	48.7	50.5	46.8	49.5
Ours L	61.8	61.8	47.3	50.7	47.0	50.1
4M-7 XL	62.1	61.2	48.6	51.0	48.1	50.1
Ours XL	63.9	63.9	48.6	52.5	48.4	51.3

Unimodal transfer

Method	Pre-training data	Enc. param.	IN1K	ADE20K	NYUv2-D	ARKS
MAE B [38]	IN1K		84.2	46.1	89.1	30.9
DeiT III B [83]	IN21K		85.4	49.0	87.4	36.1
MultiMAE B [7]	IN1K		84.0	46.2	89.0	34.2
DINOv2 B [65]	LVD142M	86M	85.3	51.6	92.2	38.1
4M-7 B [62]	CC12M		84.5	50.1	92.0	40.3
4M-7 B (Ours)	COYO		84.4	49.4	91.4	38.6
Ours B	CC12M+COYO+C4		84.5	50.1	90.8	42.4
MAE L [38]	IN1K		86.8	51.8	93.6	36.2
DeiT III L [83]	IN21K		87.0	52.0	89.6	40.3
DINOv2 L [65]	LVD142M	303M	86.7	53.4	94.1	42.8
4M-7 L [62]	CC12M		86.6	53.4	94.4	46.8
4M-7 L (Ours)	COYO		86.7	53.5	94.3	45.2
Ours L	CC12M+COYO+C4		86.5	53.4	93.7	47.0
DINOv2 g [65]	LVD142M	1.1B	88.0	58.7	92.5	45.3
4M-7 XL [62]	CC12M		87.0	55.0	96.1	48.1
4M-7 XL (Ours)	COYO	1.2B	87.1	56.1	96.5	47.3
Ours XL	CC12M+COYO+C4		87.1	56.0	96.5	48.4

looking correct vs being correct

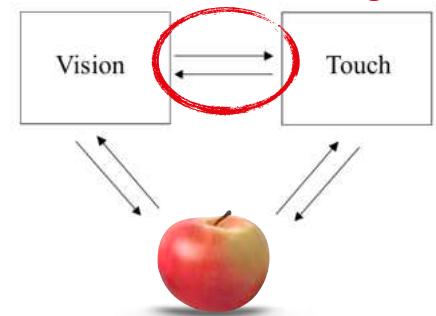
Text Prompt: a giant gorilla at the top of the Empire State building.



4M without co-training with LLM/T5



4M with co-training with LLM/T5

Cross-Modal Learning

Multimodality as self-Supervision

Learn from the entire world/internet
with few modalities



Learn from the test space only
with rich modalities



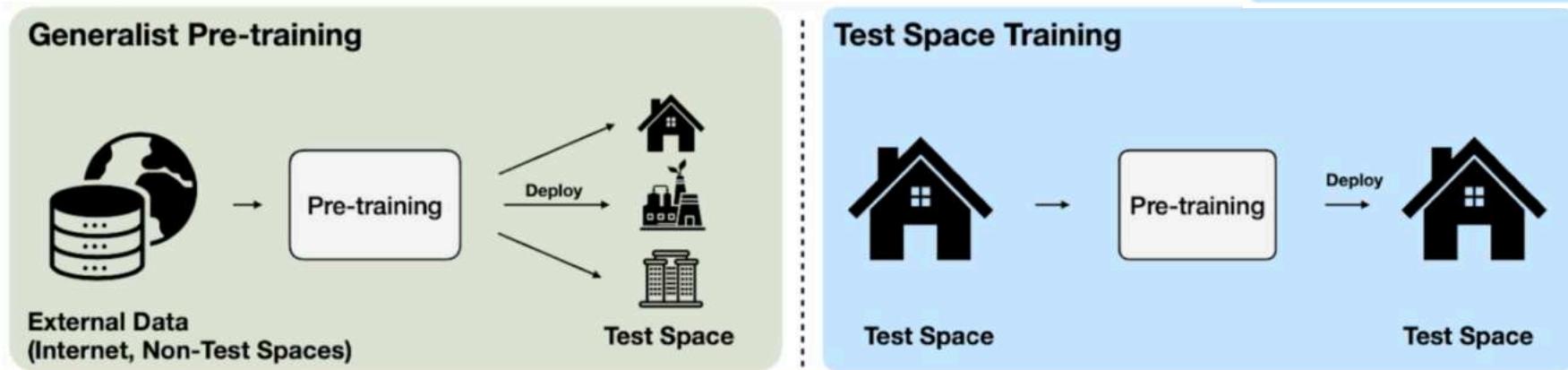
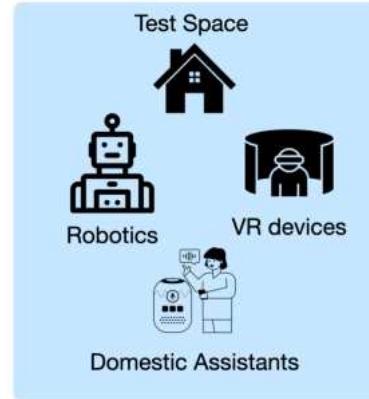
- Limit the world to the test space and “overfit” to it.
- Can we perfectly solve vision there?

EPFL Multimodality as self-Supervision

78

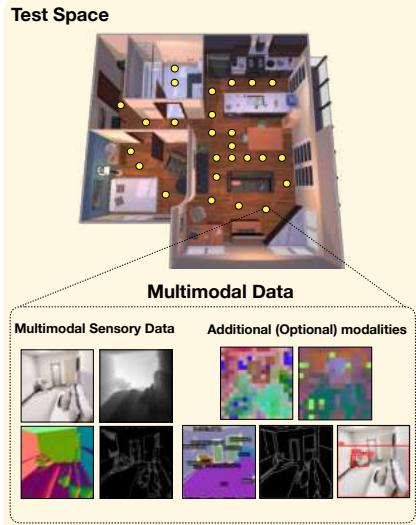
Zamir

- Test-Space Training (TST): Investigates the role of
 - 1) **specialization**, in contrast to generalization.
 - 2) **internet data** in training (multimodal) FMs.

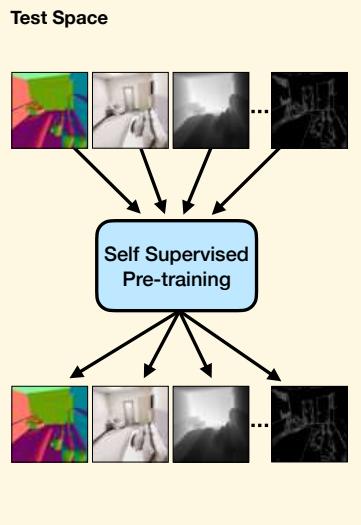


Test-Space Training

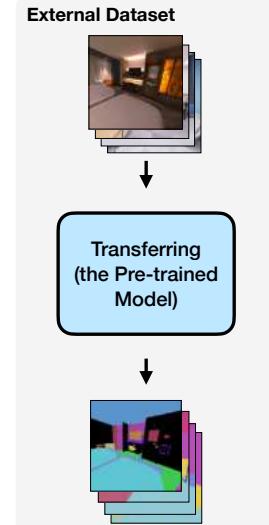
1. Data Collection



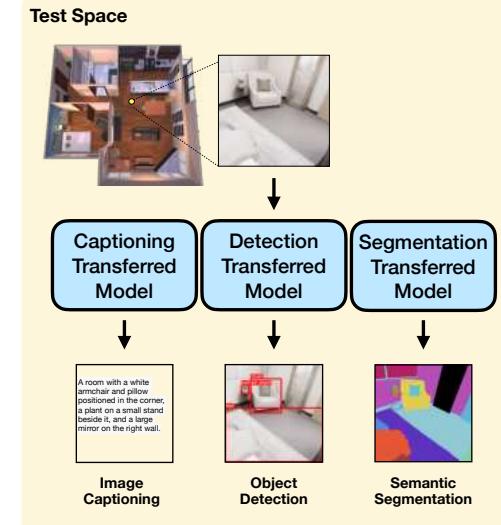
2. Pre-training



3. Transfer



4. Deployment

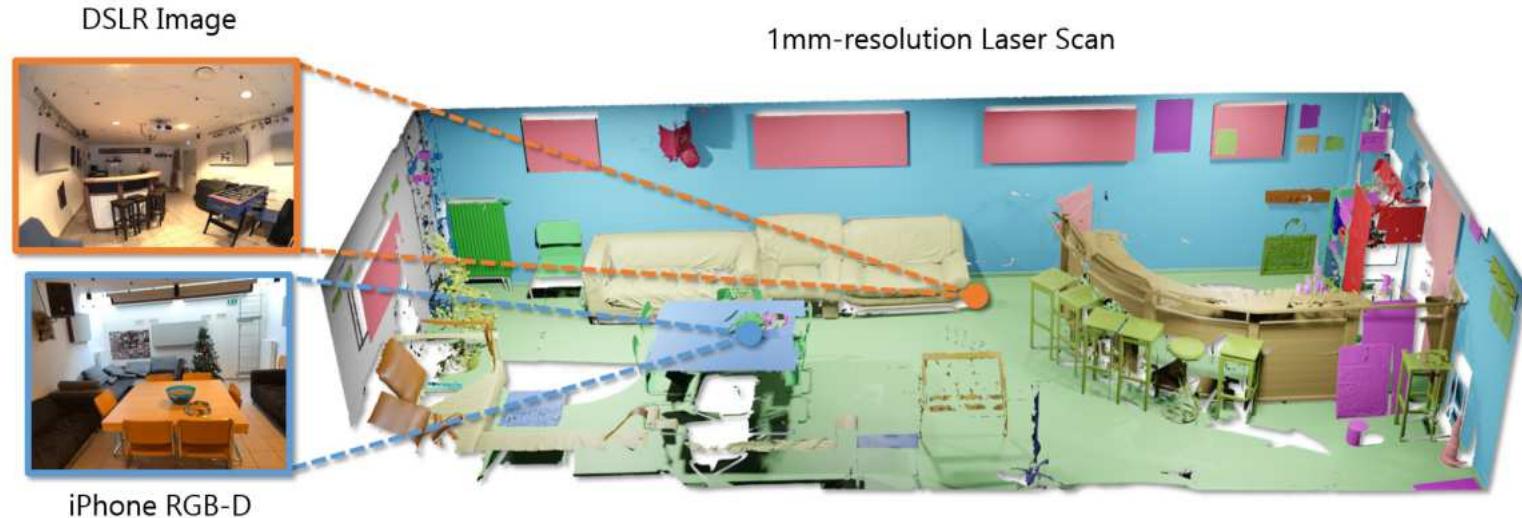


Experimental results

In Scannet++, Replica, THOR

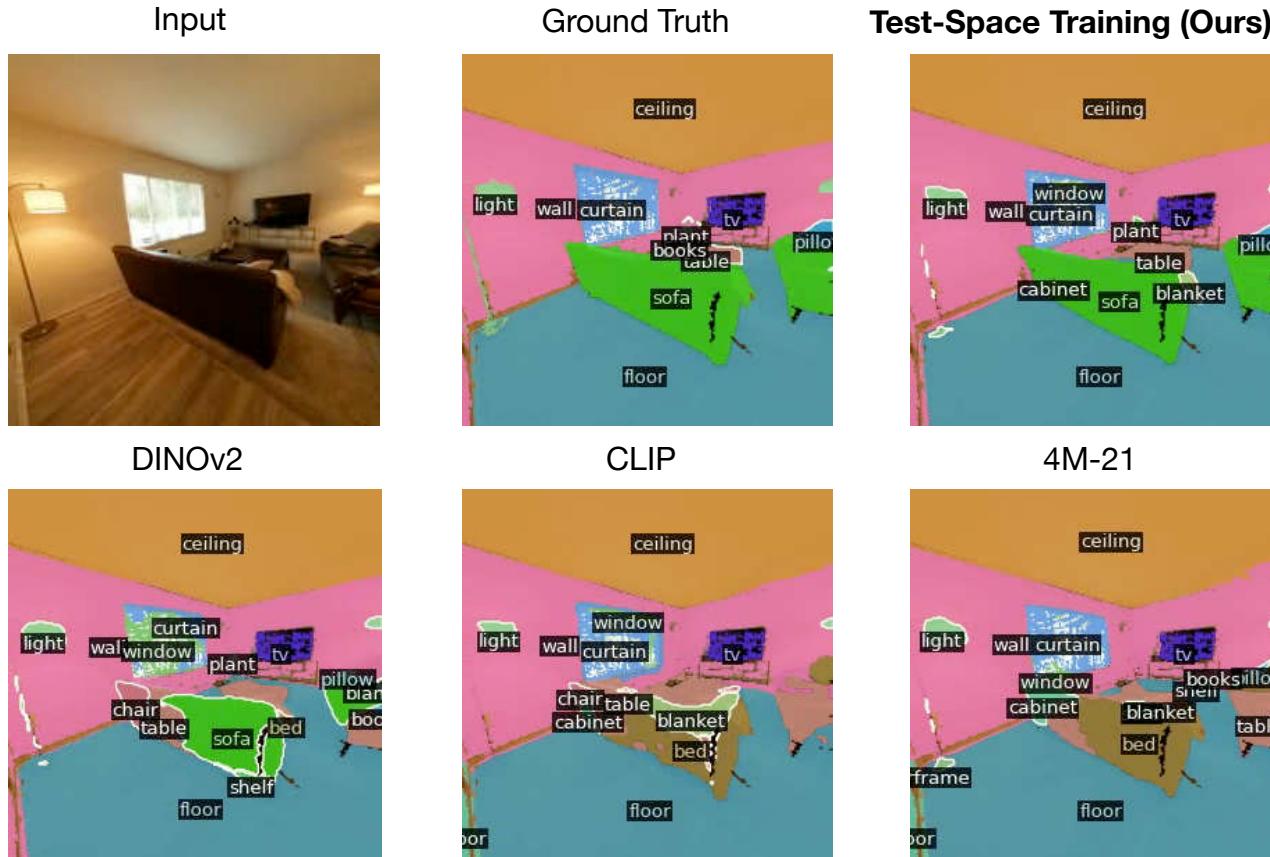
On semantic segmentation, detection, captioning.

Scannet++¹



1. Yeshwanth et al. 2023

Semantic Segmentation Vs. internet-based generalists



Semantic Segmentation

Vs. task specialists

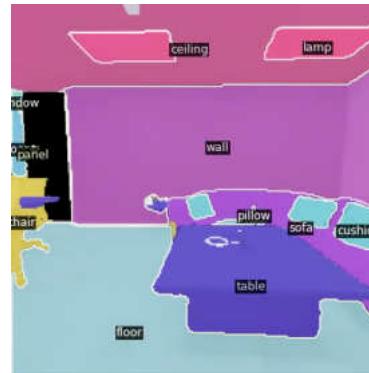
Input



Ground Truth



Test-Space Training (Ours)



Mask2Former



A small, cozy living room with a brown leather sofa, a wooden coffee table, and a television on a stand. A window is on the left, and a lamp is on the right.



Detection

Input



Ground Truth



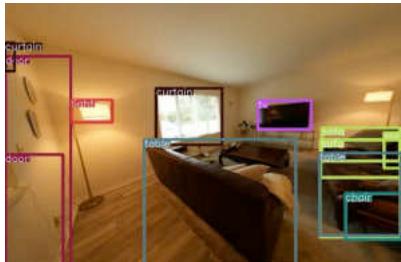
Test-Space Training (Ours)



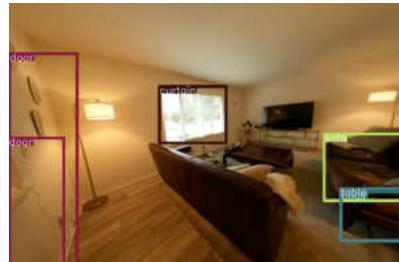
CLIP



4M-21



ViTDet



Scratch



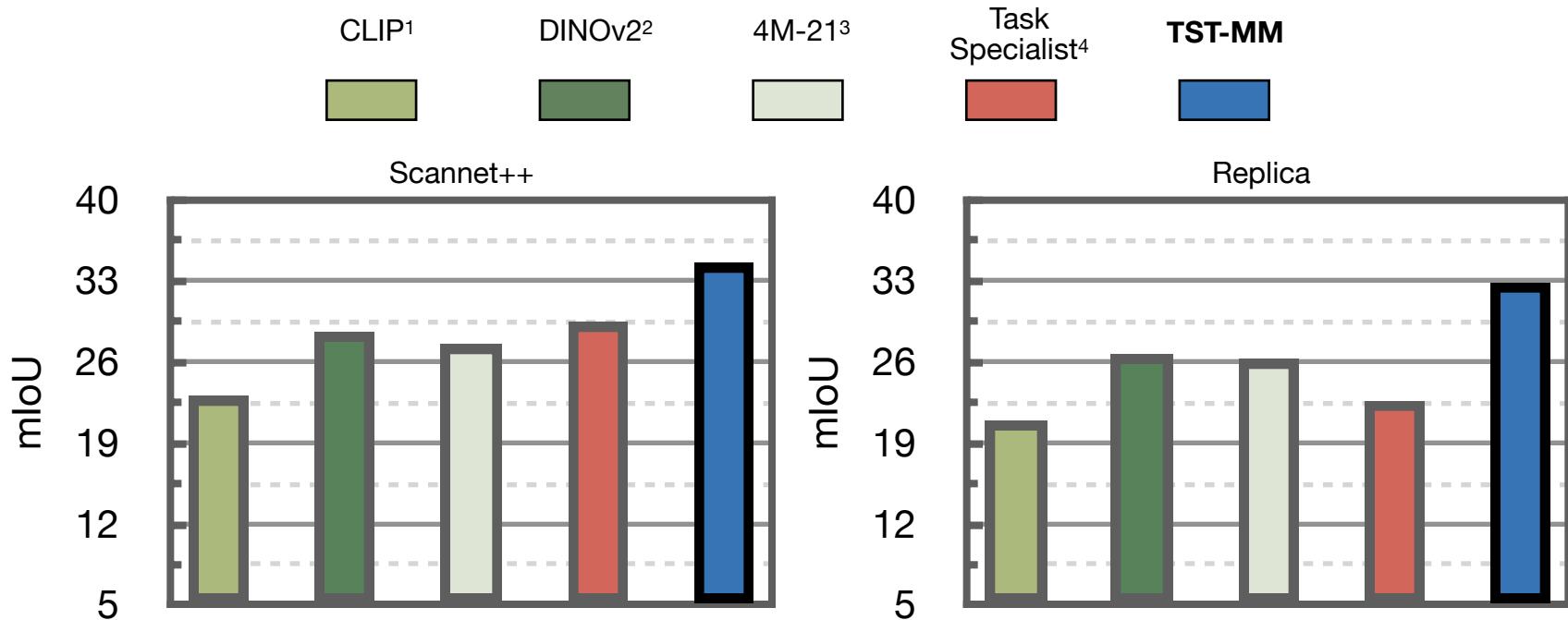
DINOv2



Quantitative comparison

Semantic Segmentation

- TST outperforms internet based generalists^[1,2,3], and task specialists^[4,5].



1. Radford et al. 2021

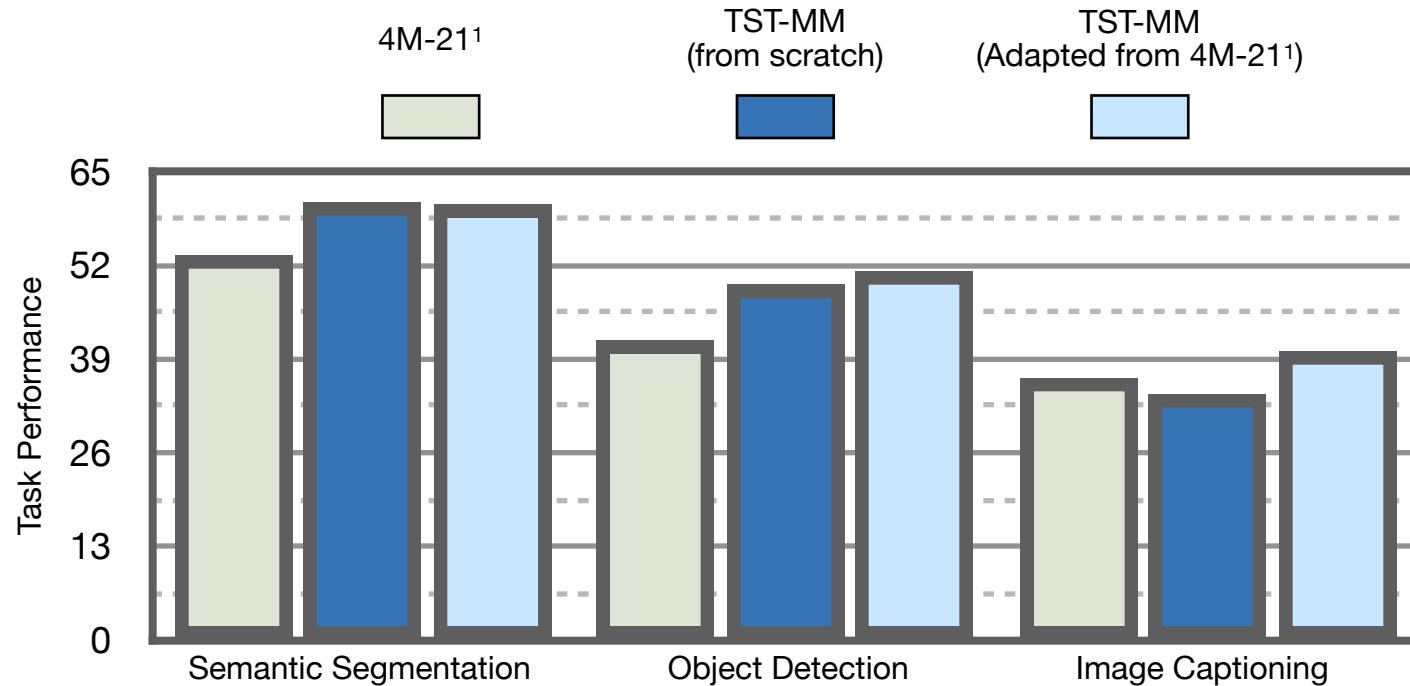
2. Squab et al. 2023

3. Bachmann*, Kar*, Mizrahi* et al. 2024

4. Cheng et al. 2022

Adaptation

Adapt a pre-trained generalist vs. train from scratch

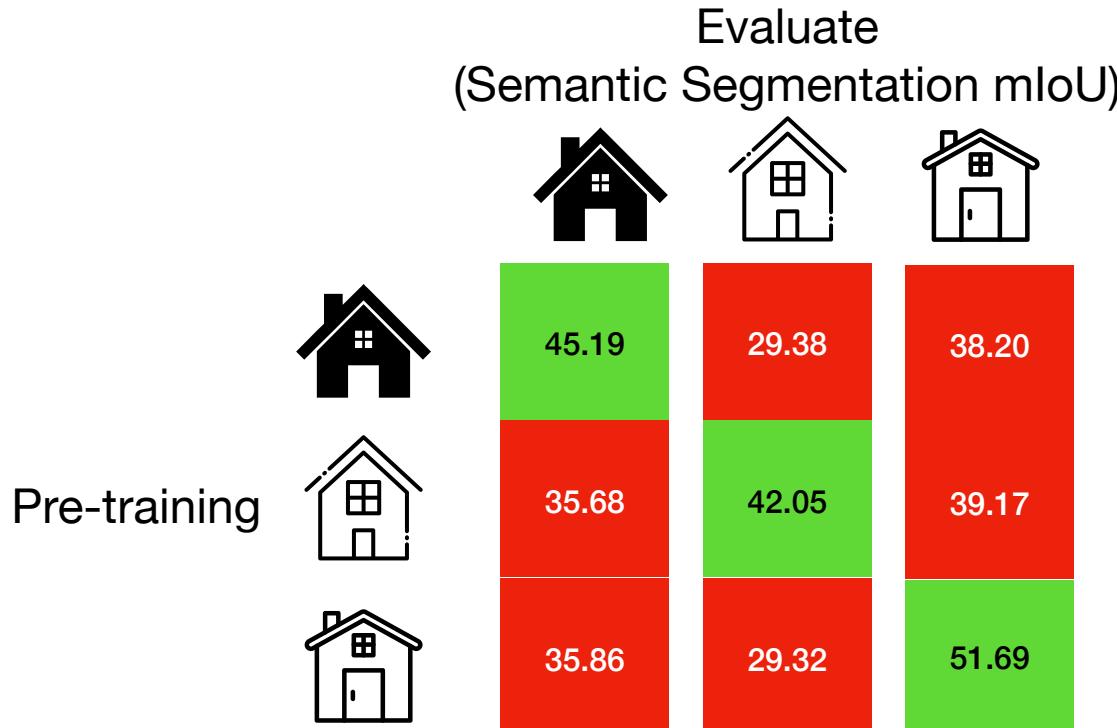


Analysis

Analysis 1. Is “specialization” actually happening?

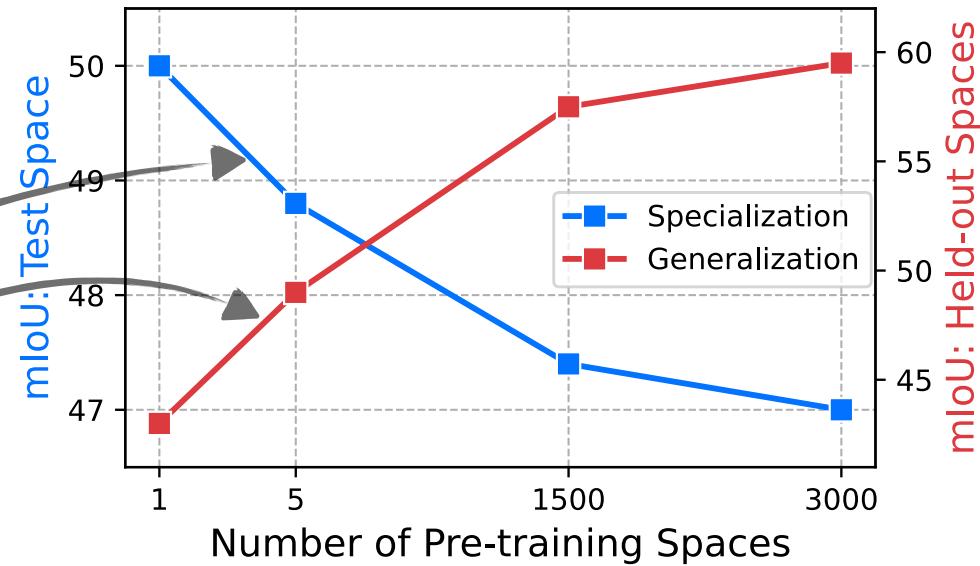
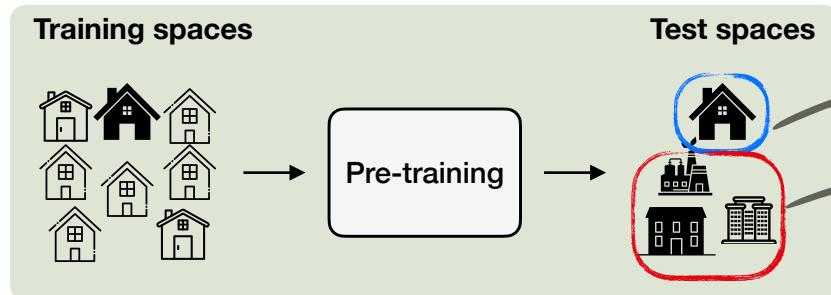
Analysis

Analysis 1. Is “specialization” actually happening?



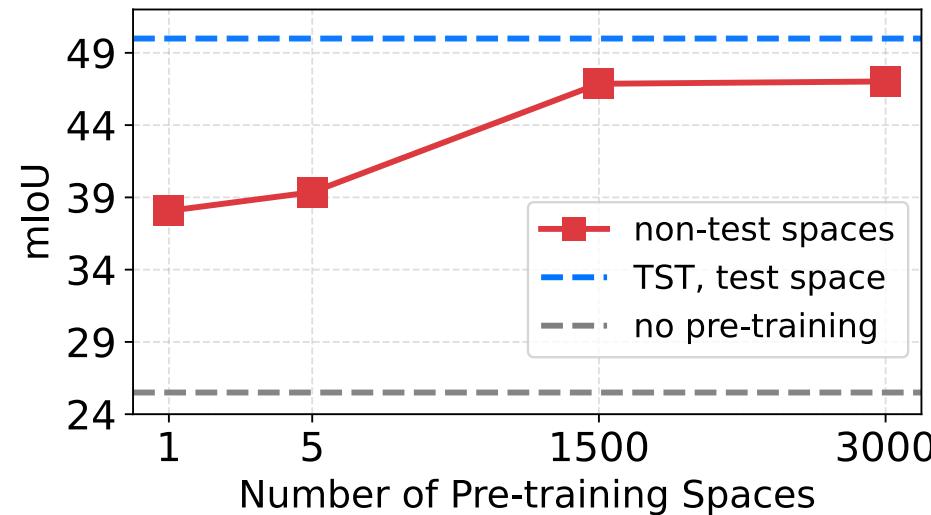
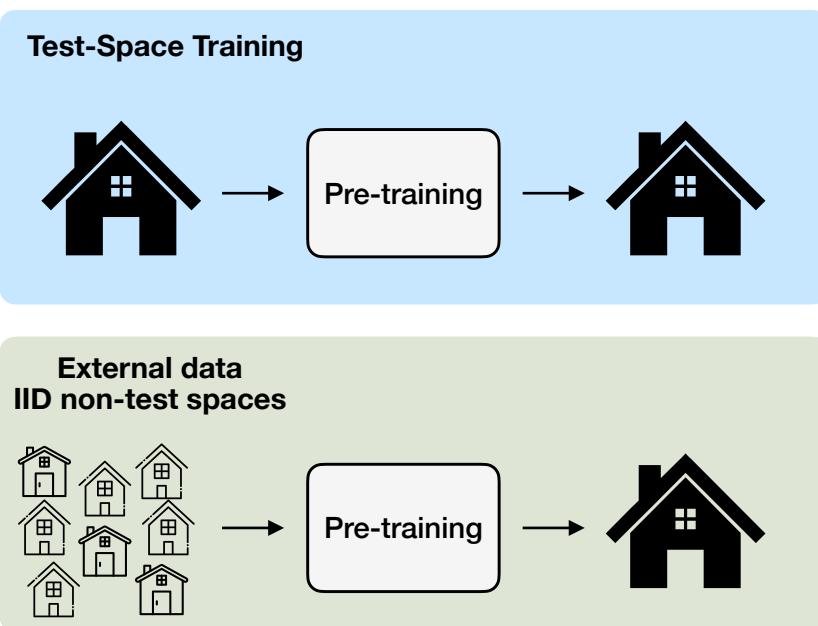
Analysis

Analysis 2. Specialization-generalization tradeoff



Analysis

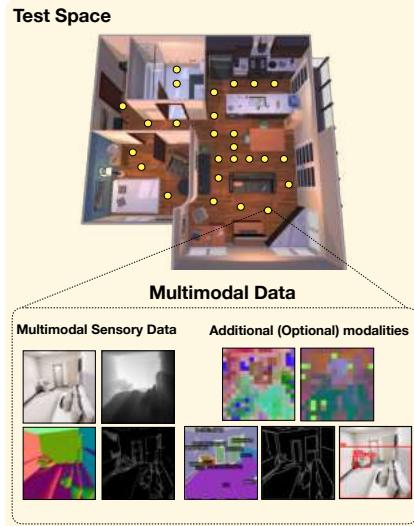
Analysis 3. How much external data is the test-space data worth?



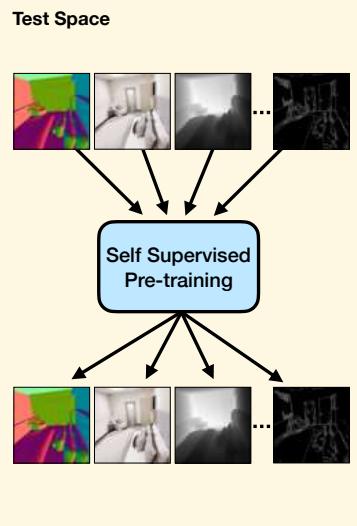
Analysis

Analysis 4. What about other self-supervised objectives?

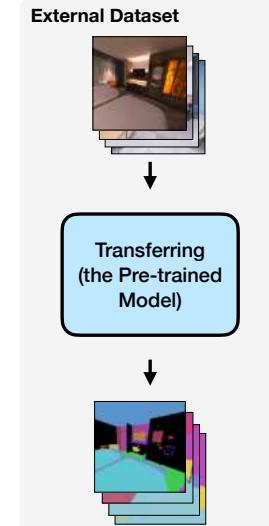
1. Data Collection



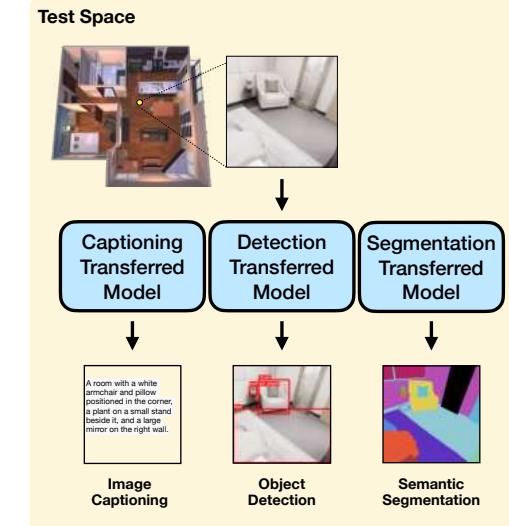
2. Pre-training



3. Transfer

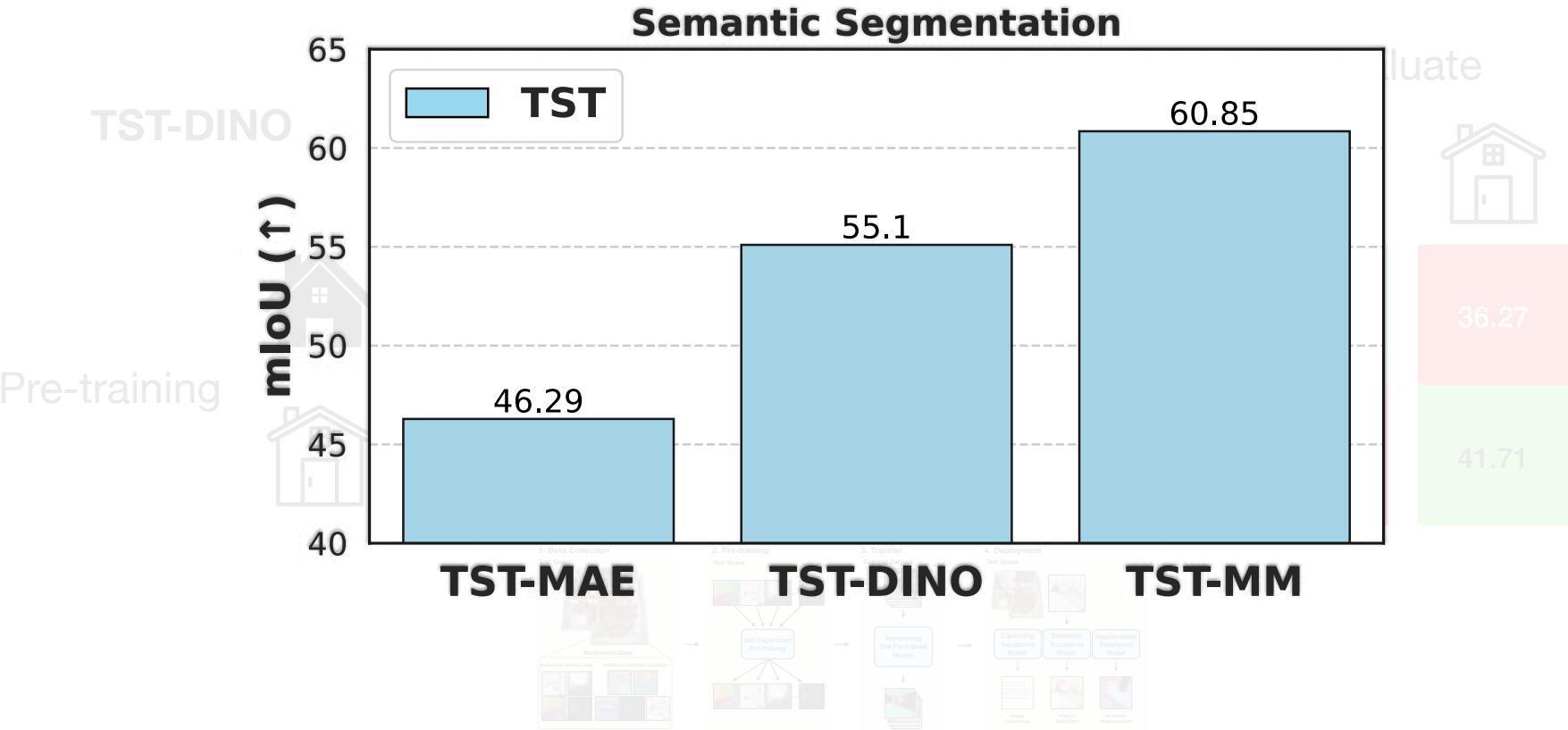


4. Deployment



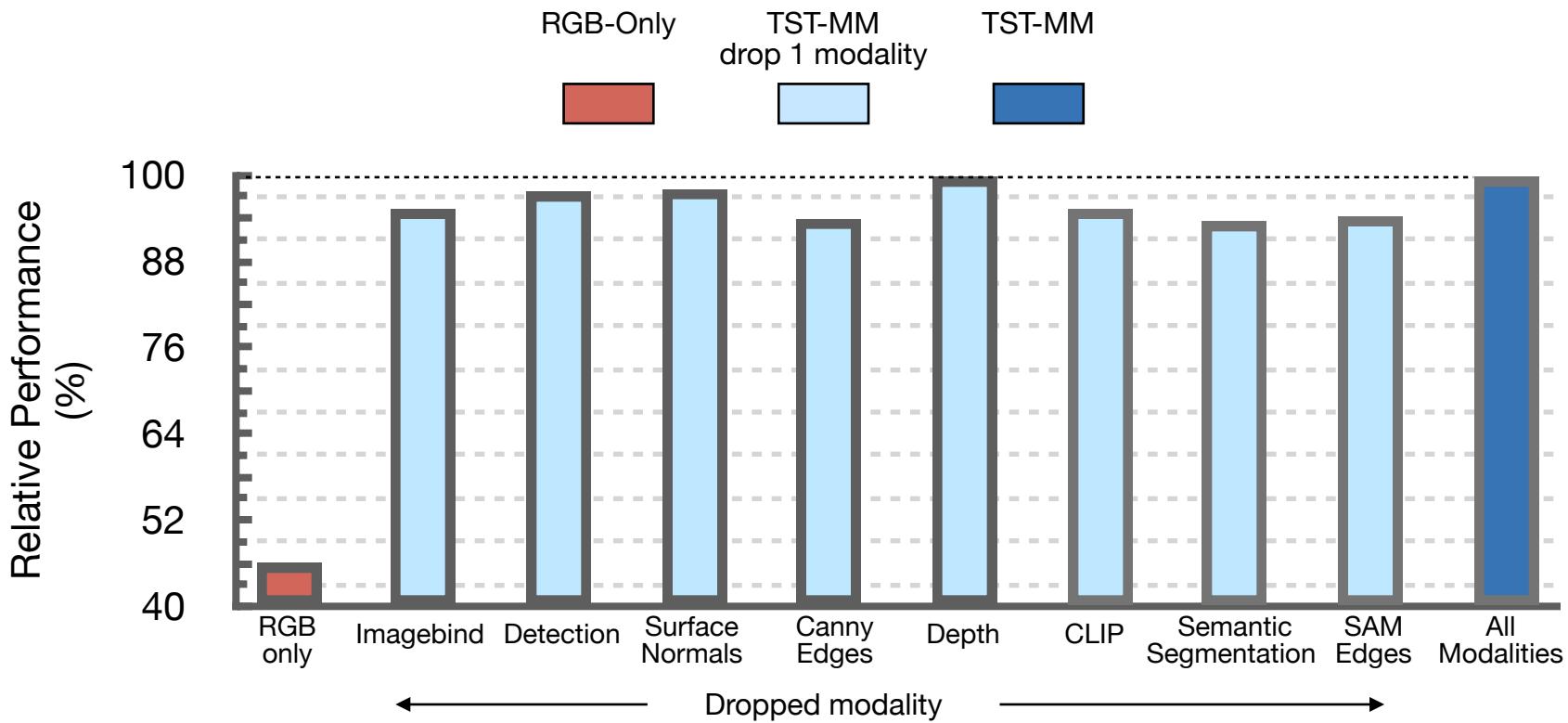
Analysis

Analysis 4. What about other self-supervised objectives?



Analysis

Analysis 5. Is one modality doing most of the job?



Test-Space Training

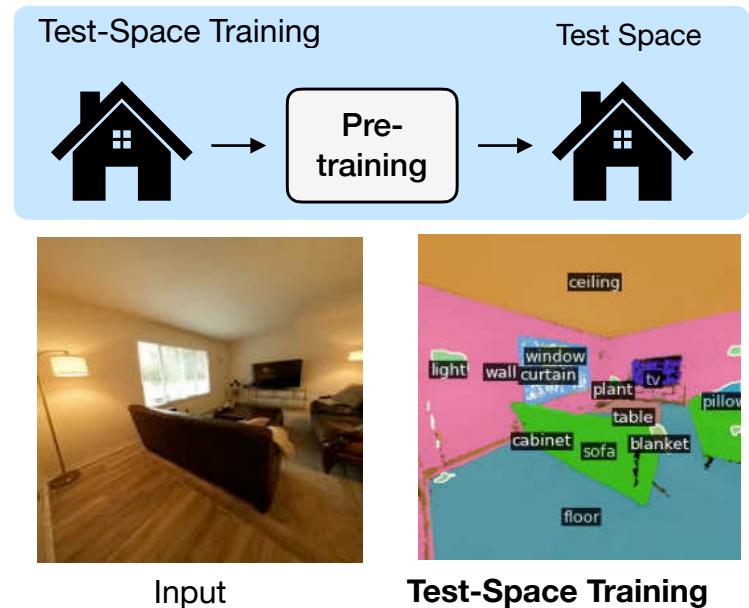
Self-Supervised Specialization of Vision Models to the Test Environment

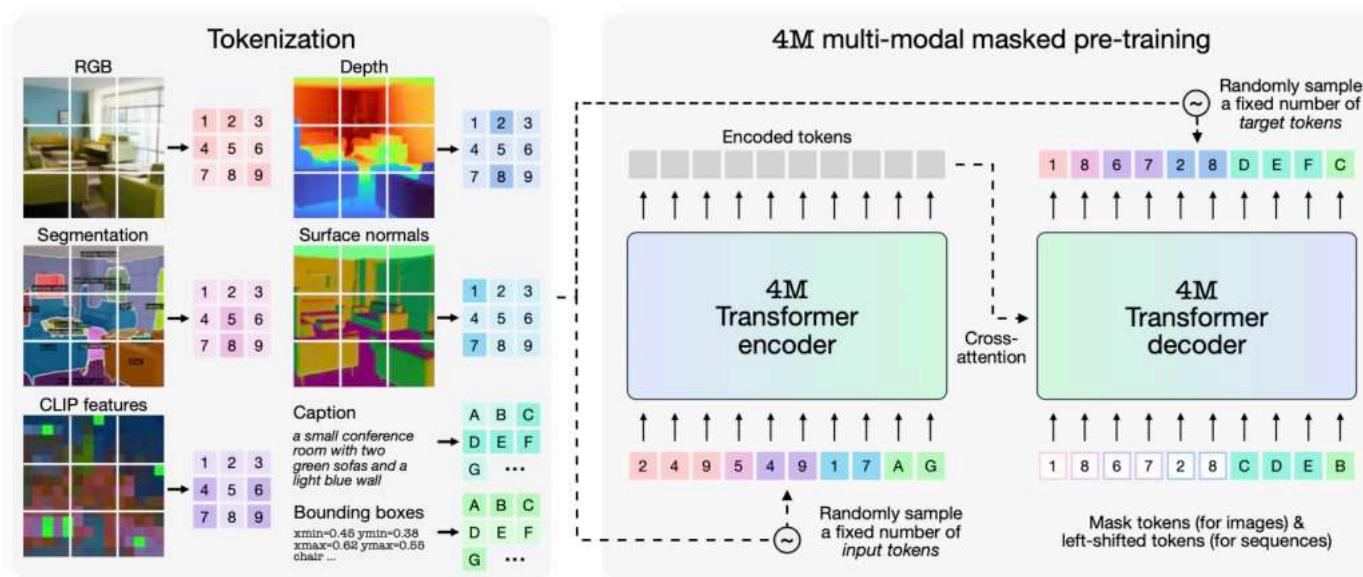
Kunal Pratap Singh*, Ali Garjani*,

Muhammad Uzair Khattak, Rishabh Singh, Jason Toskov,

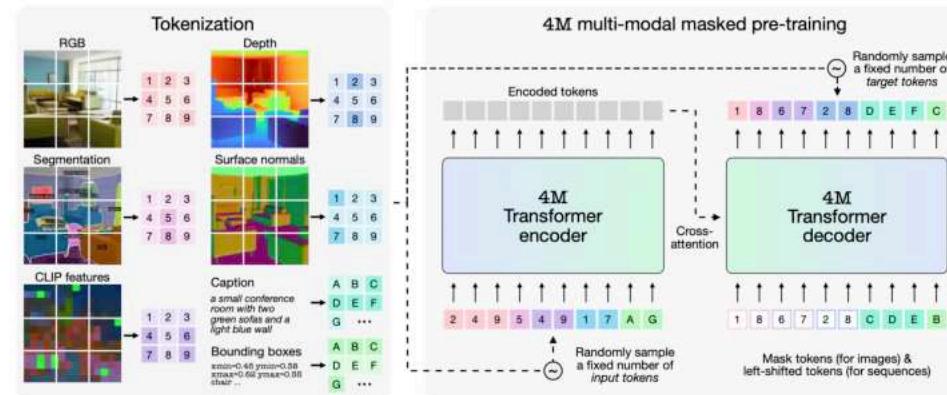
Andrei Atanov, Oğuzhan Fatih Kar, Amir Zamir

93

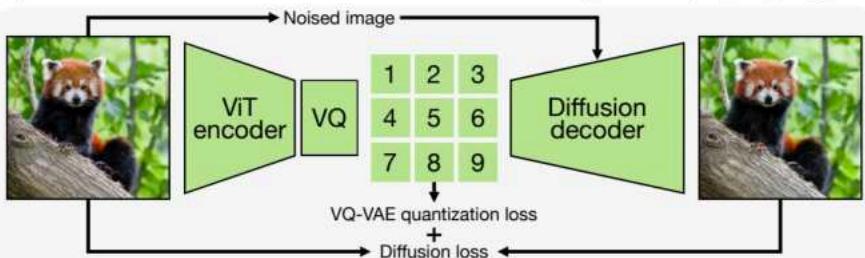




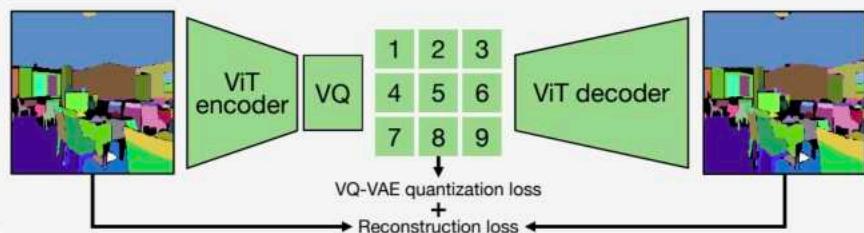
Tokenization



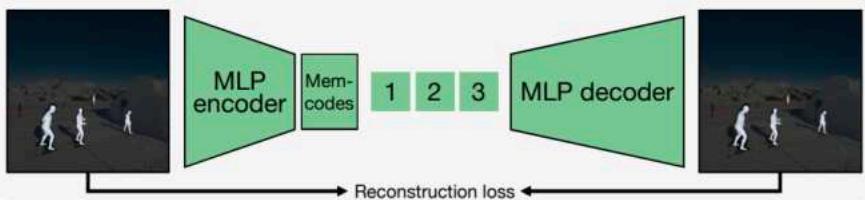
Spatial discrete VAE with diffusion decoder: RGB, normal, depth, edges



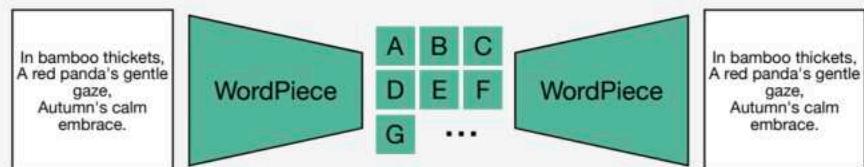
Spatial discrete VAE: Segmentation, CLIP, DINOv2, ImageBind, SAM inst.



MLP discrete VAE: Human poses, DINOv2 & ImageBind global tokens



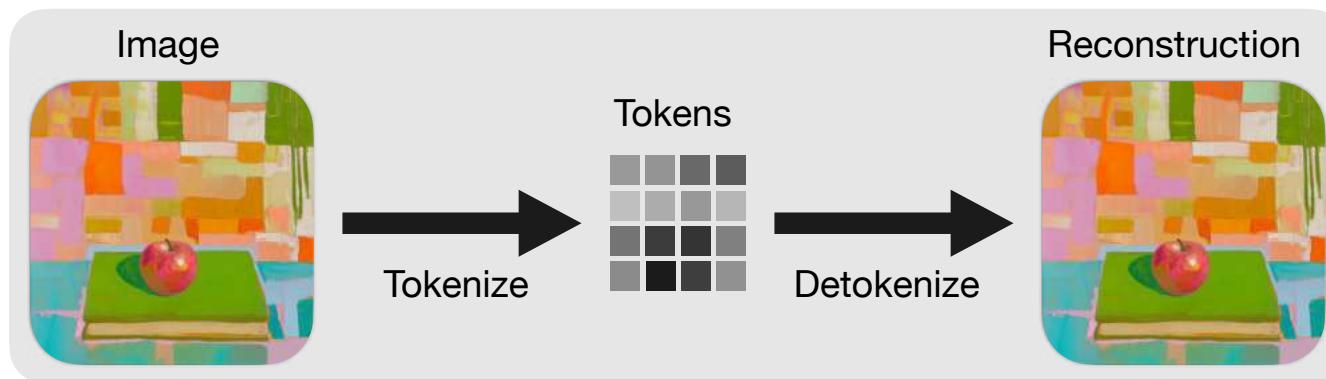
Sequence tokenizer: Text, bounding boxes, metadata, color palette



Token-based generation

Common way to perform generation:

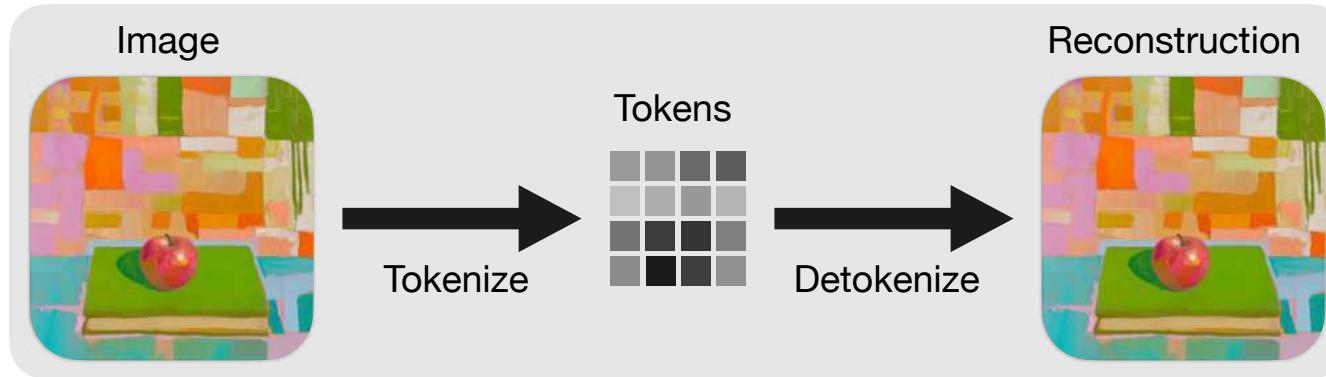
Stage 1: Train *tokenizer* with autoencoding objective



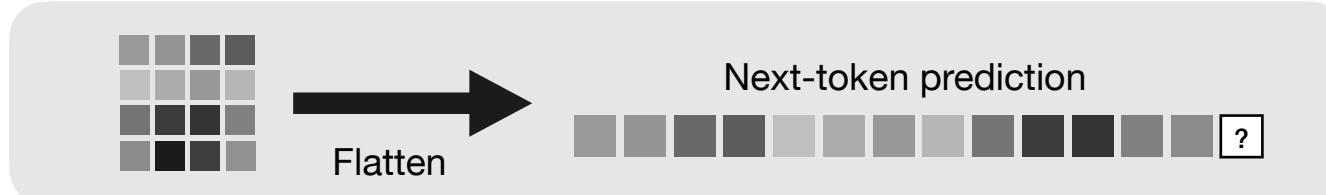
Token-based generation

Common way to perform generation:

Stage 1: Train *tokenizer* with autoencoding objective

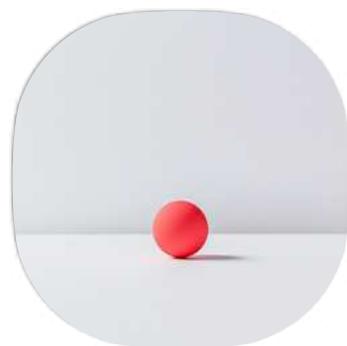


Stage 2: Perform *next-token prediction* on image tokens

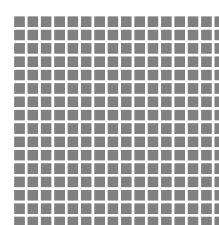
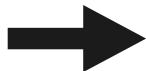


Token-based generation

Common 2D grid tokenizers images represented with **a fixed number of tokens, regardless of complexity.**



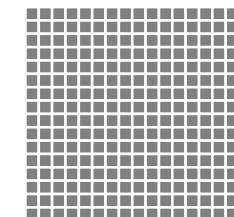
e.g. 256x256 pixels



e.g. 16x16 tokens



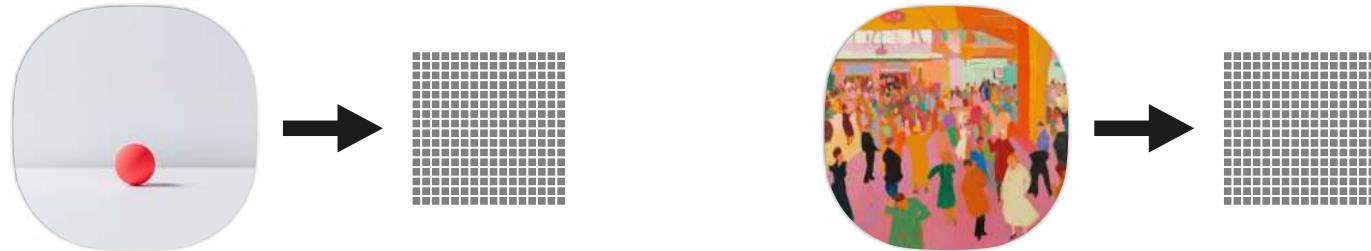
e.g. 256x256 pixels



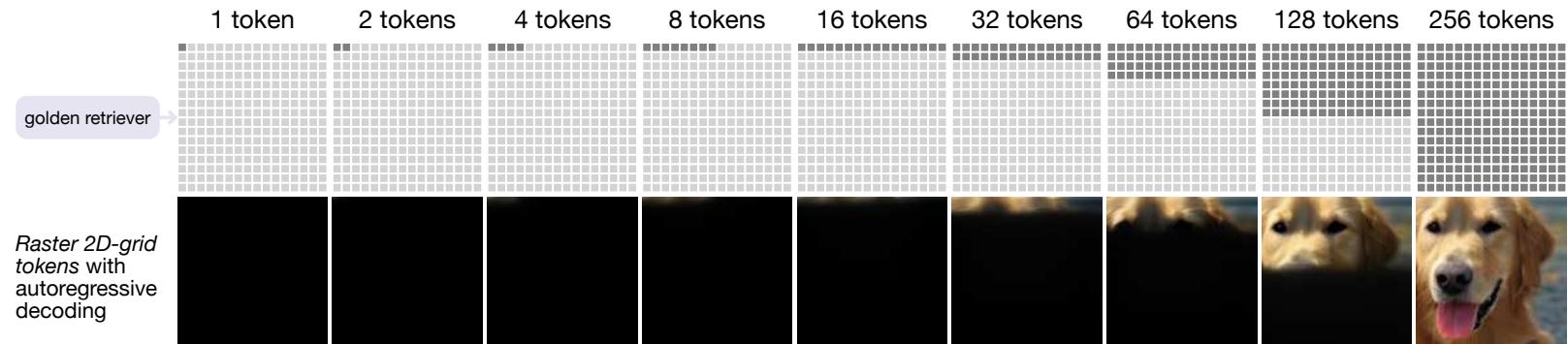
e.g. 16x16 tokens

Token-based generation

Common 2D grid tokenizers images represented with **a fixed number of tokens, regardless of complexity.**



Autoregressive generation is performed ~patch-by-patch.



Abstraction — Compression

Do we need to model **every detail, all the time?**



golden retriever

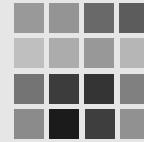
FlexTok

*Flexible-length
1D token
sequences with
autoregressive
decoding*



FlexTok overview

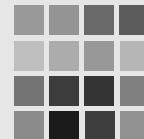
Classical 2D grid tokenizers



Tokens represent
local image
information

FlexTok overview

Classical 2D grid tokenizers



Tokens represent
local image
information

FlexTok 1D flexible length tokenizer



High-level
semantics

Details

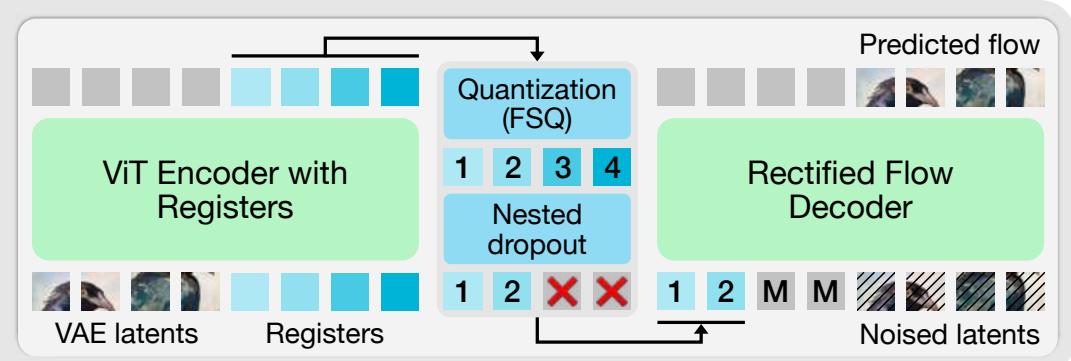


FlexTok method

Overview

Stage 1

FlexTok tokenizer training

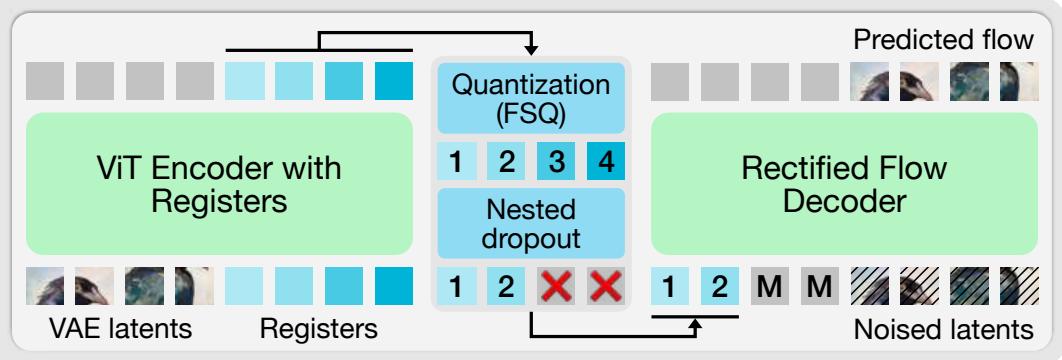


FlexTok method

Overview

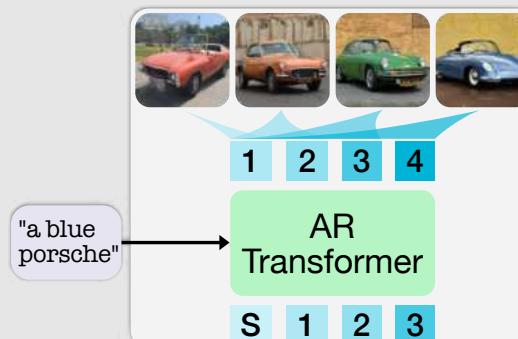
Stage 1

FlexTok tokenizer training



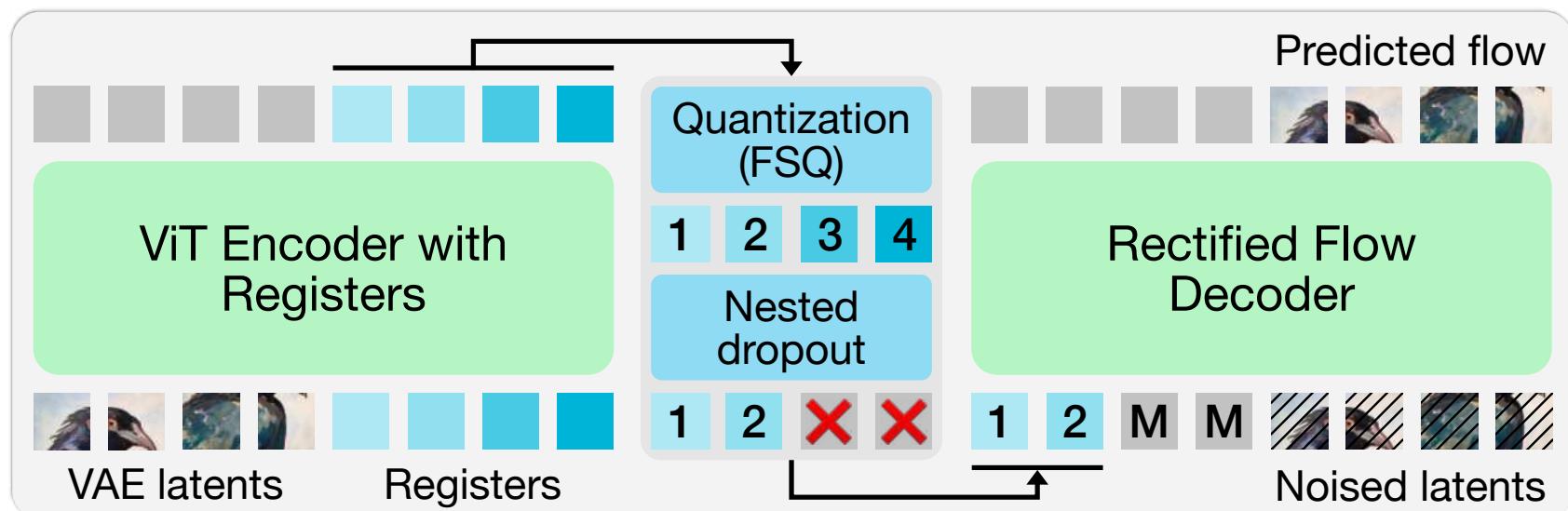
Stage 2

Autoregressive generation
using FlexTok tokens



FlexTok method

Stage 1: Tokenizer training



FlexTok reconstruction

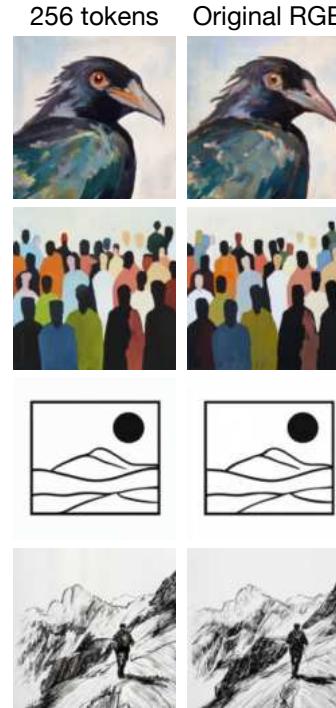
Specify a coarse-to-fine "visual vocabulary"

Original RGB



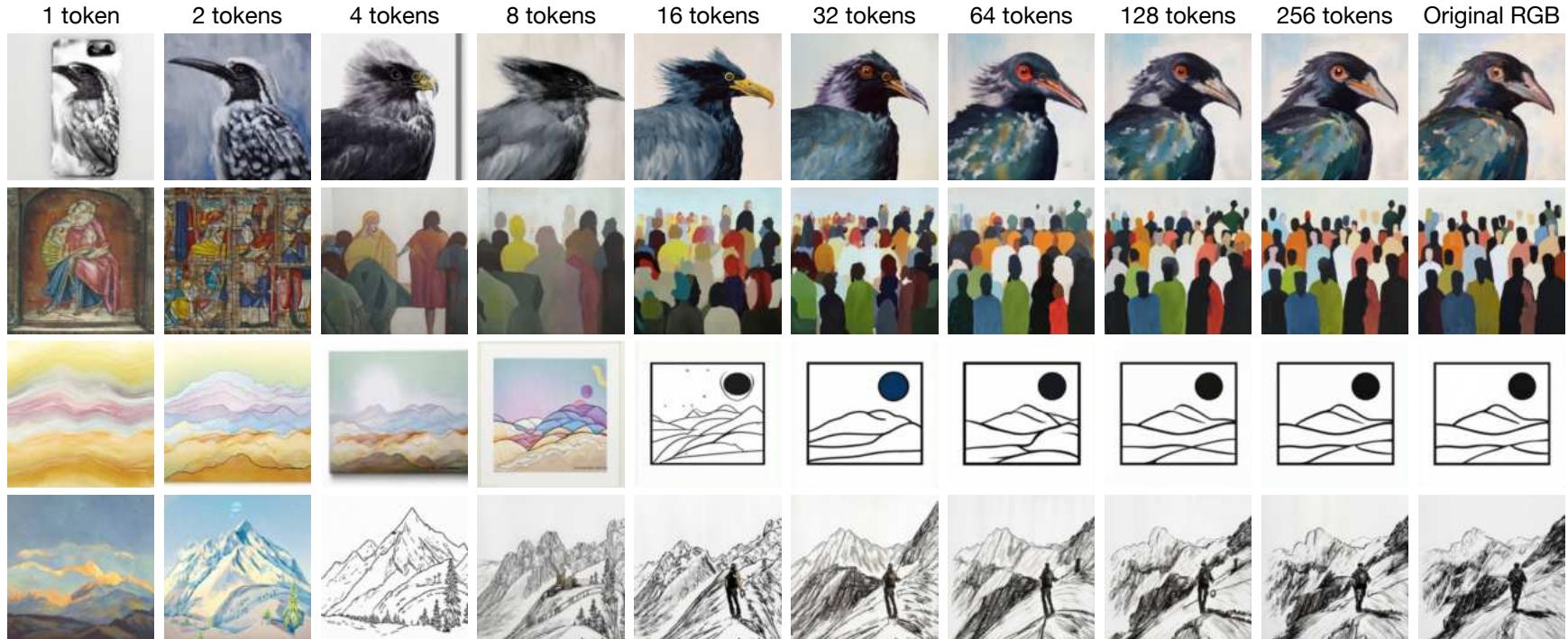
FlexTok reconstruction

Specify a coarse-to-fine "visual vocabulary"



FlexTok reconstruction

Specify a coarse-to-fine "visual vocabulary"



Autoregressive generation

Class-to-image



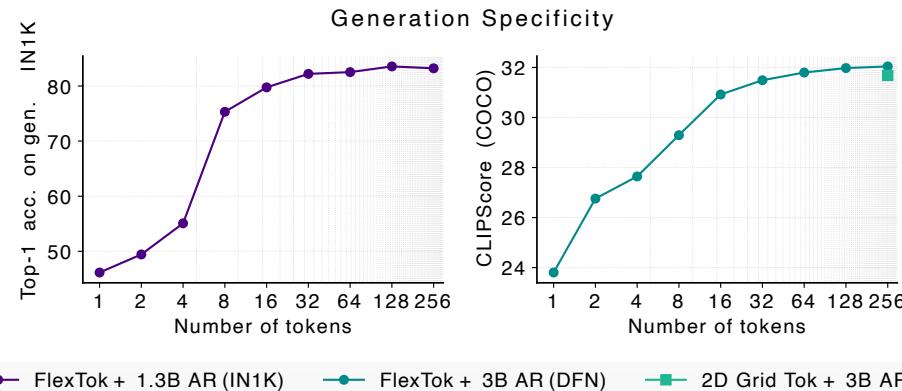
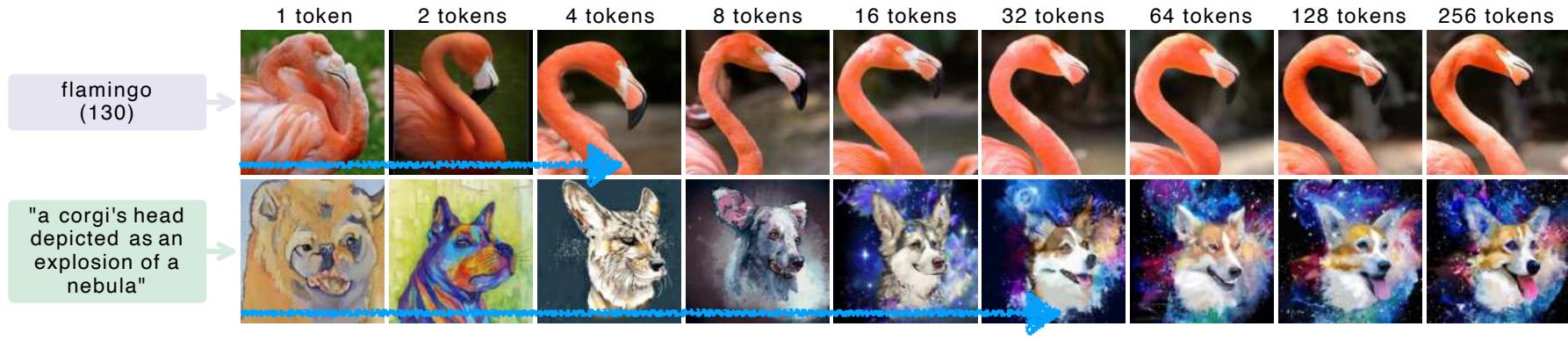
Autoregressive generation

Text-to-image



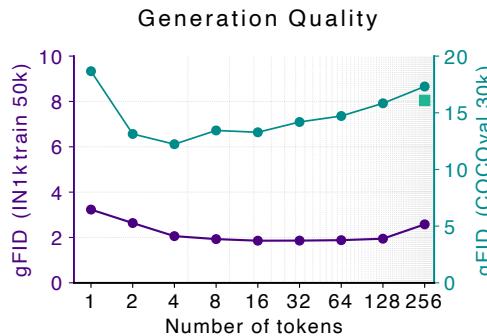
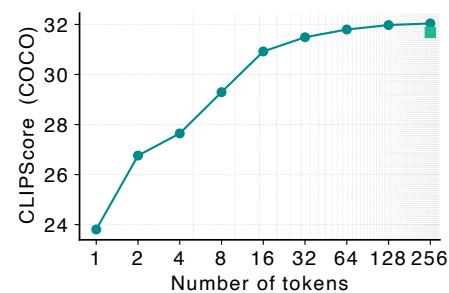
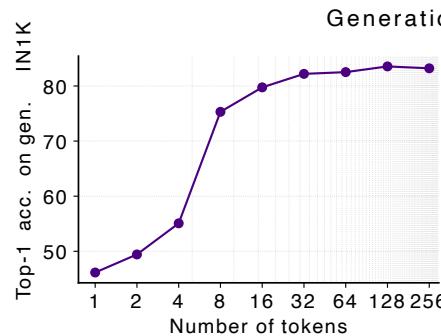
Autoregressive generation

Adaptive conditioning alignment



Autoregressive generation

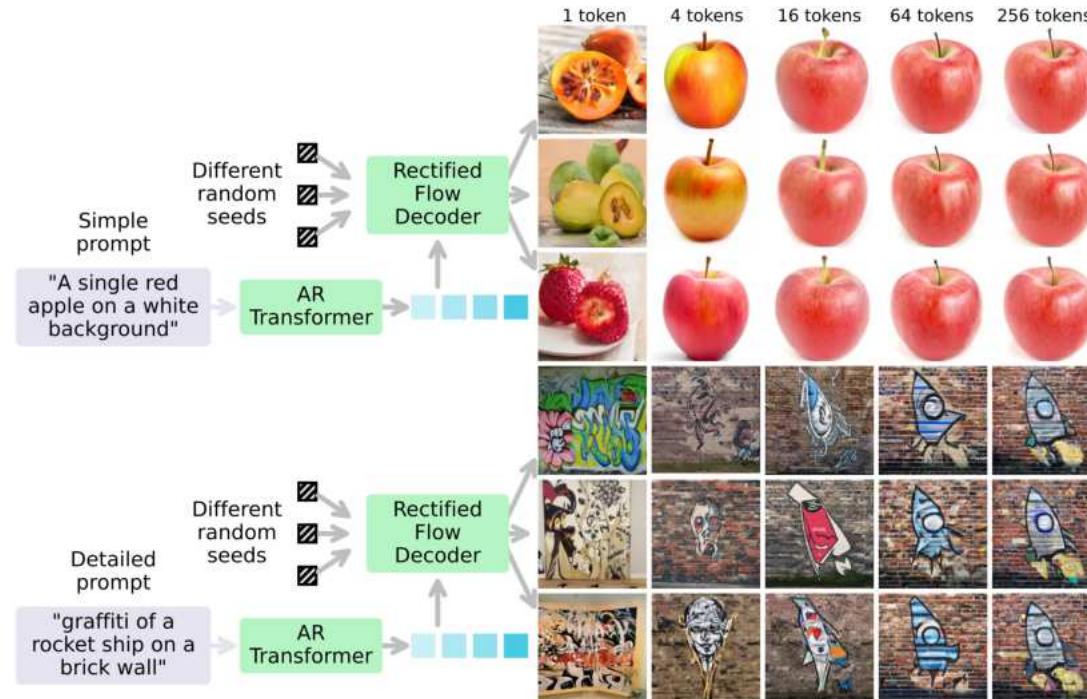
Adaptive conditioning alignment



● FlexTok + 1.3B AR (IN1K) ● FlexTok + 3B AR (DFN) ■ 2D Grid Tok + 3B AR (DFN)

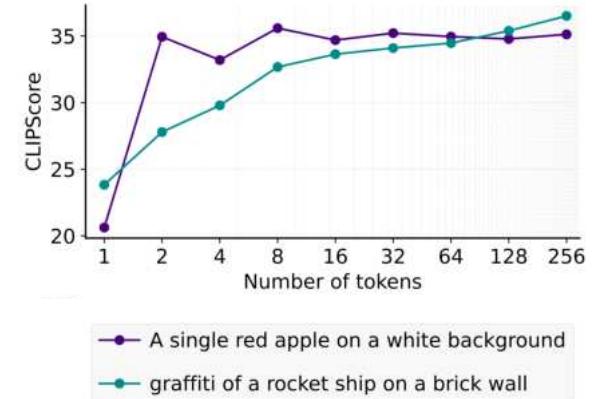
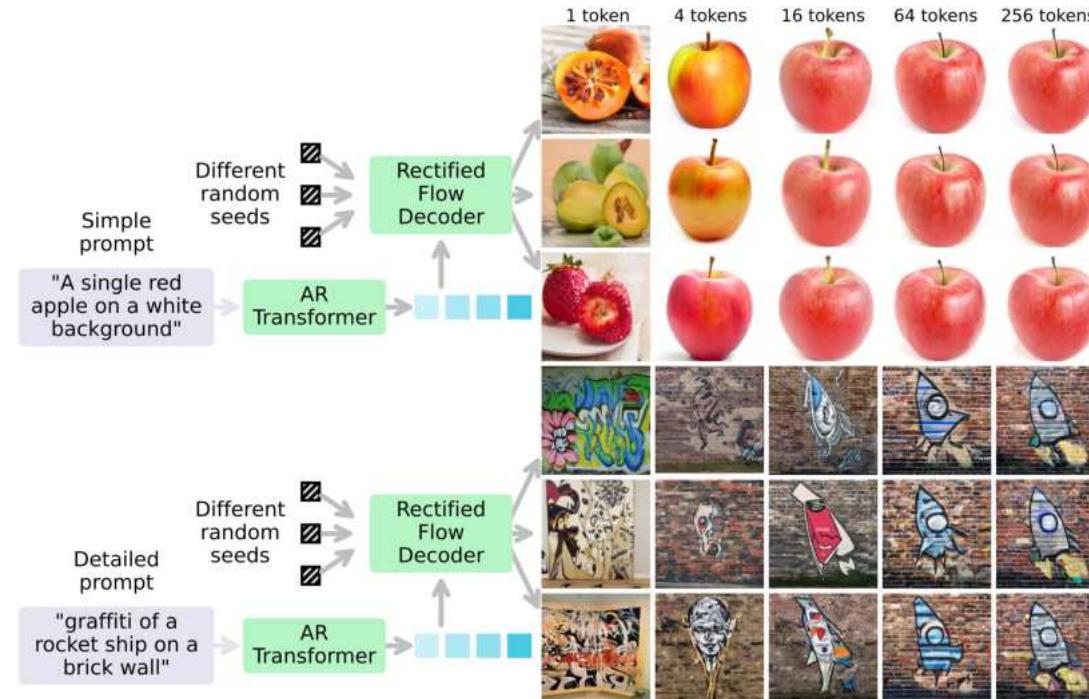
Autoregressive generation

Image generation with simple and complex prompts



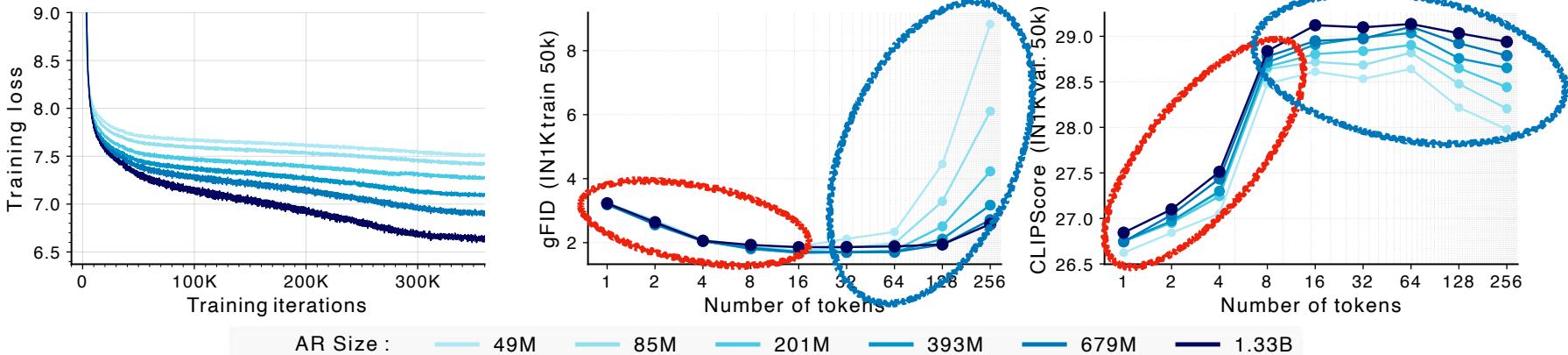
Autoregressive generation

Image generation with simple and complex prompts



Scaling autoregressive generation

- Prediction quality for first ~8 tokens is independent of model size
- Scaling AR model improves quality and alignment when predicting >32 tokens





Resampling Images into 1D Token Sequences of Flexible Length

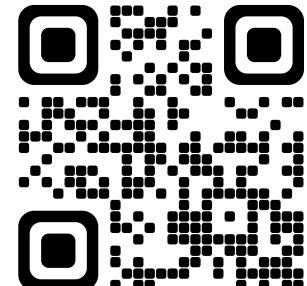
Roman Bachmann^{1,2*}, Jesse Allardice^{1*}, David Mizrahi^{1*}, Enrico Fini¹, Oğuzhan Fatih Kar², Elmira Amirloo¹, Alaaeldin El-Nouby¹, Amir Zamir², Afshin Dehghan¹



¹Apple, ²EPFL
*Equal contribution

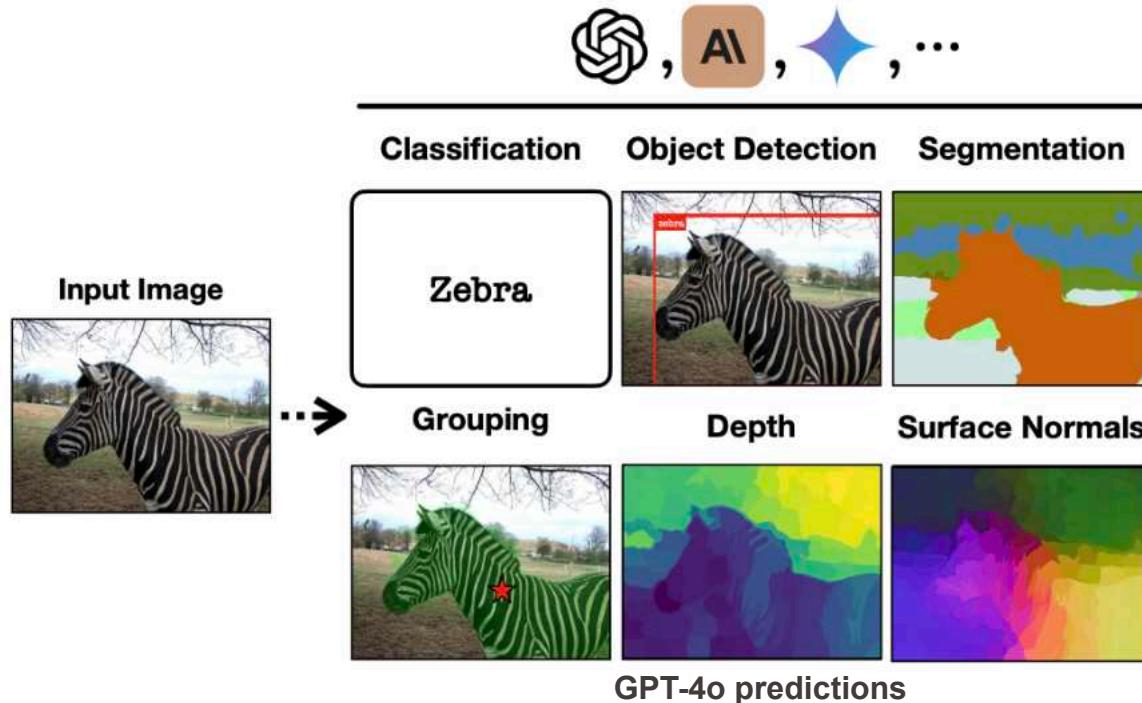
Summary:

- FlexTok: 1D flexible-length tokenizer
- Specify "visual vocabulary", enabling coarse-to-fine generation
- Adaptive: Number of tokens to predict depends on complexity
- Future directions into flexible temporal and multimodal compression, as well as long-term reasoning in abstract spaces



Benchmarking popular multimodal FMs

How well does GPT-4o understand vision?

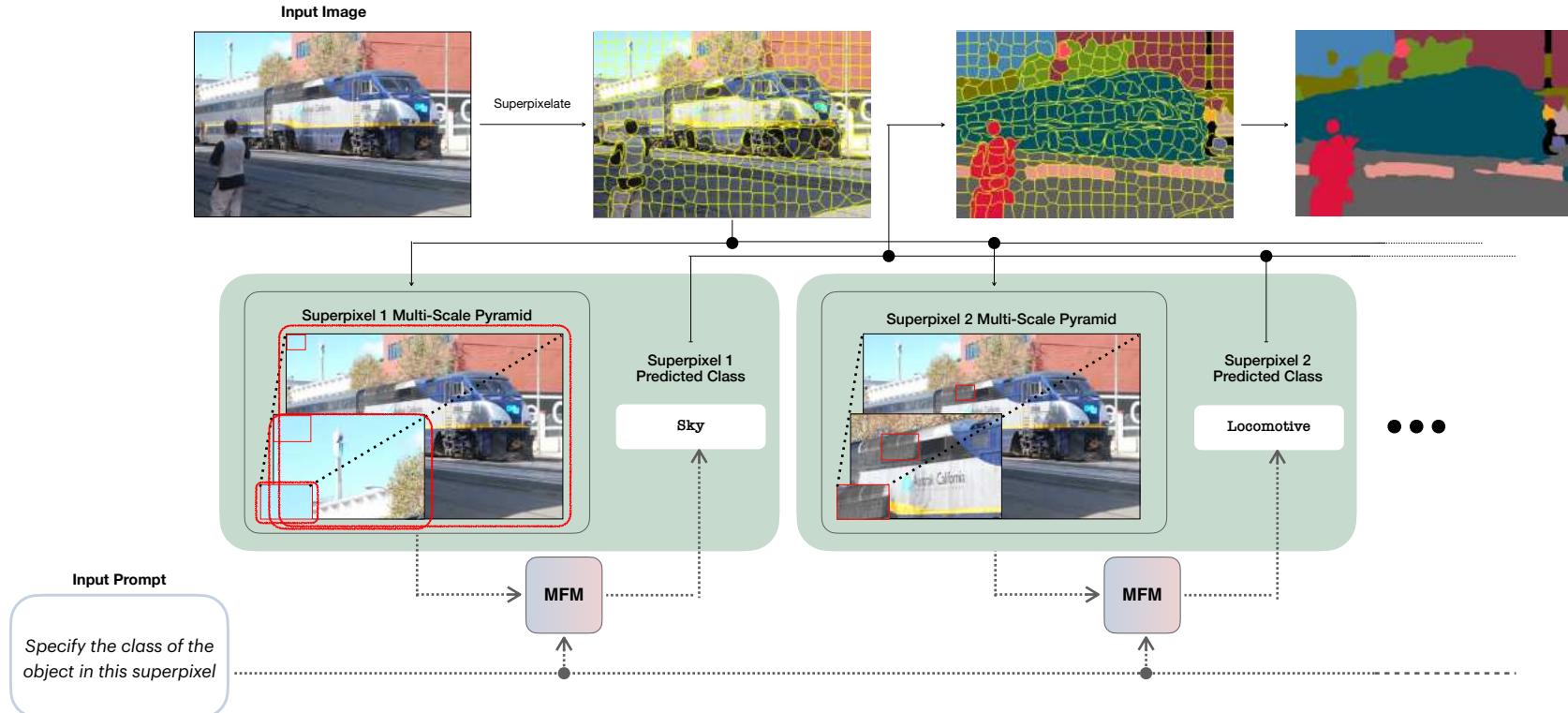


<https://fm-vision-evals.epfl.ch/>

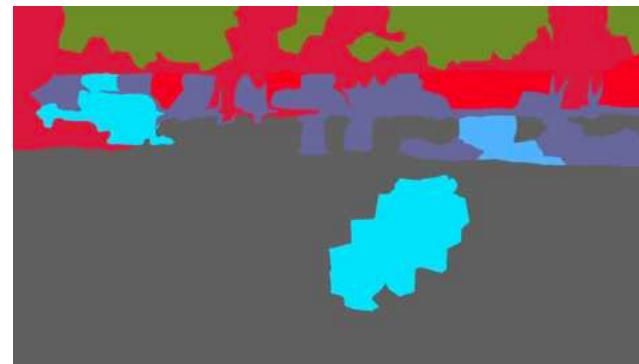
How Well Does GPT-4o Understand Vision? Evaluating Multimodal Foundation Models on Standard Computer Vision Tasks, Ramachandran, Garjani, Bachmann, Atanov *, Kar *, Zamir *. arxiv 2025.

How to extract a non-textual task from chatbots?

e.g., semantic segmentation from chatGPT?

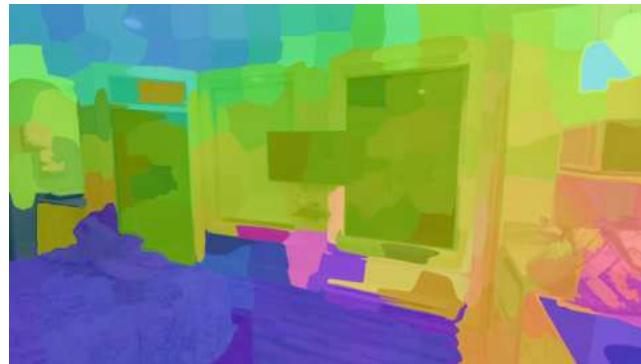


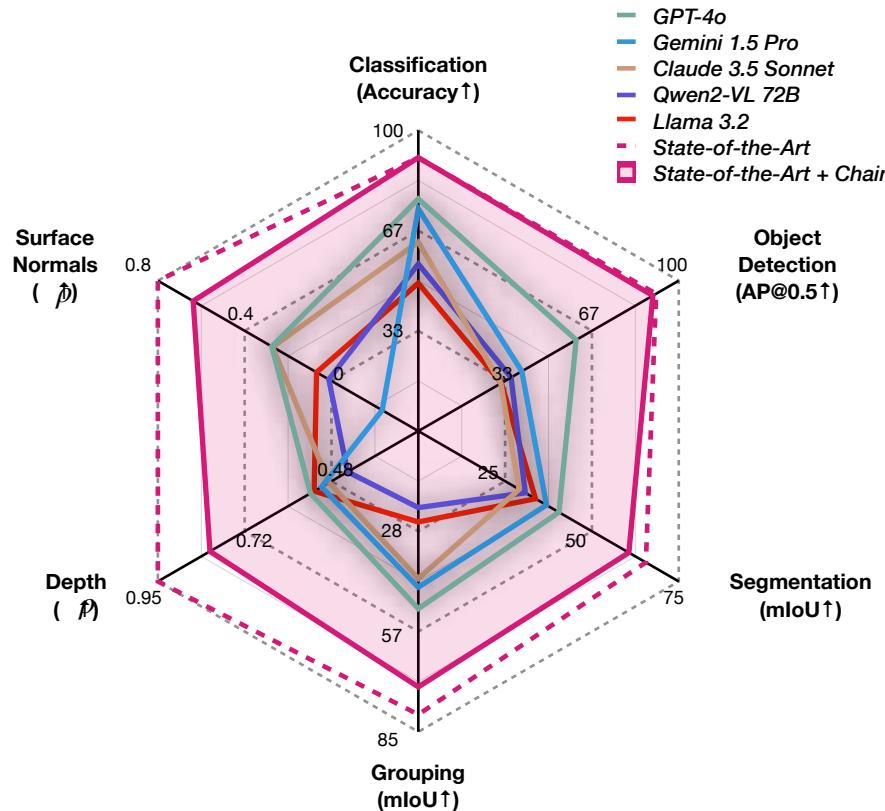
Predictions (GPT-4o)



Predictions

(GPT-4o)



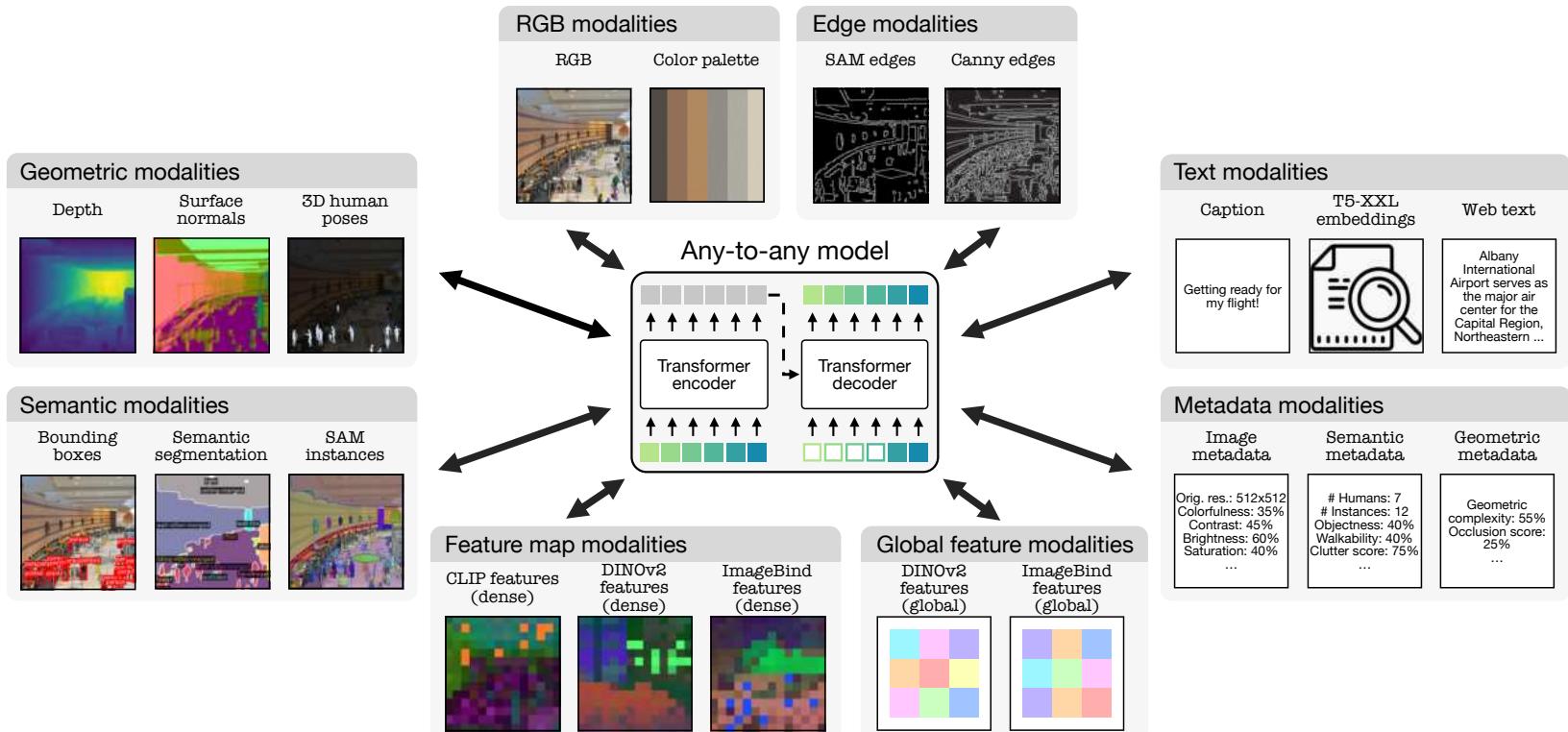


Key Takeaways

- **Not state-of-the-art** but **respectable generalists**.
- **Stronger at semantic tasks** than geometric tasks.
- **GPT-4o outperforms** other models across most tasks.

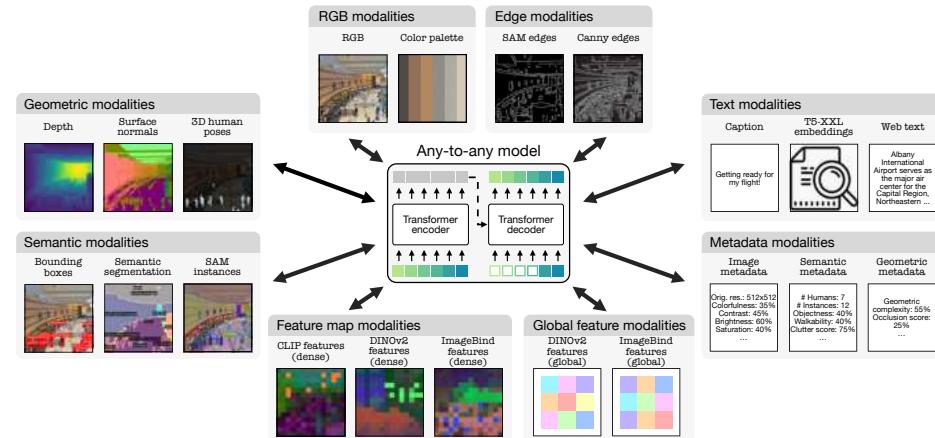
<https://fm-vision-evals.epfl.ch/>

How Well Does GPT-4o Understand Vision? Evaluating Multimodal Foundation Models on Standard Computer Vision Tasks, Ramachandran, Garjani, Bachmann, Atanov *, Kar *, Zamir *. arxiv 2025.



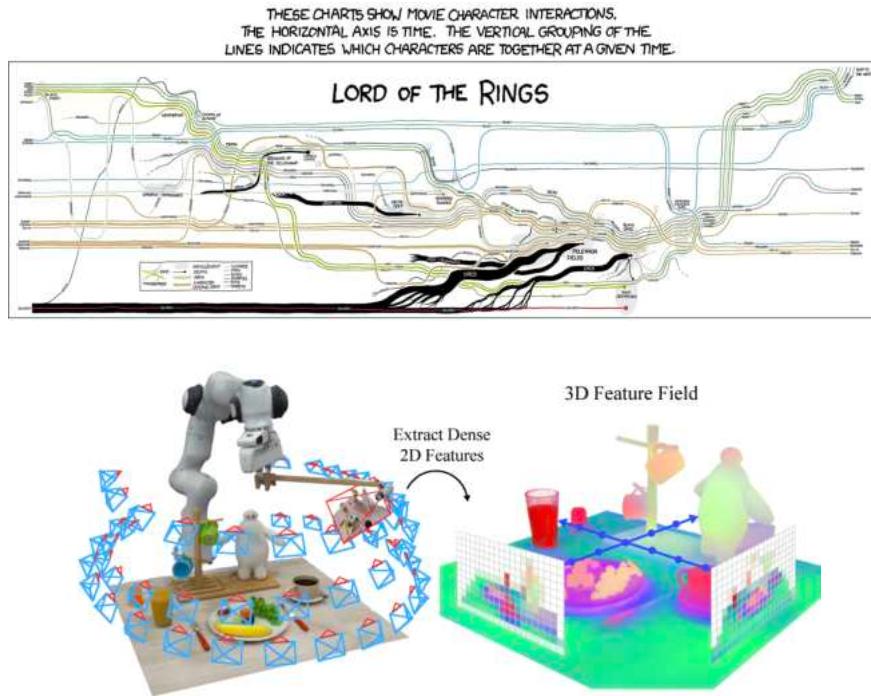
- 4M: Massively Multimodal Masked Modeling, Mizrahi, Bachmann, Kar, Yeo, Gao, Dehghan, Zamir. NeurIPS 2023.
- MultiMAE: Multi-Modal Multi-Task Masked Autoencoders, Bachmann, Mizrahi, Atanov, Zamir. ECCV 2022
- 4M-21: An Any-to-Any Vision Model for Tens of Tasks and Modalities, Bachmann, Kar, Mizrahi, et al., 2024.

- A scalable versatile multi-modal/Multi-task foundation model
- Ultimate goal: a grounded world model. A “foundation”.



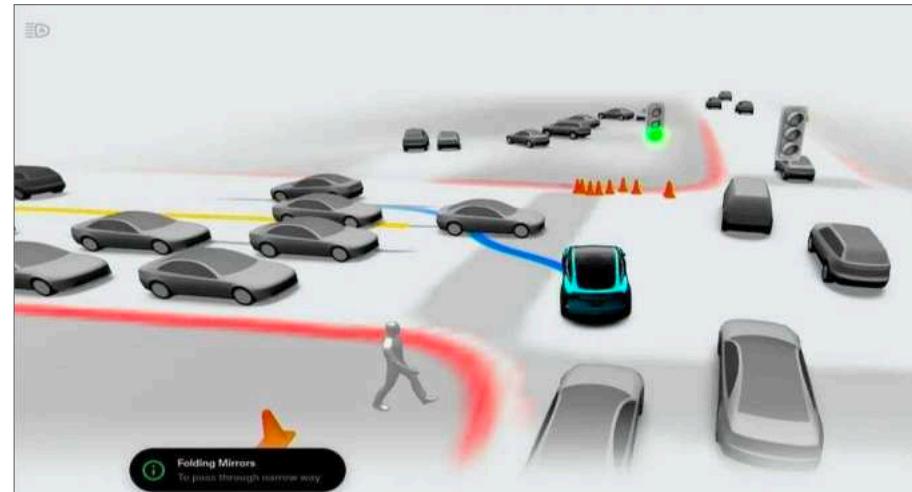
- 4M: Massively Multimodal Masked Modeling, Mizrahi, Bachmann, Kar, Yeo, Gao, Dehghan, Zamir. NeurIPS 2023.
- MultiMAE: Multi-Modal Multi-Task Masked Autoencoders, Bachmann, Mizrahi, Atanov, Zamir. ECCV 2022
- Craik, Kenneth. The nature of explanation. Vol. 445. CUP Archive, 1967.

- A scalable versatile multi-modal/Multi-task foundation model
- Ultimate goal: a grounded world model. A “foundation”.
- (Long-form) Video understanding



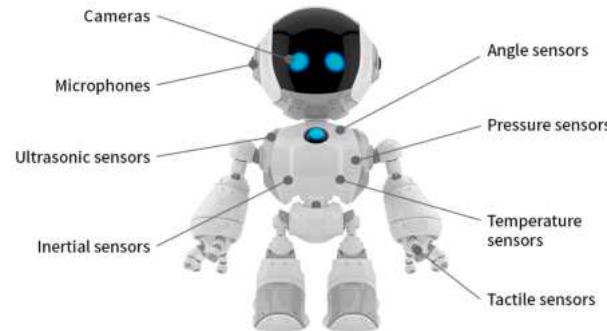
- 4M: Massively Multimodal Masked Modeling, Mizrahi, Bachmann, Kar, Yeo, Gao, Dehghan, Zamir. NeurIPS 2023.
- MultiMAE: Multi-Modal Multi-Task Masked Autoencoders, Bachmann, Mizrahi, Atanov, Zamir. ECCV 2022
- 4M-21: An Any-to-Any Vision Model for Tens of Tasks and Modalities, Bachmann, Kar, Mizrahi, et al., 2024.

- A scalable versatile **multi-modal/Multi-task foundation model**
- Ultimate goal: **a grounded world model**. A “foundation”.
- (Long-form) Video understanding
- Learning in **higher-level spaces**



- - 4M: Massively Multimodal Masked Modeling, Mizrahi, Bachmann, Kar, Yeo, Gao, Dehghan, Zamir. NeurIPS 2023.
 - MultiMAE: Multi-Modal Multi-Task Masked Autoencoders, Bachmann, Mizrahi, Atanov, Zamir. ECCV 2022
 - 4M-21: An Any-to-Any Vision Model for Tens of Tasks and Modalities, Bachmann, Kar, Mizrahi, et al., 2024.

- A scalable versatile **multi-modal/Multi-task foundation model**
- Ultimate goal: **a grounded world model. A “foundation”.**
- (Long-form) Video understanding
- Learning in higher-level spaces
- **Physical/MM self-supervision**



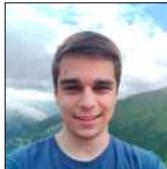
- 4M: Massively Multimodal Masked Modeling, Mizrahi, Bachmann, Kar, Yeo, Gao, Dehghan, Zamir. NeurIPS 2023.
- MultiMAE: Multi-Modal Multi-Task Masked Autoencoders, Bachmann, Mizrahi, Atanov, Zamir. ECCV 2022
- 4M-21: An Any-to-Any Vision Model for Tens of Tasks and Modalities, Bachmann, Kar, Mizrahi, et al., 2024.

- A scalable versatile **multi-modal/Multi-task foundation model**
- Ultimate goal: **a grounded world model. A “foundation”.**
- (Long-form) Video understanding
- Learning in higher-level spaces
- Physical/MM self-supervision
- Multimodal in-context learning
- Reasoning
- Co-training
- Inducing emergence

- 4M: Massively Multimodal Masked Modeling, Mizrahi, Bachmann, Kar, Yeo, Gao, Dehghan, Zamir. NeurIPS 2023.
- MultiMAE: Multi-Modal Multi-Task Masked Autoencoders, Bachmann, Mizrahi, Atanov, Zamir. ECCV 2022
- 4M-21: An Any-to-Any Vision Model for Tens of Tasks and Modalities, Bachmann, Kar, Mizrahi, et al., 2024.



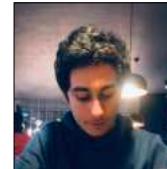
Roman
Bachmann



David
Mizrahi



Oguzhan
Kar



Ali
Garjani



Mingfei
Gao



David
Griffiths



Sogand
Salehi



Andrei
Atanov



Jiawei Fu



Rishabh
Singh



Isabella
Yu



Andrew
Spielberg



Jiming Hu



Teresa
Yeo



Afshin
Dehghan



Amir
Zamir

Multimodal Learning

<https://4m.epfl.ch/>

<https://flextok.epfl.ch/>

<https://fm-vision-evals.epfl.ch/>

<https://amirzamir.com/>

EPFL