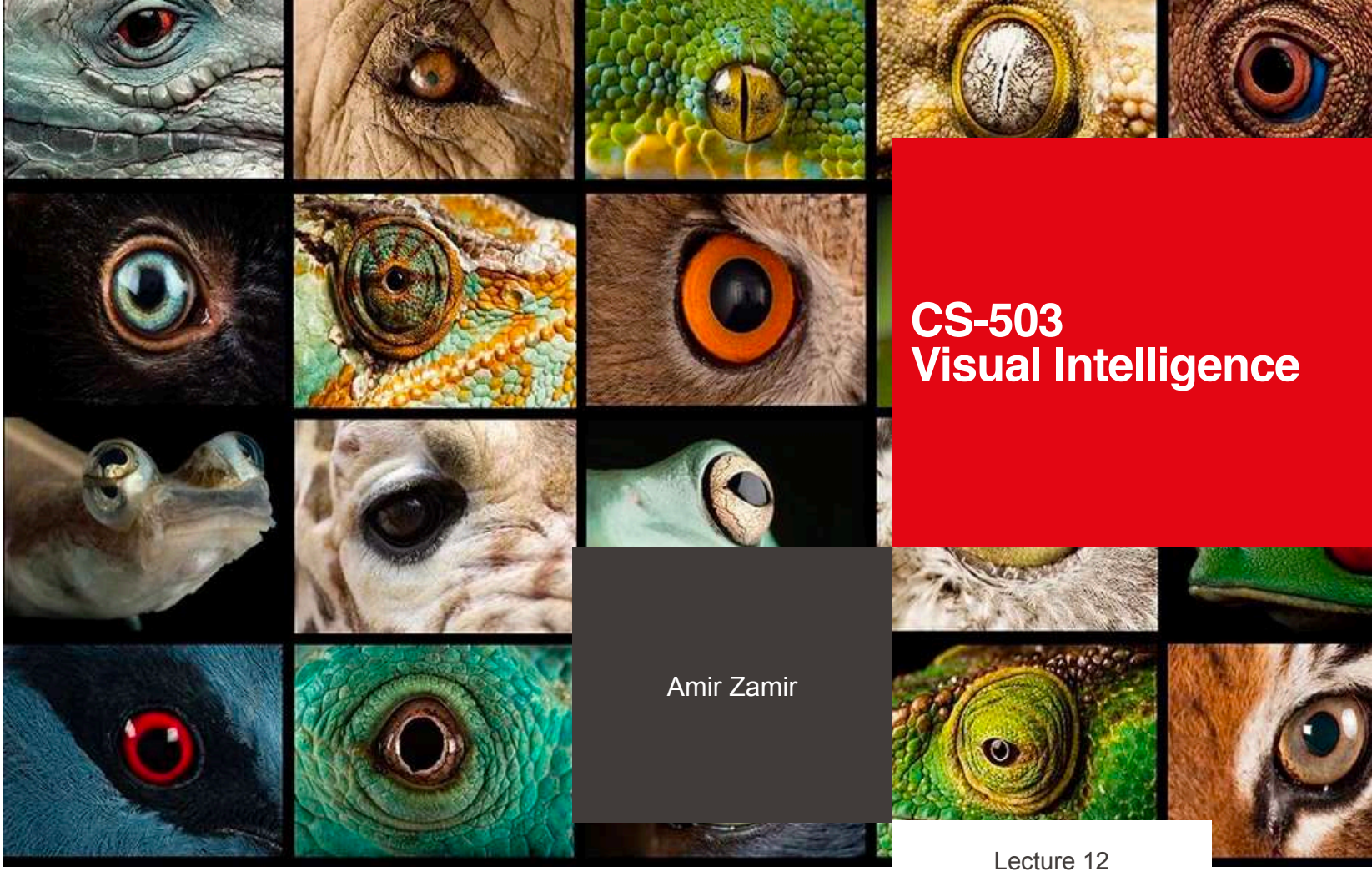


CS-503
Visual Intelligence

Amir Zamir

Lecture 12

École
polytechnique
fédérale
de Lausanne

Logistics

Logistics

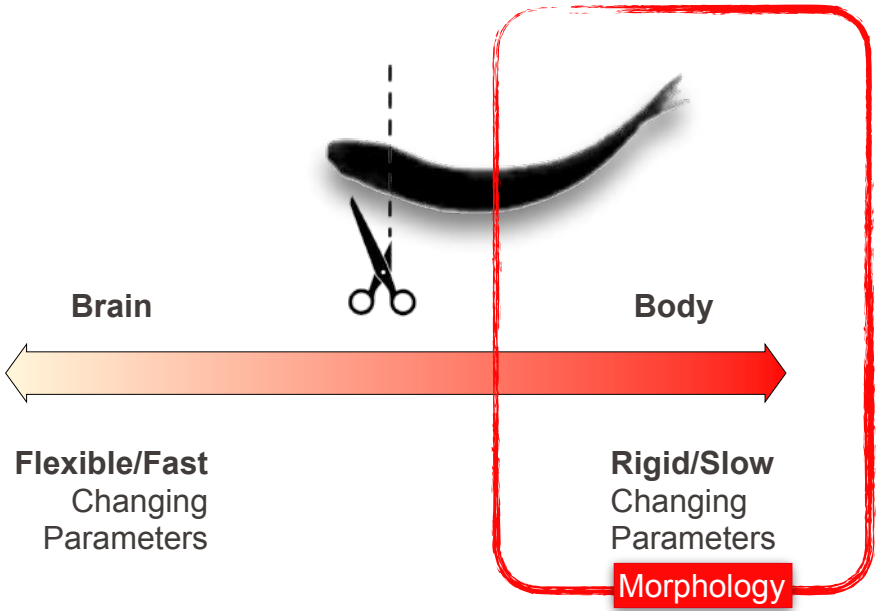
Week Num.	Date	Item
1	20.02	- lecture 1
2a	25.02	- lecture 2
2b	27.02	- lecture 3
3a	04.03	- lecture 4
3b	06.03	- lecture 5
4a	11.03	- lecture 6 (+ Q&A)
	11.03	- Transformers notebook assignment due
4b	13.03	- lecture 7
5a	18.03	- lecture 8
5b	20.03	- lecture 9
6a	25.03	- lecture 10
6b	27.03	- lecture 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

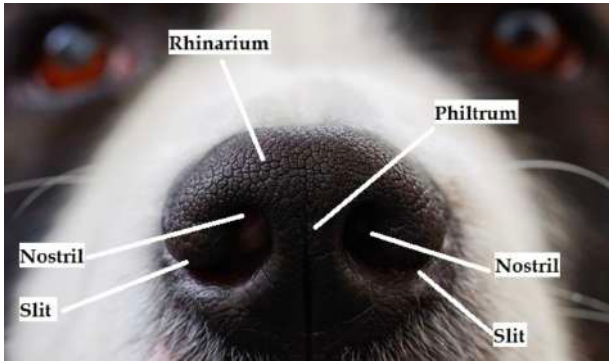
Dead Fish Swimming

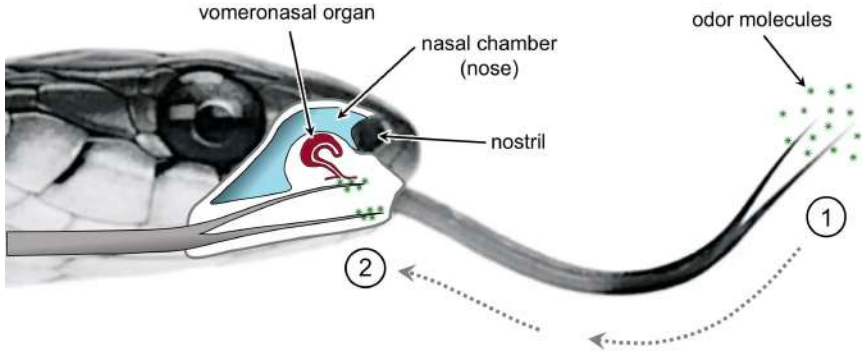
Zamir

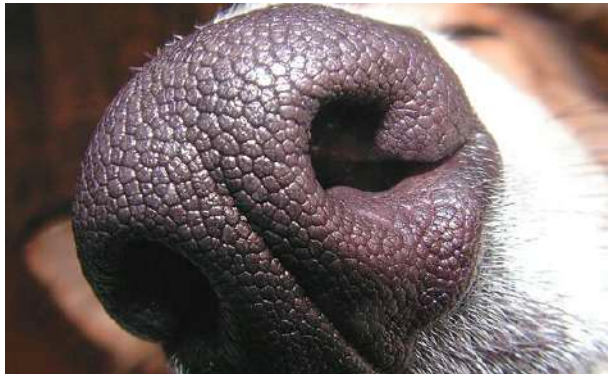
- Is the fish intelligent?
- Where is the intelligence?



Other modalities: olfactory







Multimodality

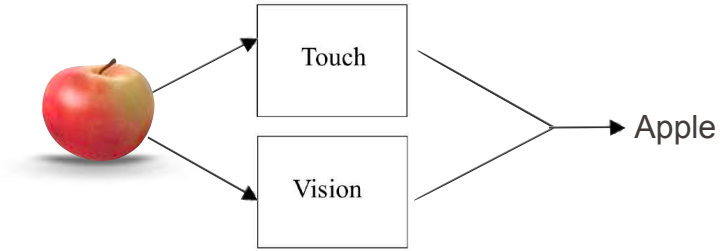
Vision: “**va**”
Audio: “**ba**”

Vision: “**ba**”
Audio: “**ba**”

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

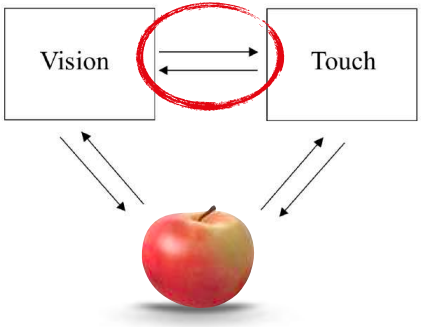
Roles of Multimodality in Learning

For sensory fusion / better inference



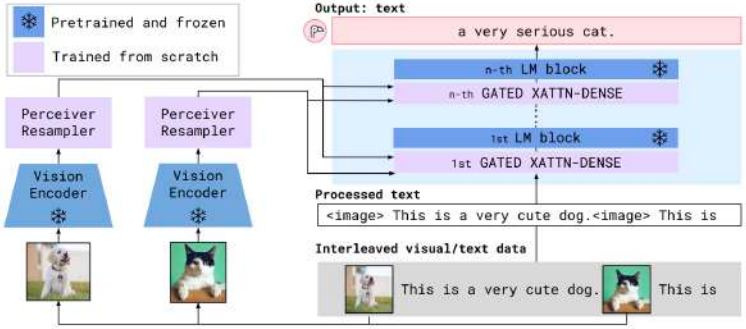
For self-supervision

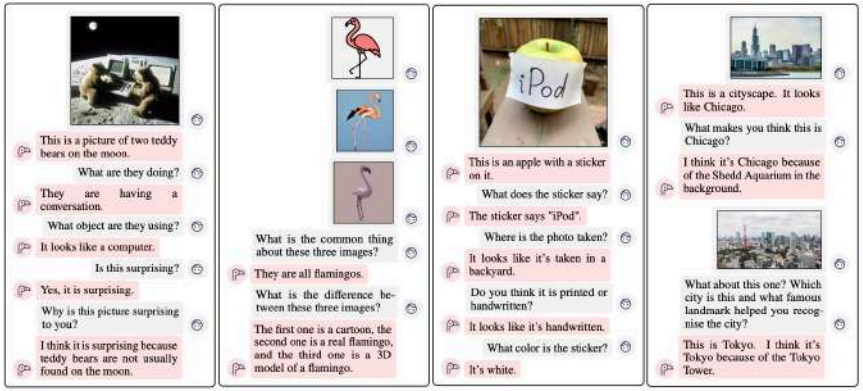
Cross-Modal Learning

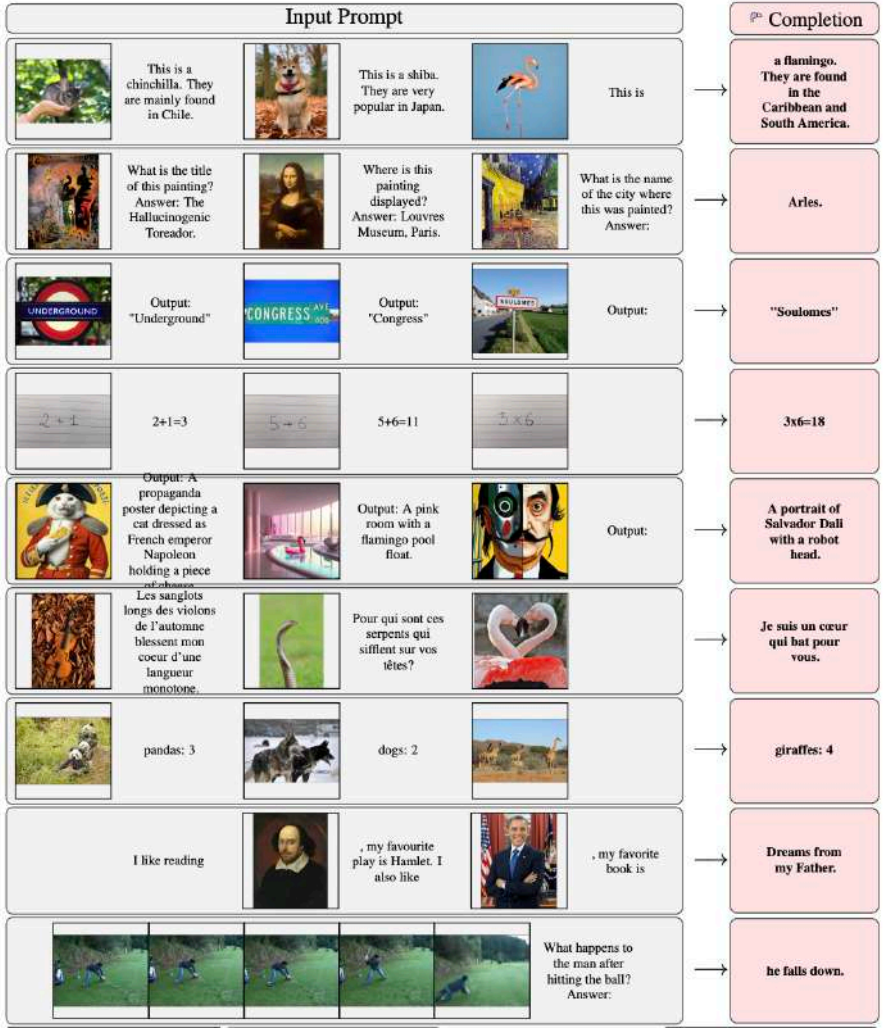


- "Six Lessons from Babies", Smith&Gasser'05:
  - *Multiple overlapping and time-locked sensory systems enable the developing system to* ***educate ("supervise") itself****.*

VLM (Vision-Language Model)
~ (RGB-Text chatbot)



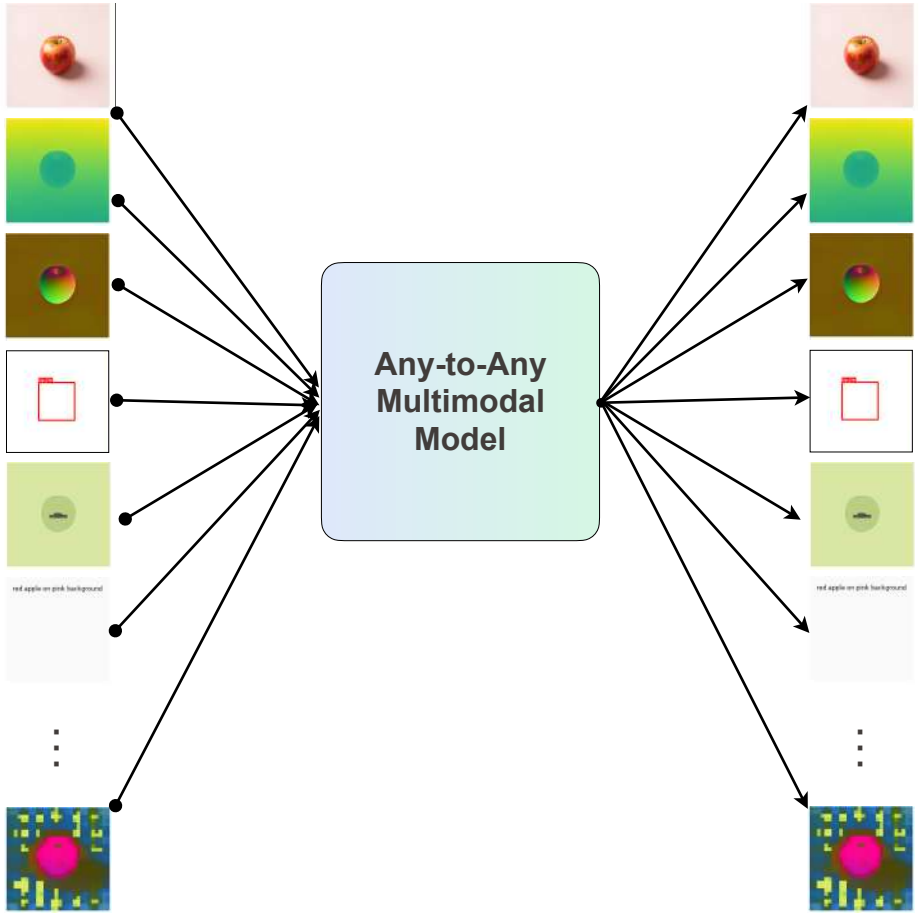




- Flamingo,, Alayrac et al, 2022.

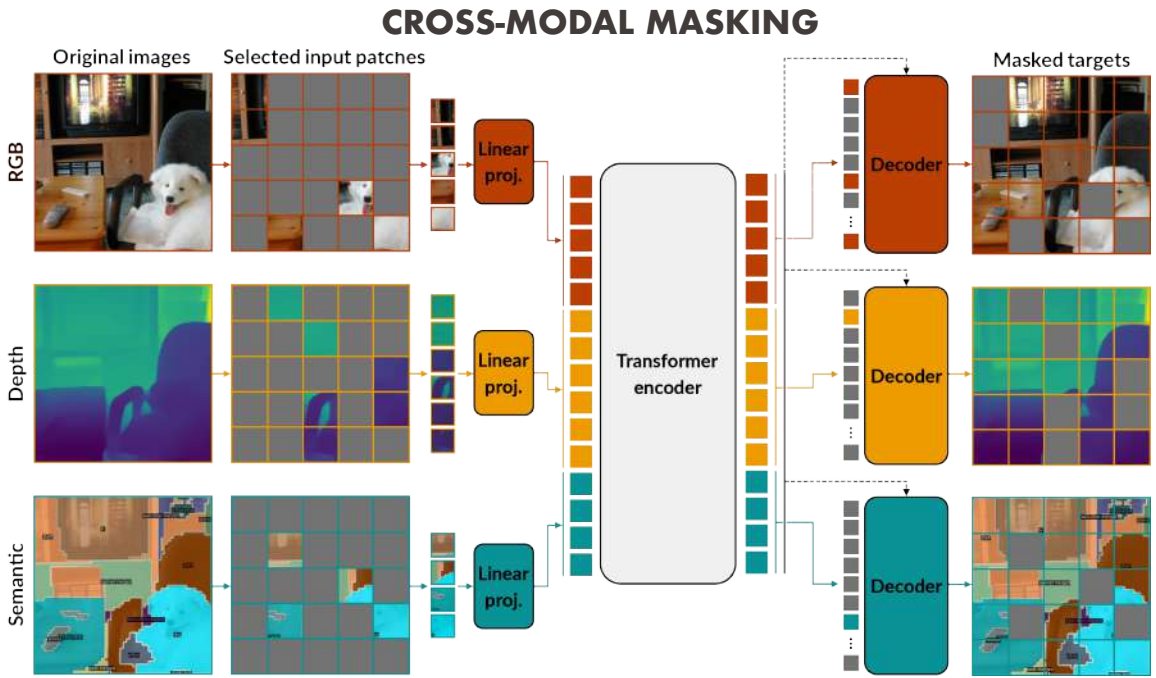
Core function: Predict anything from anything

The world

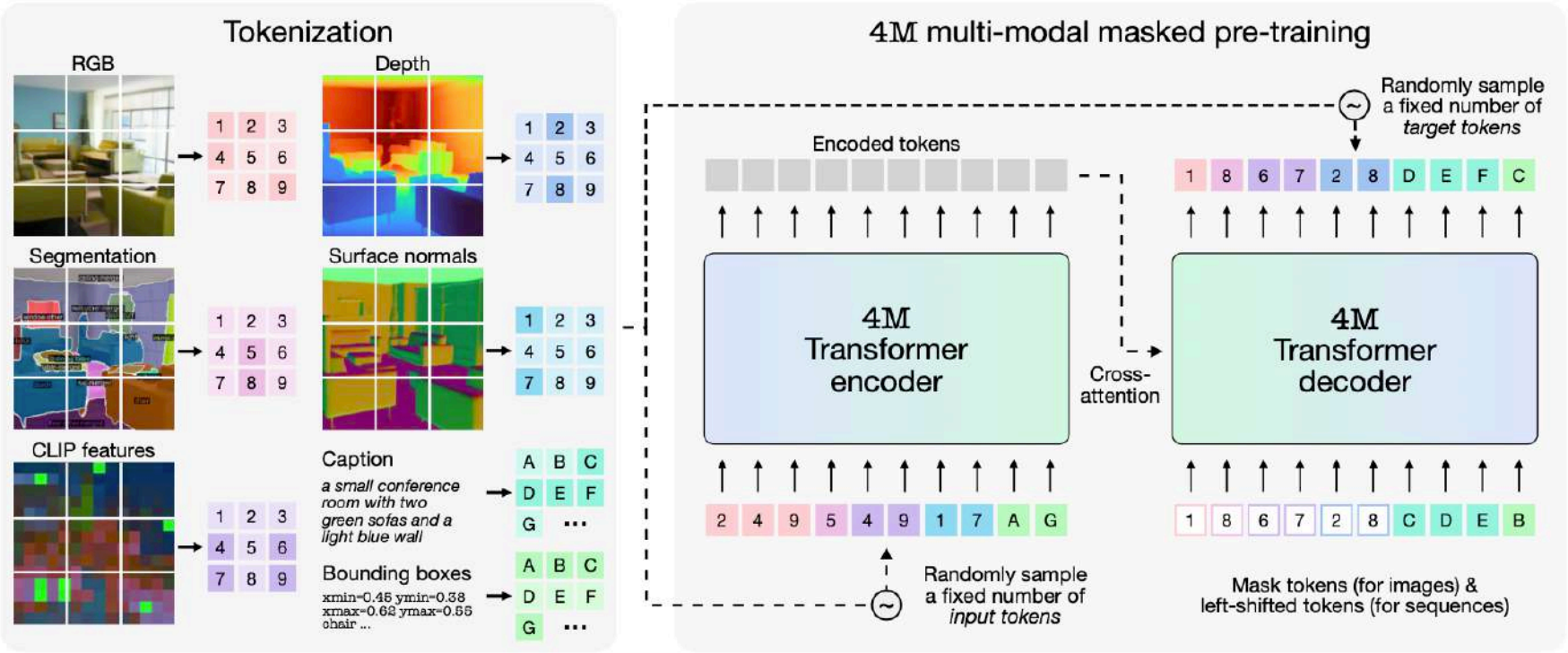


Cross-Modal Masked Modeling

MultiMAE: Multi-Modal Multi-Task Masked Autoencoders



MultiMAE: Multi-Modal Multi-Task Masked Autoencoders, Bachmann, Mizrahi, Atanov, Zamir. ECCV 2022



- **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.

- Mizrahi, Bachmann, Kar, Yeo, Gao, Dehghan, Zamir. NeurIPS 2023.

https://4m.epfl.ch/

modalities

RGB modalities

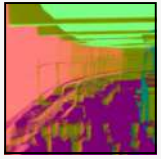
RGB

Color palette

Geometric modalities

Depth

Surface normals



3D human poses

Semantic modalities

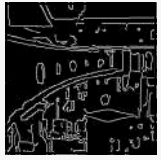
Bounding boxes

Semantic segmentation

SAM instances

Edge modalities

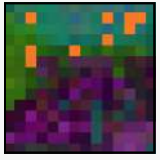
SAM edges



Canny edges

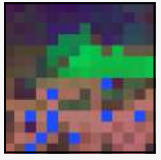
Feature map modalities

CLIP features (dense)



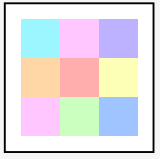
DINOv2 features (dense)

ImageBind features (dense)

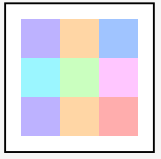


Global feature modalities

DINOv2 features (global)



ImageBind features (global)



Text modalities

Caption

Getting ready for my flight!

T5-XXL embeddings

Web text

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

Metadata modalities

Image metadata

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

Semantic metadata

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

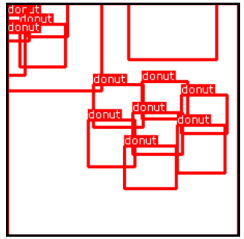
Geometric metadata

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

Out-of-the-box multi-task model

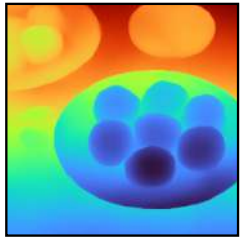
Query image

Caption

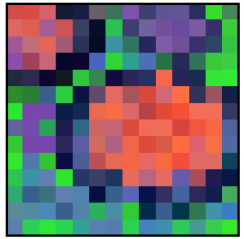


Bounding Boxes

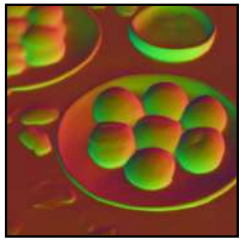
Semantic Seg.



Depth

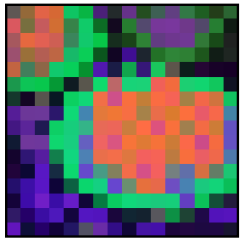


CLIP

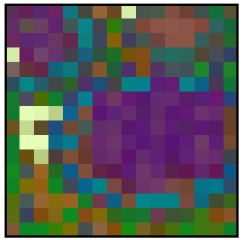


Surface Normals

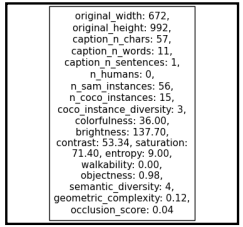
Human poses



DINOv2

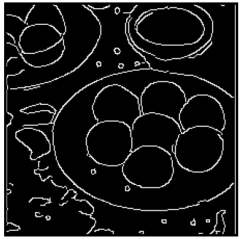


ImageBind

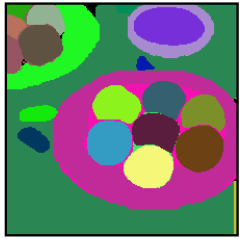


Metadata

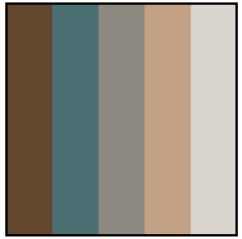
Texture Edges



SAM Edges



SAM instances

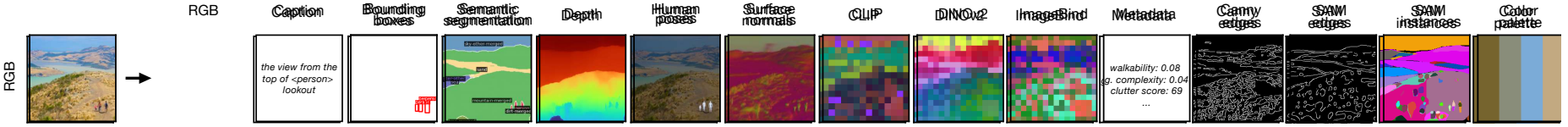


Color Palette

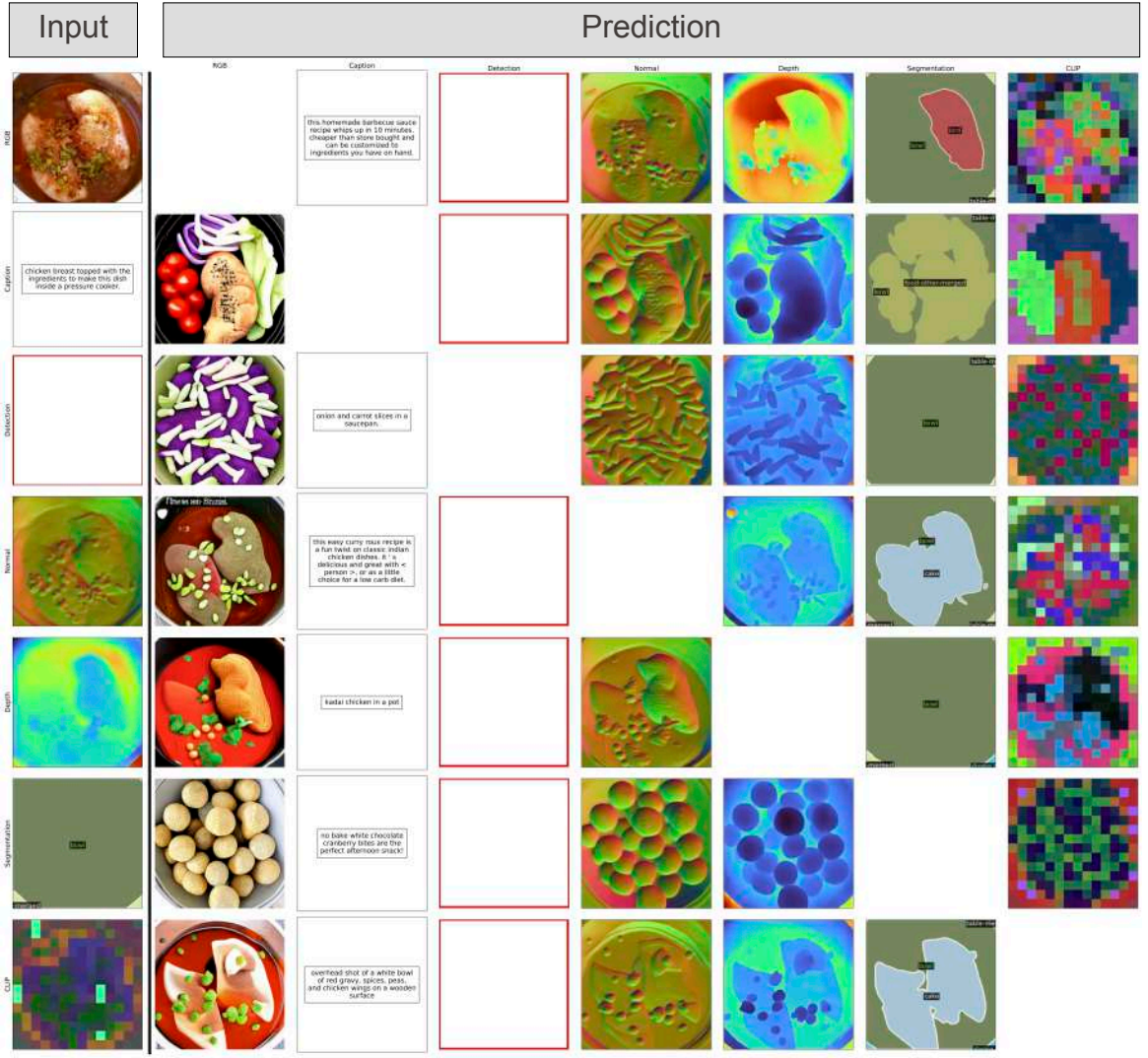
Any-to-Any generation

RGB

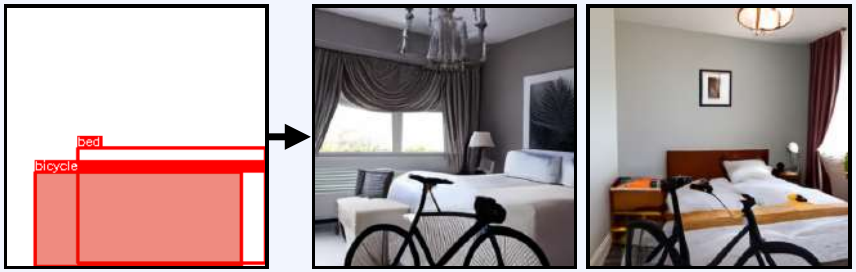
Any-to-Any generation

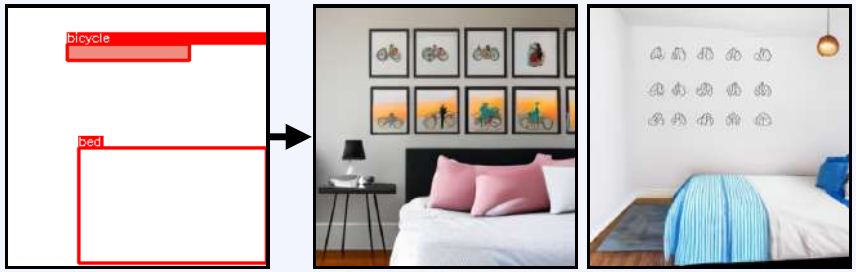


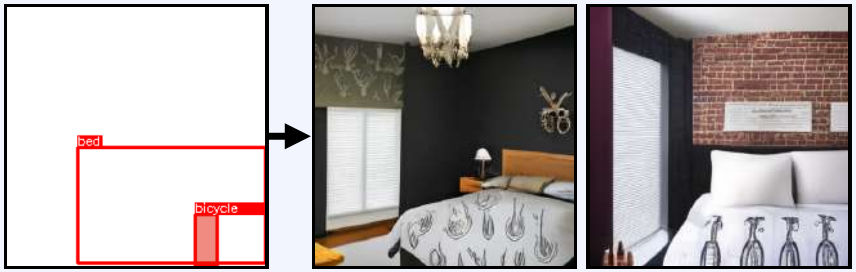
Any-to-Any generation



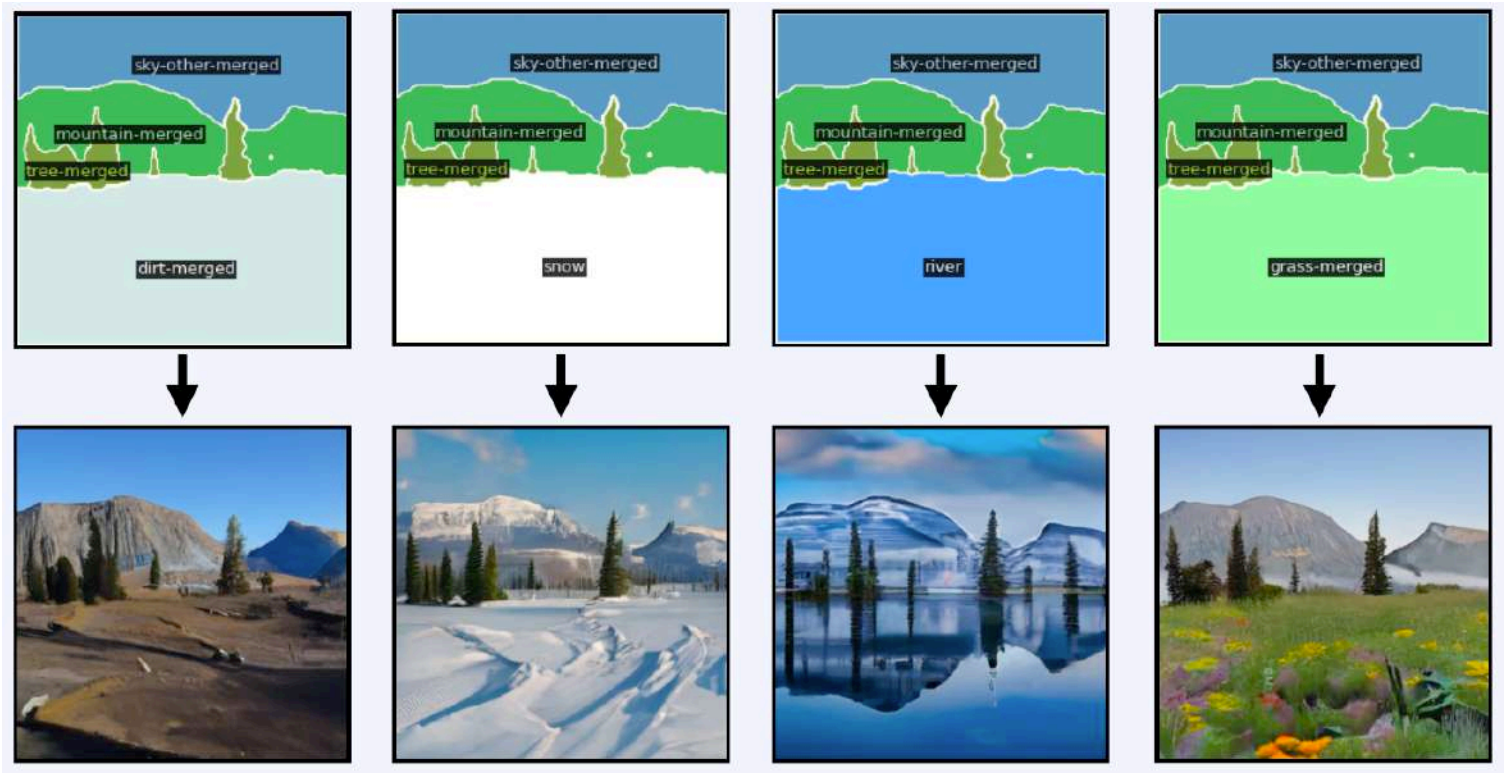
Probing the learned model







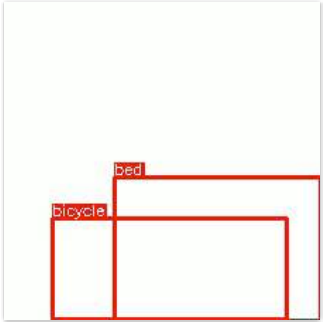
Probing the learned model



Probing the learned model

Probing the learned model

Bounding box input



Caption input

a photo of a bedroom, studio light

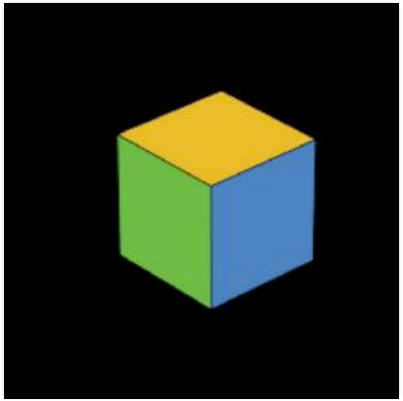
Frame-by-frame Predictions

Probing the learned model

Bounding box input

Probing the learned model

Changing SAM polygon input



+

color palette

→

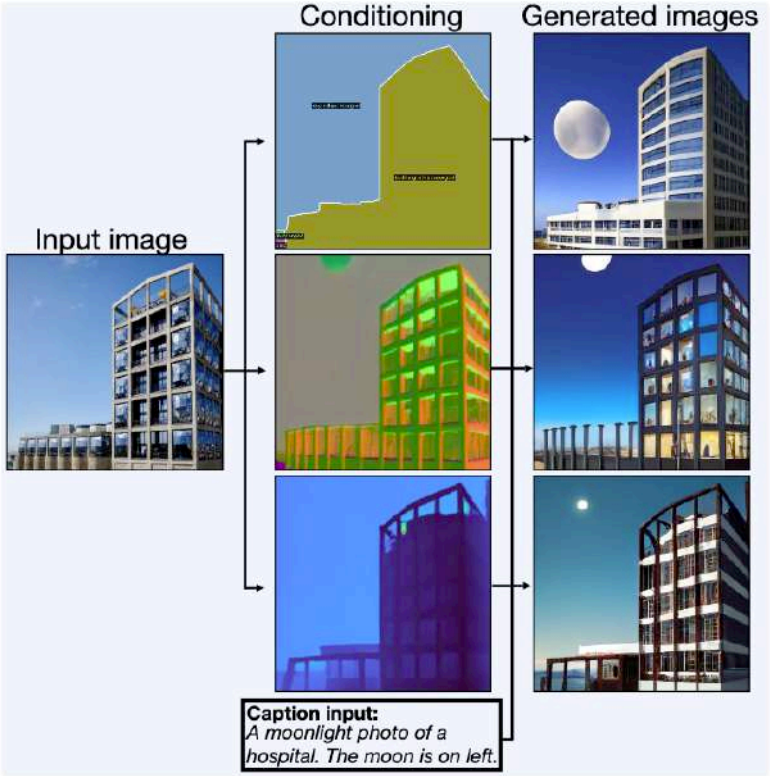
RGB prediction

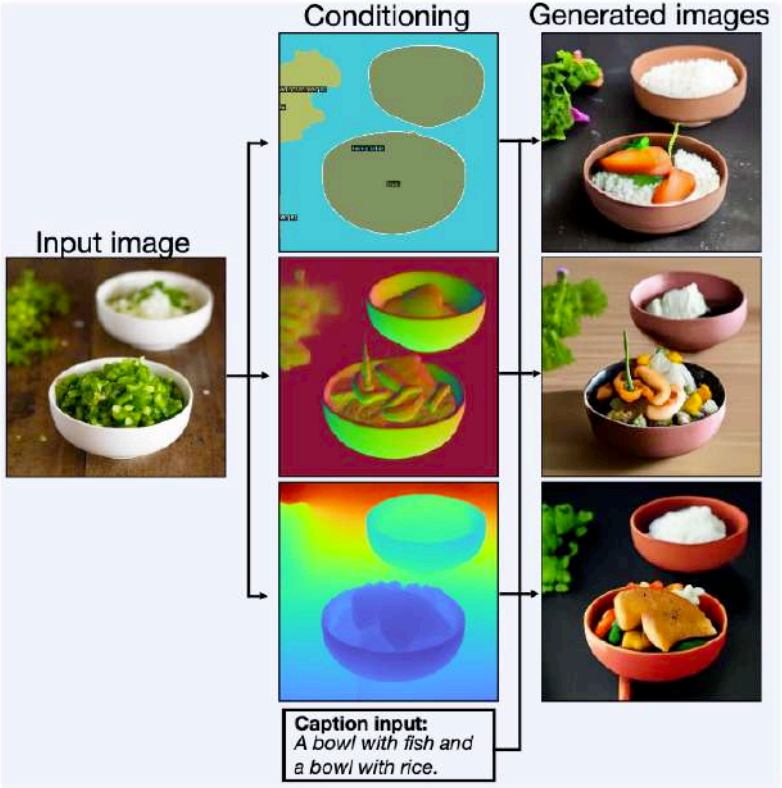
RGB prediction (with polygon overlay)

+

Fixed caption
a framed painting of mountains inside a *bedroom*

Grounded generation





multi-modal guidance & weighting

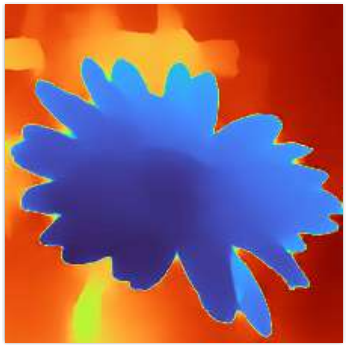
Caption input

an oil painting of a blue flower

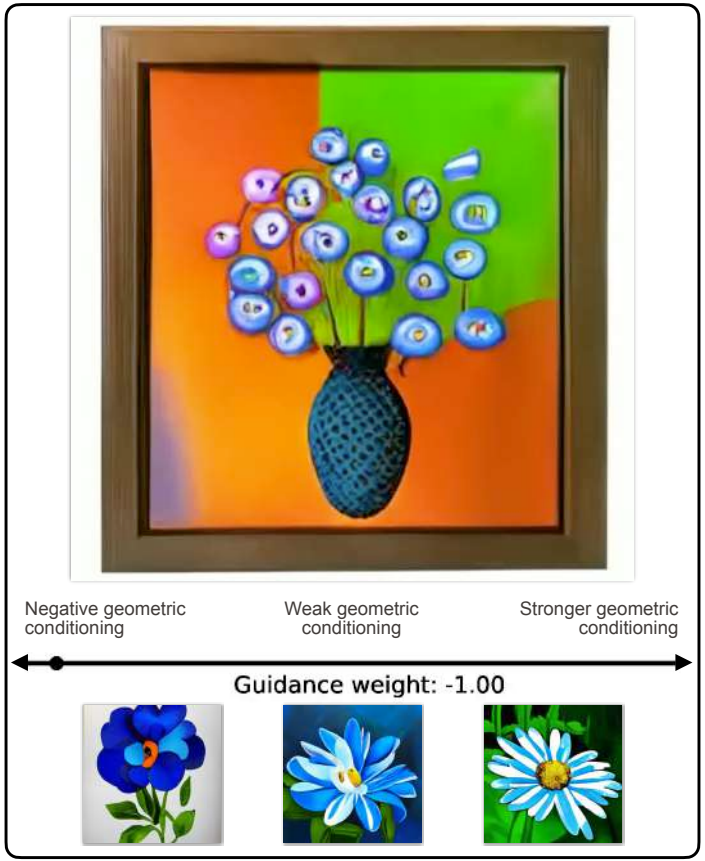
Fixed weight

2.0

Depth

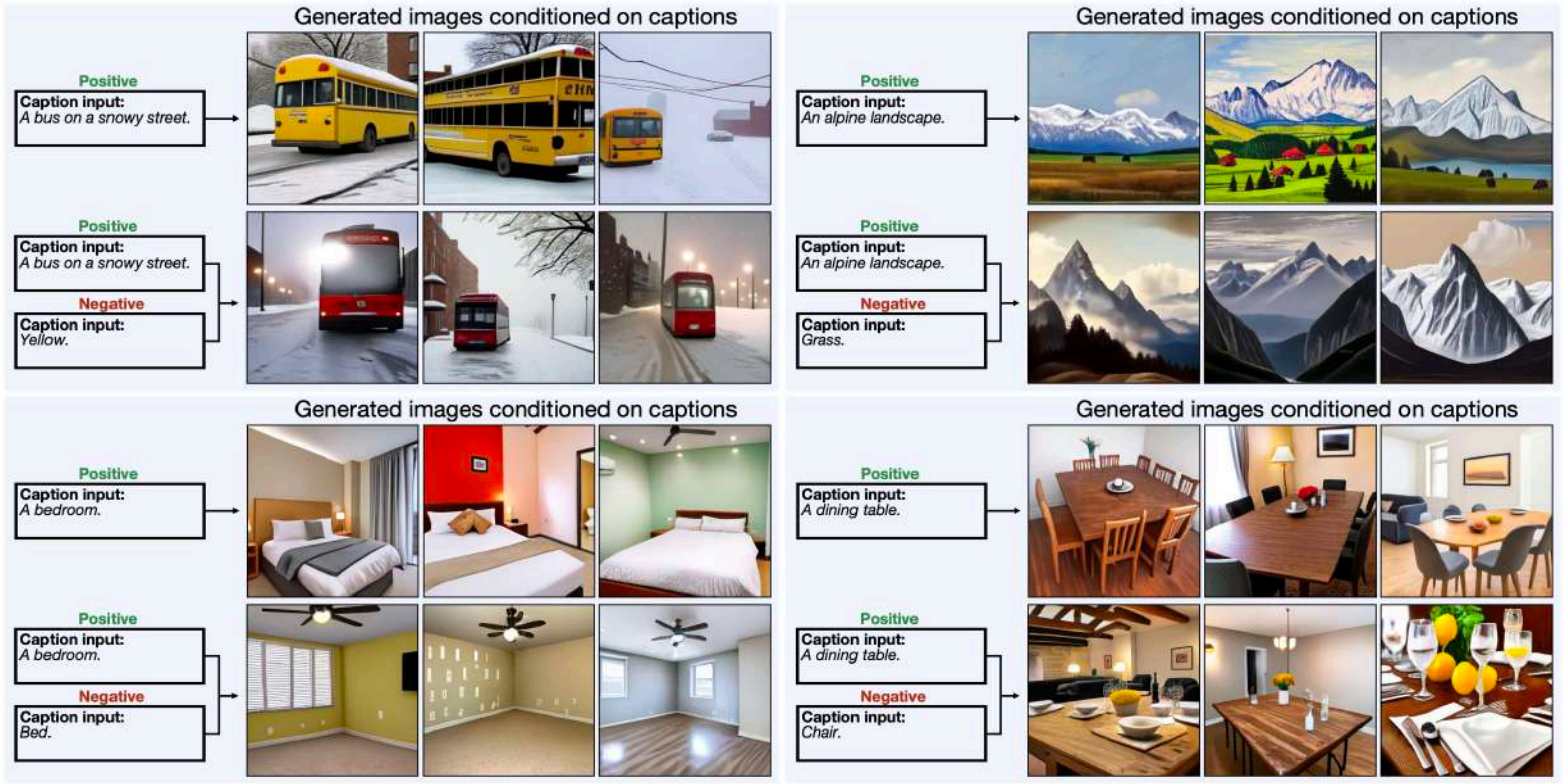


Varying weight



Classifier-free diffusion guidance. Ho and Salimans. ArXiv 2022.

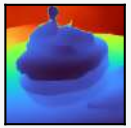
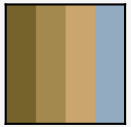
multi-modal guidance & weighting



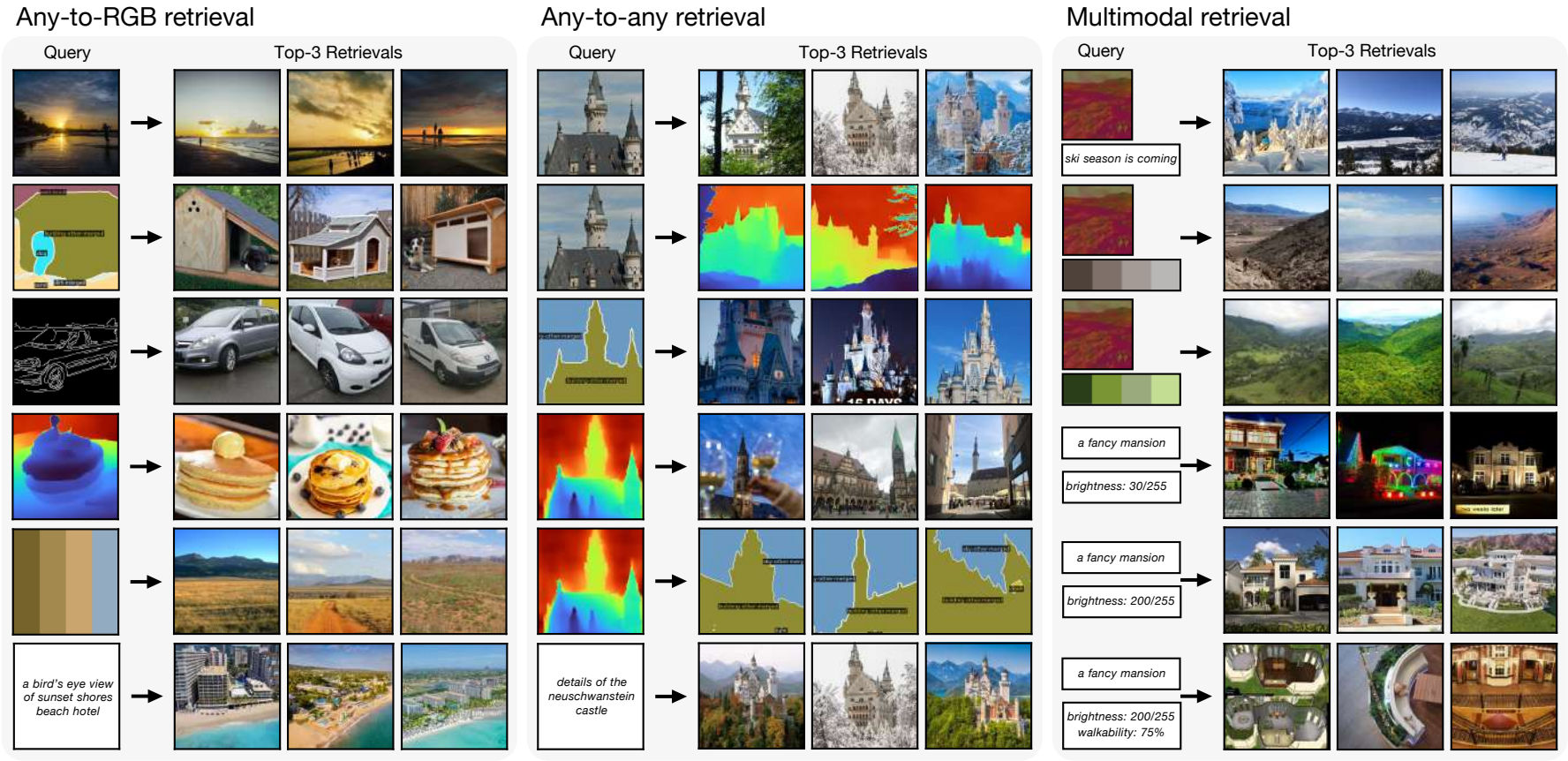
Classifier-free diffusion guidance. Ho and Salimans. ArXiv 2022.

Multimodal retrieval

Any-to-RGB retrieval

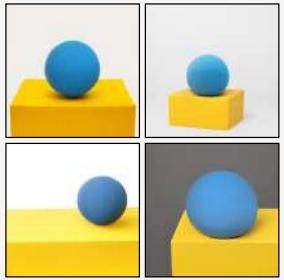
Query	Top-3 Retrievals
	
	
	
	
	
a bird's eye view of sunset shores beach hotel	

Multimodal retrieval



LLM Co-Training/Representation

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

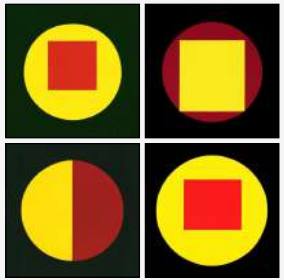


4M-7 (from caption)

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

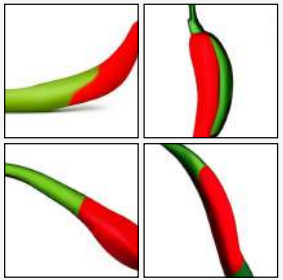
4M-7 (from caption)

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



4M-7 (from caption)

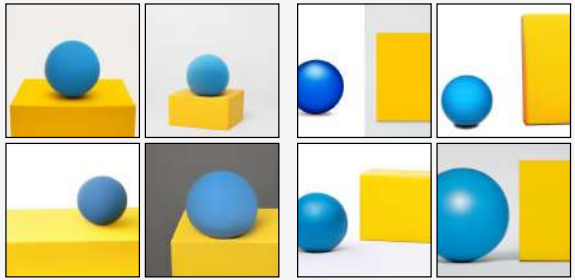
Caption input: *a green pepper to the left of a red pepper*



4M-7 (from caption)

LLM Co-Training/Representation

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

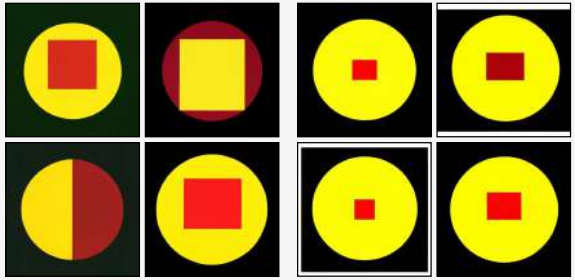


4M-7 (from caption) 4M-21 (from caption)

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

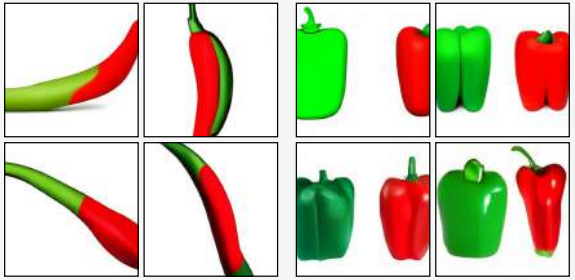
4M-7 (from caption) 4M-21 (from caption)

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



4M-7 (from caption) 4M-21 (from caption)

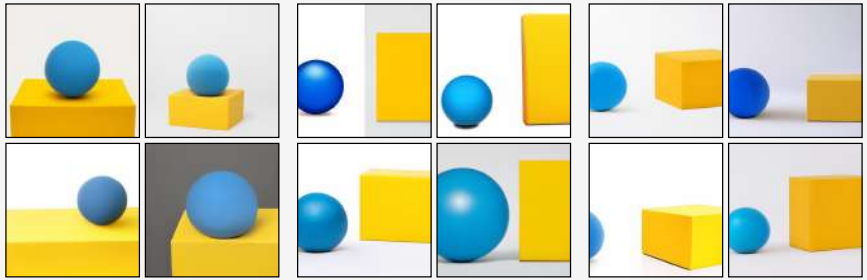
Caption input: *a green pepper to the left of a red pepper*



4M-7 (from caption) 4M-21 (from caption)

LLM Co-Training/Representation

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

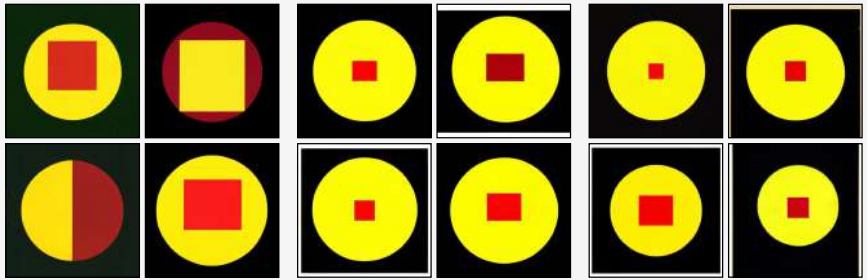


4M-7 (from caption) | 4M-21 (from caption) | 4M-21 (from T5-XXL emb.)

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

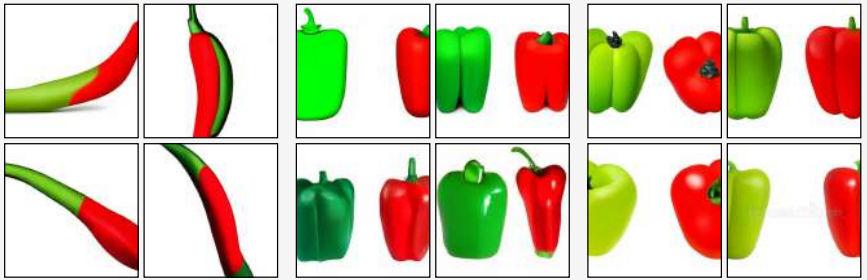
4M-7 (from caption) | 4M-21 (from caption) | 4M-21 (from T5-XXL emb.)

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



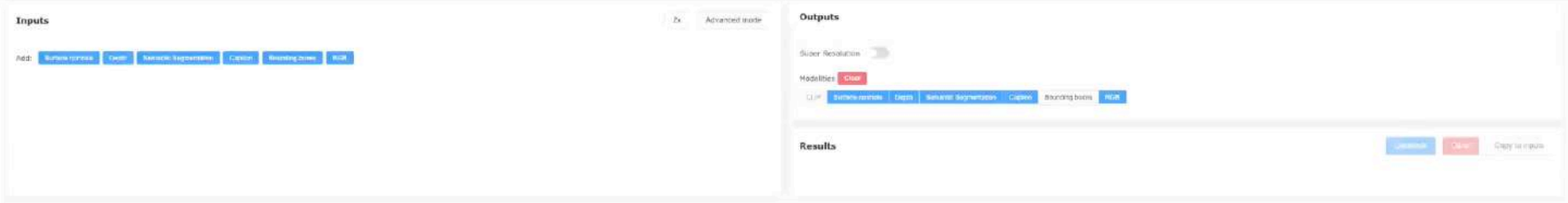
4M-7 (from caption) | 4M-21 (from caption) | 4M-21 (from T5-XXL emb.)

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



4M-7 (from caption) | 4M-21 (from caption) | 4M-21 (from T5-XXL emb.)

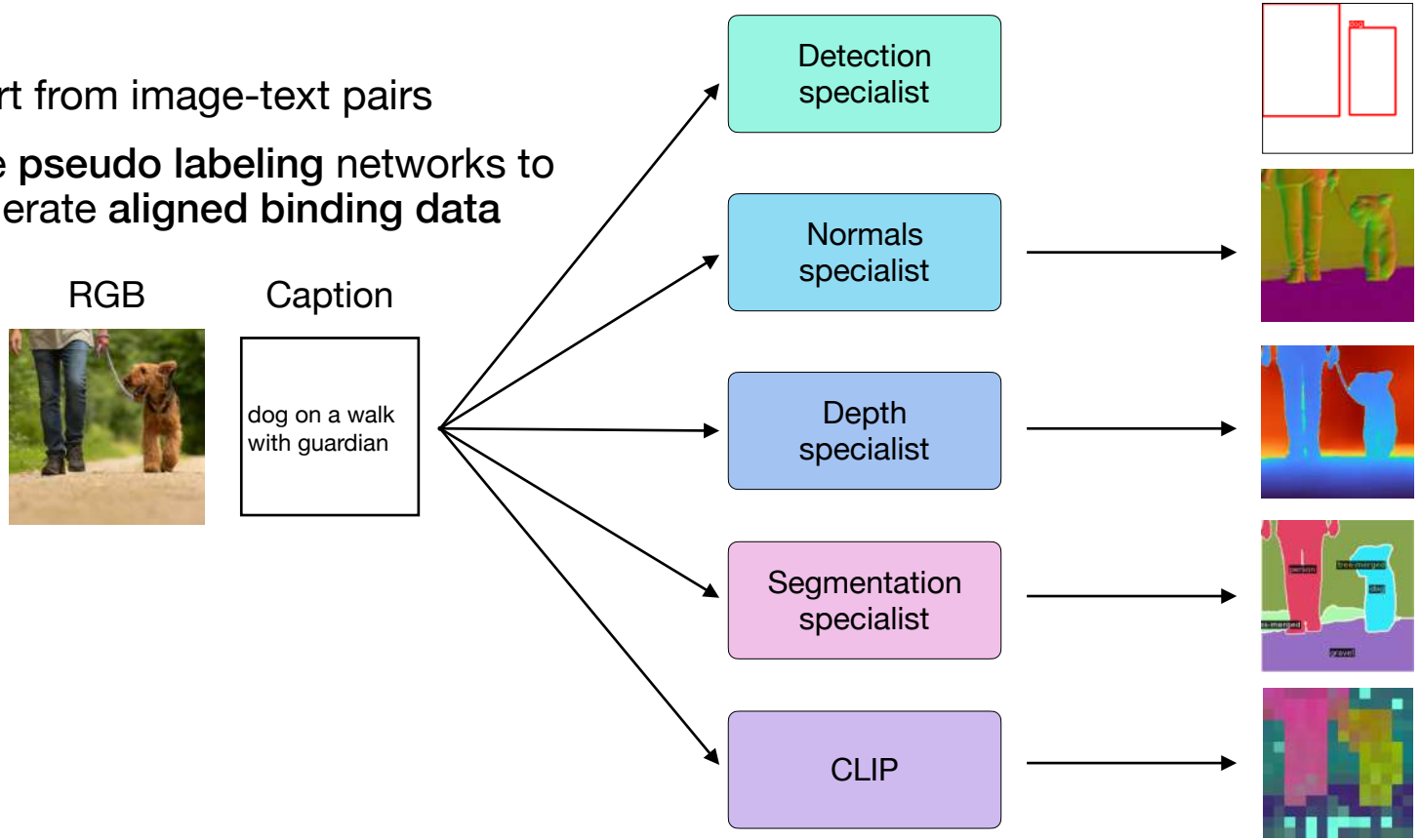
Live Demo



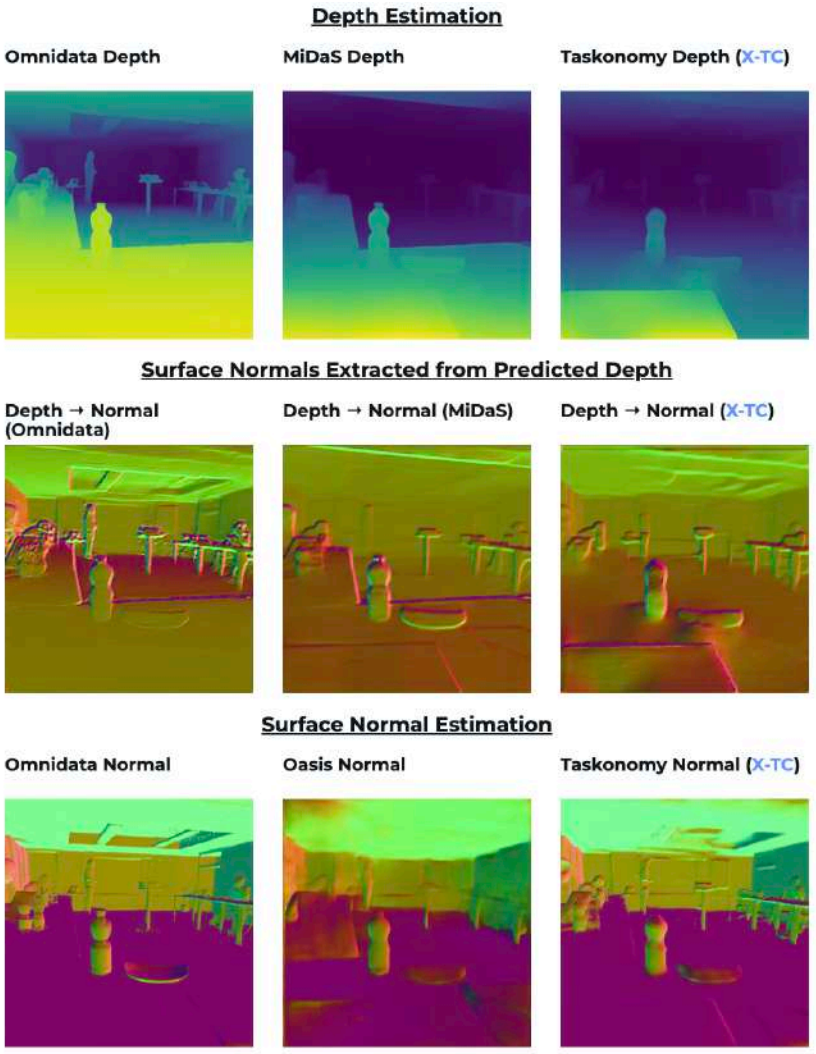
https://4m.epfl.ch/

Pseudo Labeling

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



Pseudo Labeling



https://omnidata.vision/demo/

- Omnidata, ICCV'21.
- 3D Common Corruptions CVPR'22.

Modalities

RGB modalities

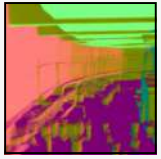
RGB

Color palette

Geometric modalities

Depth

Surface normals



3D human poses

Semantic modalities

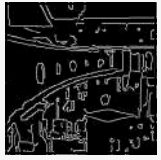
Bounding boxes

Semantic segmentation

SAM instances

Edge modalities

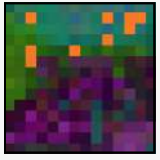
SAM edges



Canny edges

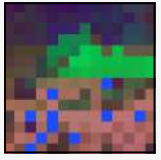
Feature map modalities

CLIP features (dense)



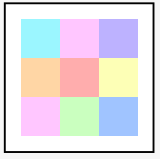
DINOv2 features (dense)

ImageBind features (dense)

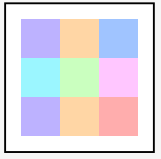


Global feature modalities

DINOv2 features (global)



ImageBind features (global)



Text modalities

Caption

Getting ready for my flight!

T5-XXL embeddings

Web text

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

Metadata modalities

Image metadata

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

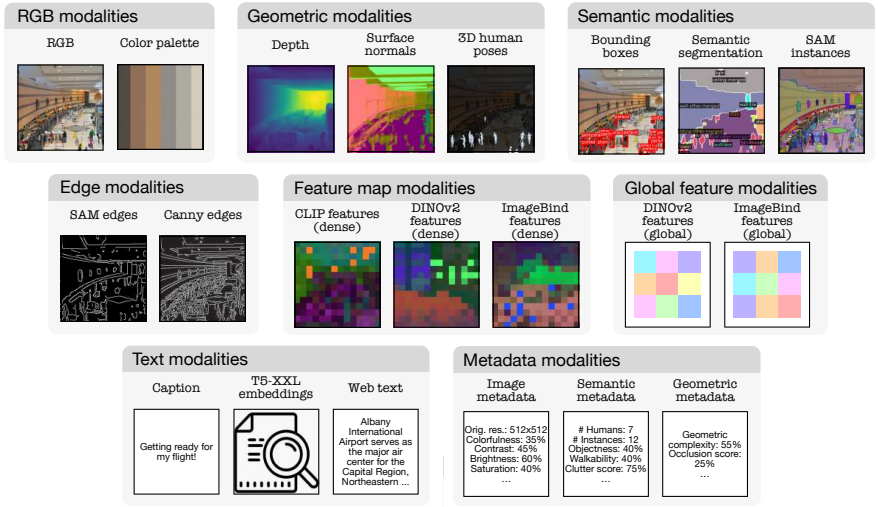
Semantic metadata

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

Geometric metadata

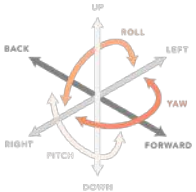
Geometric complexity: 55%
Occlusion score: 25%
...

Modalities

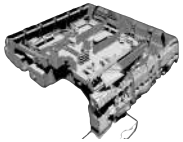


Planned

Motion



IMU & Motor control



3D & space-level fusion

Sketch

Video

Multi-view

Quantitative comparisons

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 [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	108.3
	4M-7 L [62]	21.5	0.69	47.2	✗	✗	✗
	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**	✗	✗	✗
	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.**

Quantitative comparisons

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 [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	108.3
	4M-7 L [62]	21.5	0.69	47.2	✗	✗	✗
	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**	✗	✗	✗
	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

	NYUv2-S mIoU ↑		Hypersim mIoU ↑		ARKitScenes AP^{3D}↑	
Method	RGB	RGB-D	RGB	RGB-D	RGB	RGB-D
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 Acc.↑	ADE20K mIoU↑	NYUv2-D δ_1 acc.↑	ARKS AP^{3D}↑
MAE B [38]	IN1K	86M	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		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	303M	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		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	1.2B	87.0	55.0	96.1	48.1
4M-7 XL (Ours)	COYO		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

Reasoning & Test-Time Compute

Solving a task via coding

Query: How many muffins can each kid have for it to be fair?

Generated Code

```
def execute_command(image):
    image_patch = ImagePatch(image)
    muffin_patches = image_patch.find("muffin")
    kid_patches = image_patch.find("kid")
    return str(len(muffin_patches) // len(kid_patches))
```

Execution

```
muffin_patches =
image_patch.find("muffin")
```

```
kid_patches =
image_patch.find("kid")
```

```
▶len(muffin_patches)=8
▶len(kid_patches)=2
▶8//2 = 4
Result:4
```

Query: Drink with zero alcohol

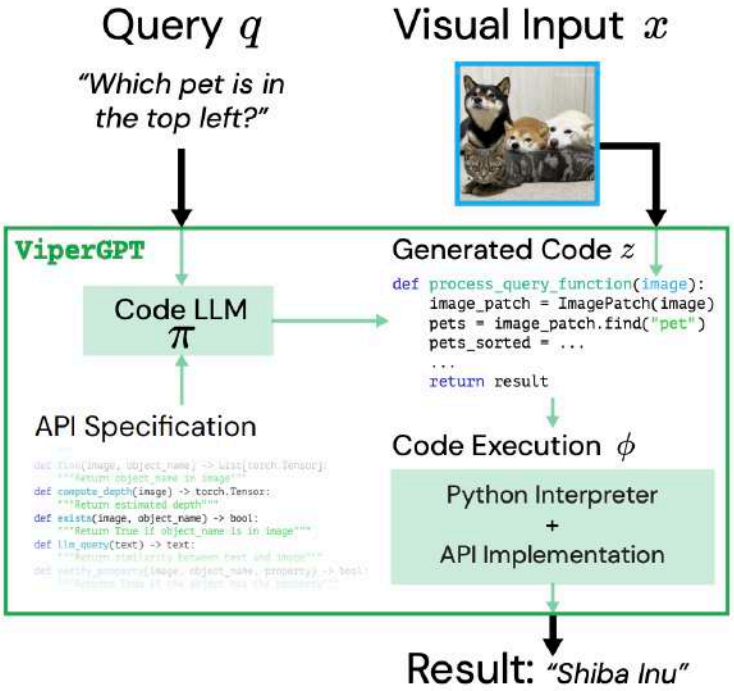
```
def execute_command(image):
    image_patch = ImagePatch(image)
    drink_patches = image_patch.find("drink")
    for drink_patch in drink_patches:
        drink_name = drink_patch.simple_query("What is this?")
        alcoholic = llm_query(f"Does the {drink_name} have alcohol?")
        if alcoholic == "no":
            return drink_patch
    return None
```

```
▶drink_patches=
```

```
▶drink_name = 'tullamore dew'
▶alcoholic = 'yes'
▶drink_name = 'bacardi'
▶alcoholic = 'yes'
▶drink_name = 'gin'
▶alcoholic = 'yes'
▶drink_name = 'dr pepper'
▶alcoholic = 'no'
Result:
```

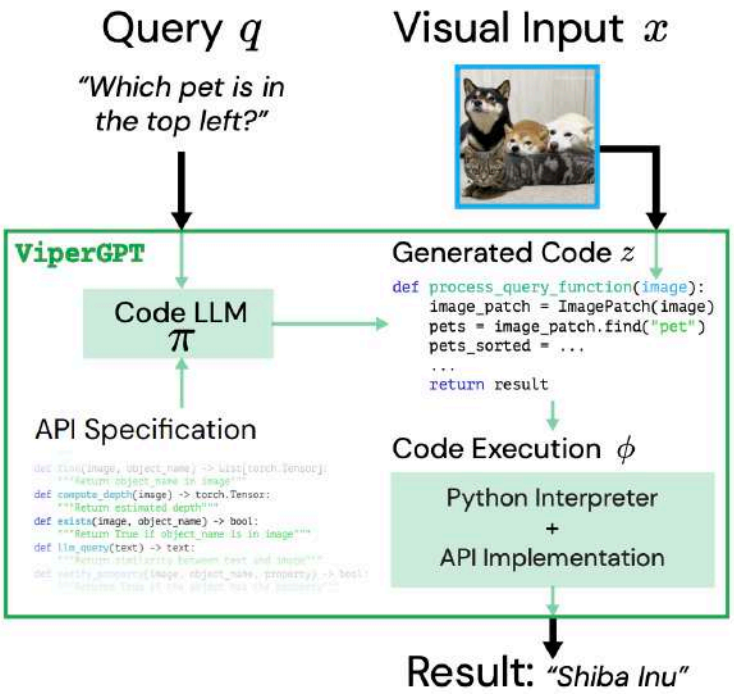
ViperGPT: ViperGPT: Visual Inference via Python Execution for Reasoning
https://viper.cs.columbia.edu/

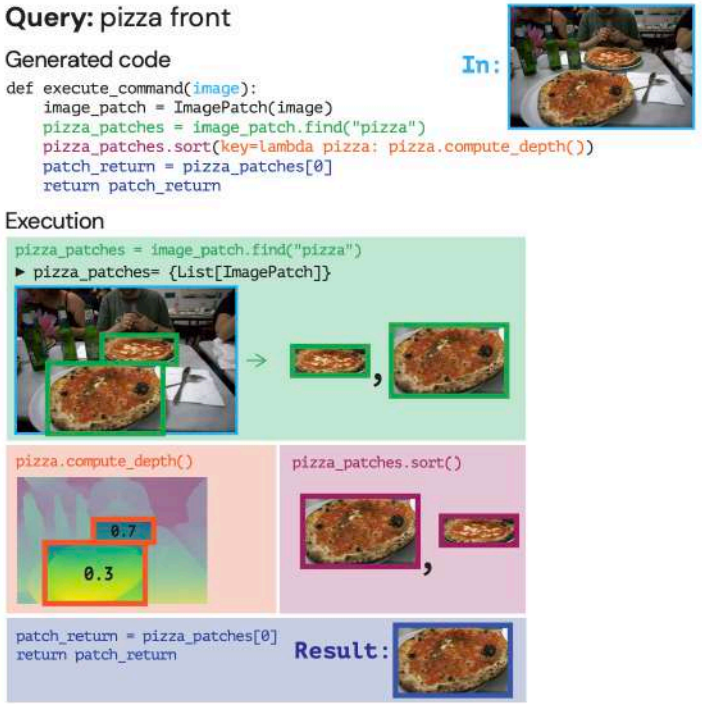
Solving a task via coding



ViperGPT: ViperGPT: Visual Inference via Python Execution for Reasoning
https://viper.cs.columbia.edu/

Solving a task via coding





ViperGPT: ViperGPT: Visual Inference via Python Execution for Reasoning
https://viper.cs.columbia.edu/

Solving a task via coding

```python
def incr_list(l: list):
    """Return list with elements incremented by 1.
    >>> incr_list([1, 2, 3])
    [2, 3, 4]
    >>> incr_list([5, 3, 5, 2, 3, 3, 9, 0, 123])
    [6, 4, 6, 3, 4, 4, 10, 1, 124]
    """
    return [i + 1 for i in l]
```

```python
def solution(lst):
    """Given a non-empty list of integers, return the sum of all of the odd elements
    that are in even positions.

    Examples
    solution([5, 8, 7, 1]) ==>12
    solution([3, 3, 3, 3, 3]) ==>9
    solution([30, 13, 24, 321]) ==>0
    """
    return sum(lst[i] for i in range(0,len(lst)) if i % 2 == 0 and lst[i] % 2 == 1)
```

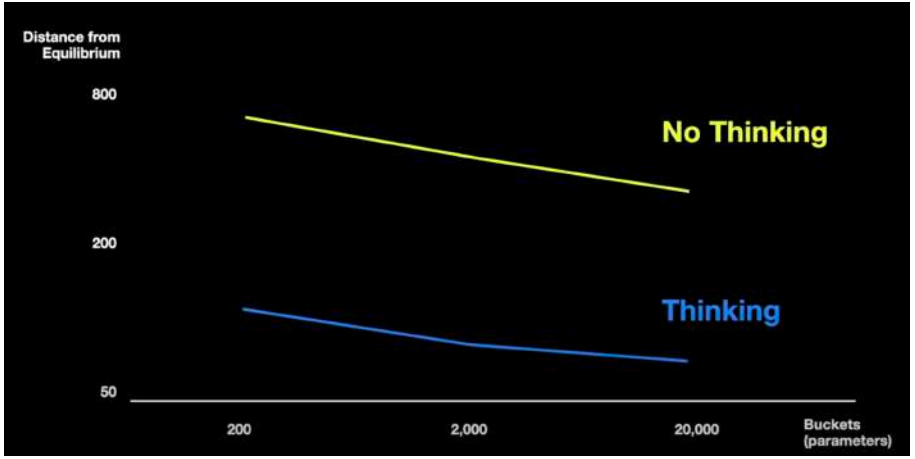
```python
def encode_cyclic(s: str):
    """
    returns encoded string by cycling groups of three characters.
    """
    # split string to groups. Each of length 3.
    groups = [s[(3 * i):min((3 * i + 3), len(s))] for i in range((len(s) + 2) // 3)]
    # cycle elements in each group. Unless group has fewer elements than 3.
    groups = [(group[1:] + group[0]) if len(group) == 3 else group for group in groups]
    return "".join(groups)

def decode_cyclic(s: str):
    """
    takes as input string encoded with encode_cyclic function. Returns decoded string.
    """
    # split string to groups. Each of length 3.
    groups = [s[(3 * i):min((3 * i + 3), len(s))] for i in range((len(s) + 2) // 3)]
    # cycle elements in each group.
    groups = [(group[-1] + group[:-1]) if len(group) == 3 else group for group in groups]
    return "".join(groups)
```

Figure 2. Three example problems from the HumanEval dataset, where the probabilities that a single sample from Codex-12B passes unit tests are 0.9, 0.17, and 0.005. The prompt provided to the model is shown with a white background, and a successful model-generated completion is shown in a yellow background. Though not a guarantee for problem novelty, all problems were hand-written and not programmatically copied from existing sources. Random problems and samples can be found in Appendix B.

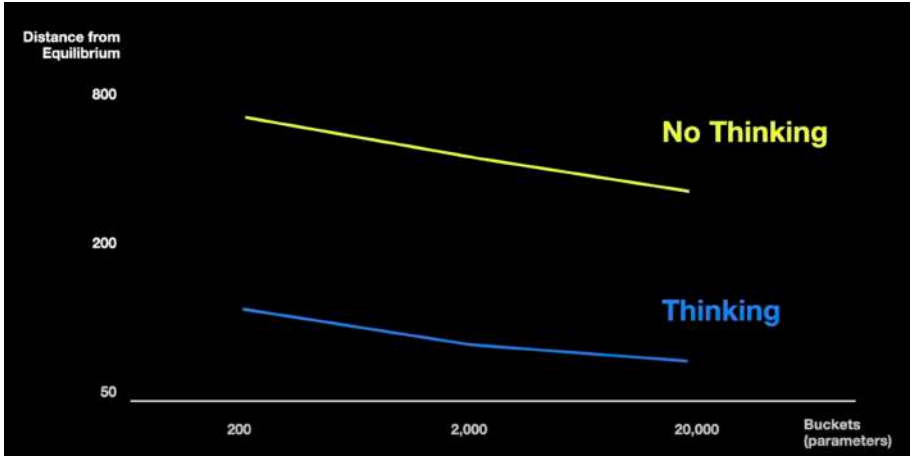
Codex: Evaluating Large Language Models Trained on Code. OpenAI

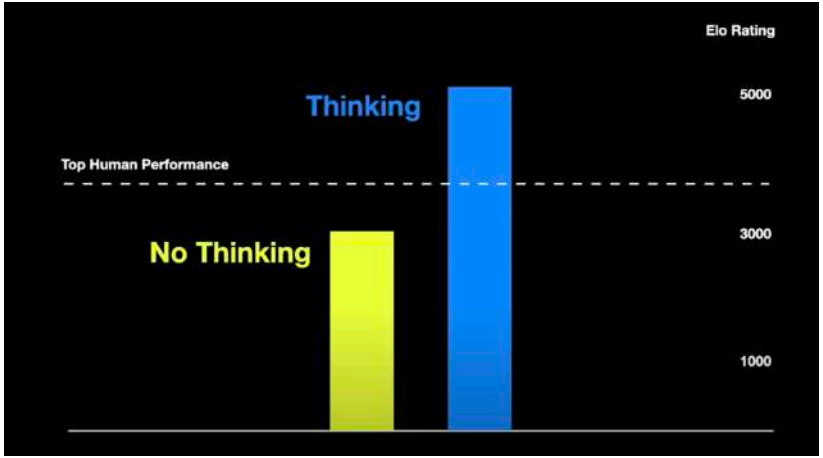
Test-Time Compute



Noam Brown

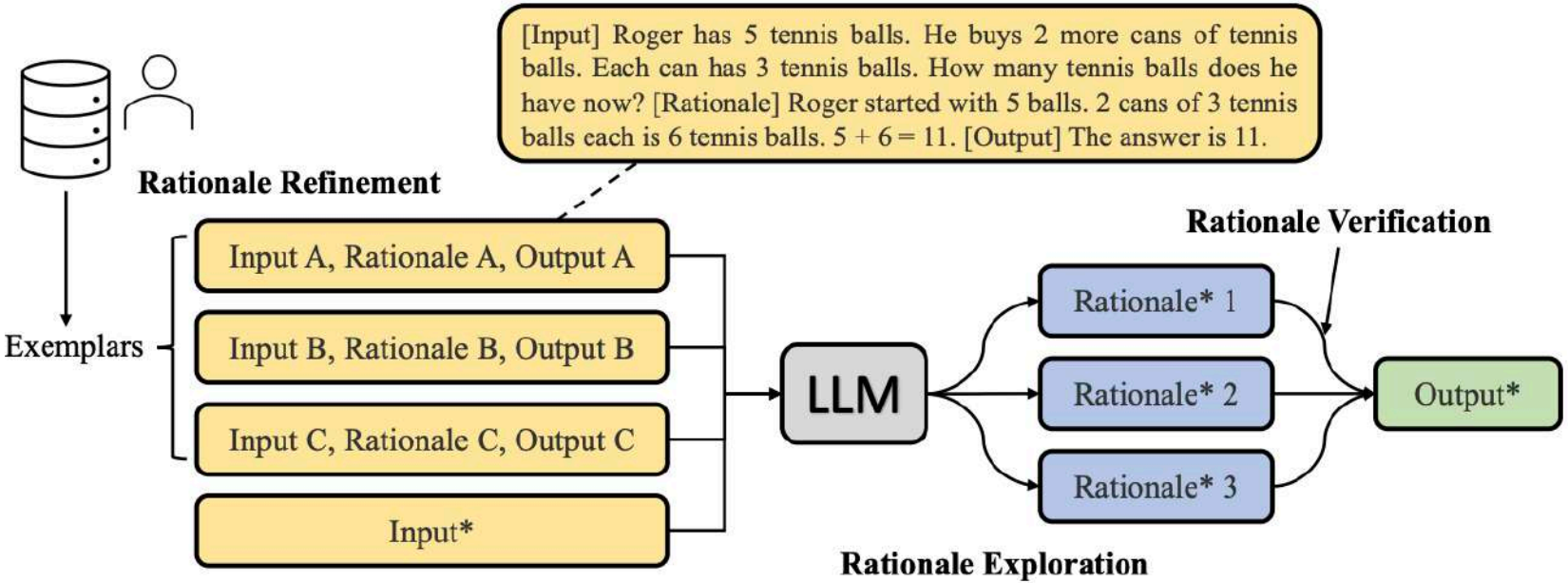
Test-Time Compute





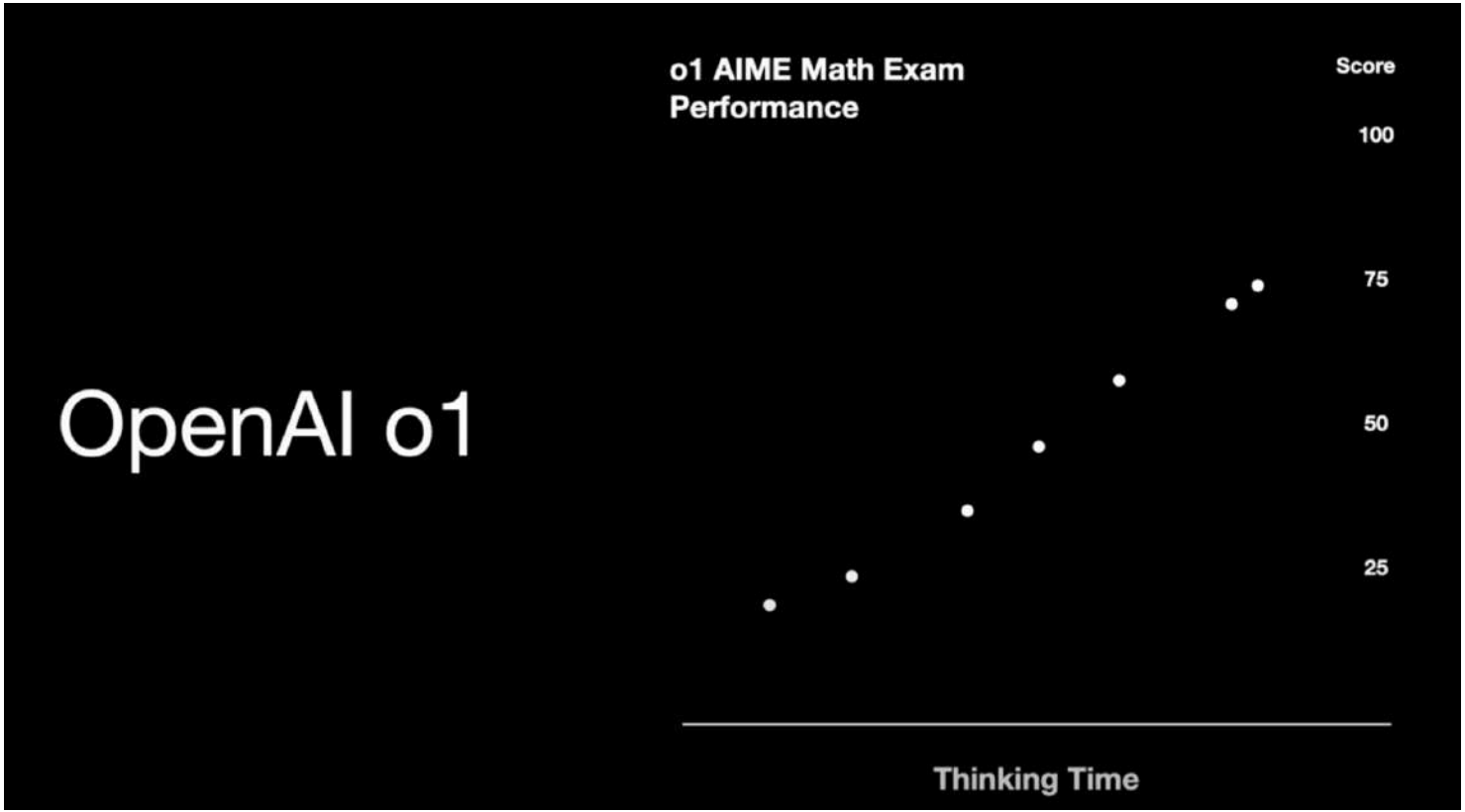
Noam Brown

Test-Time Compute



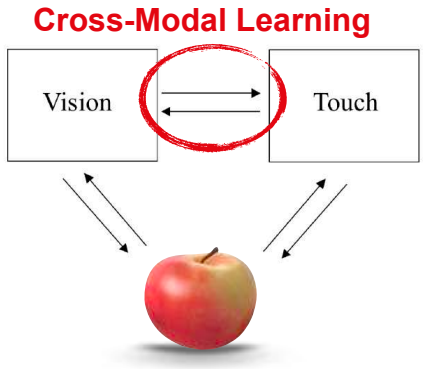
Towards Reasoning in Large Language Models: A Survey: Huang & Chang. 2023

Test-Time Compute



Noam Brown

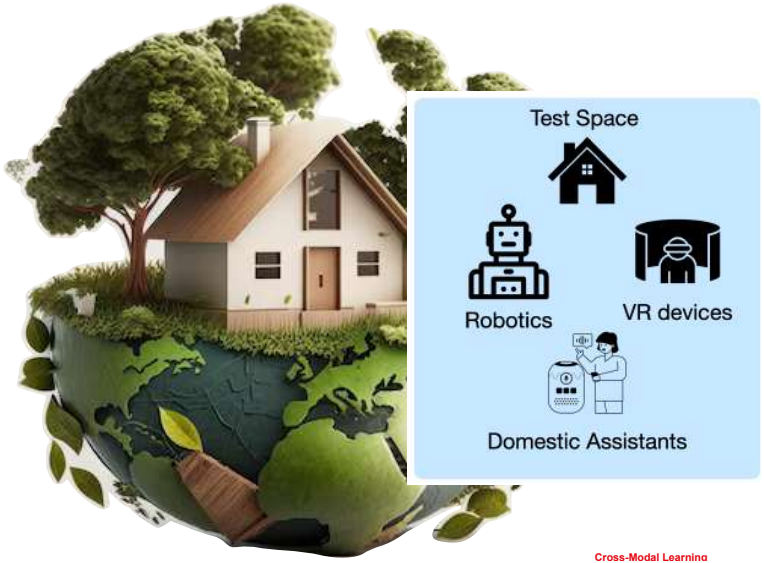
Multimodality as self-Supervision



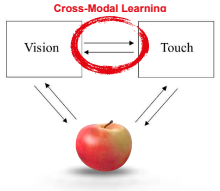
Multimodality as self-Supervision

Learn from the entire world/internet with few modalities

Learn from the test space only with rich modalities

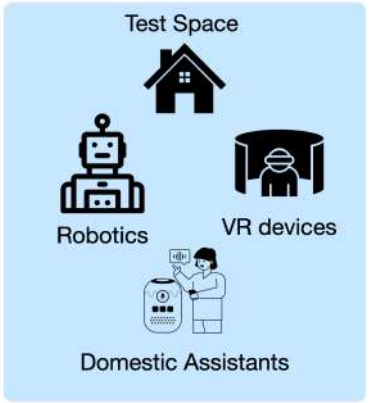


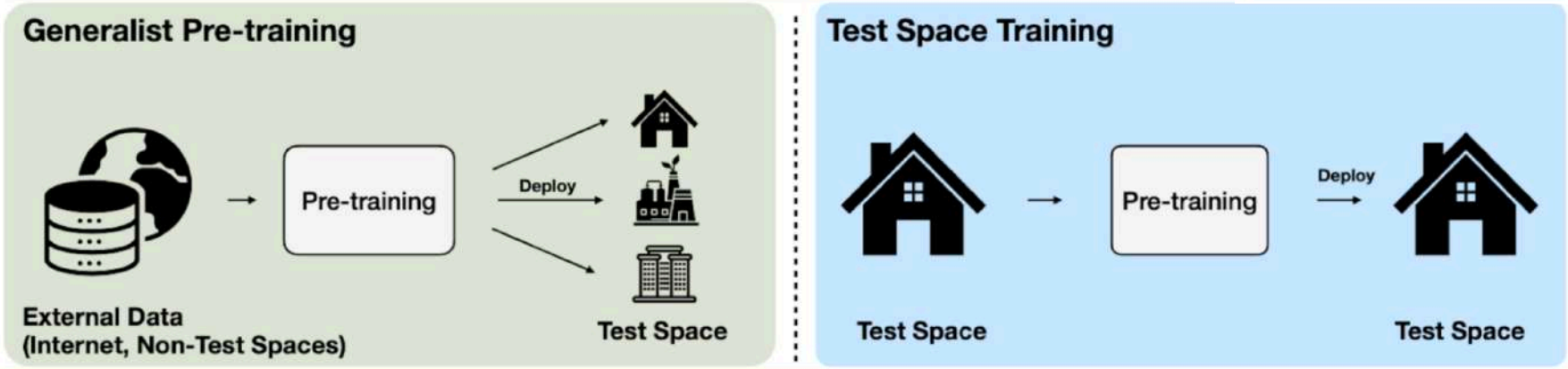
- Limit the world to the test space and "overfit" to it.
- Can we perfectly solve vision there?



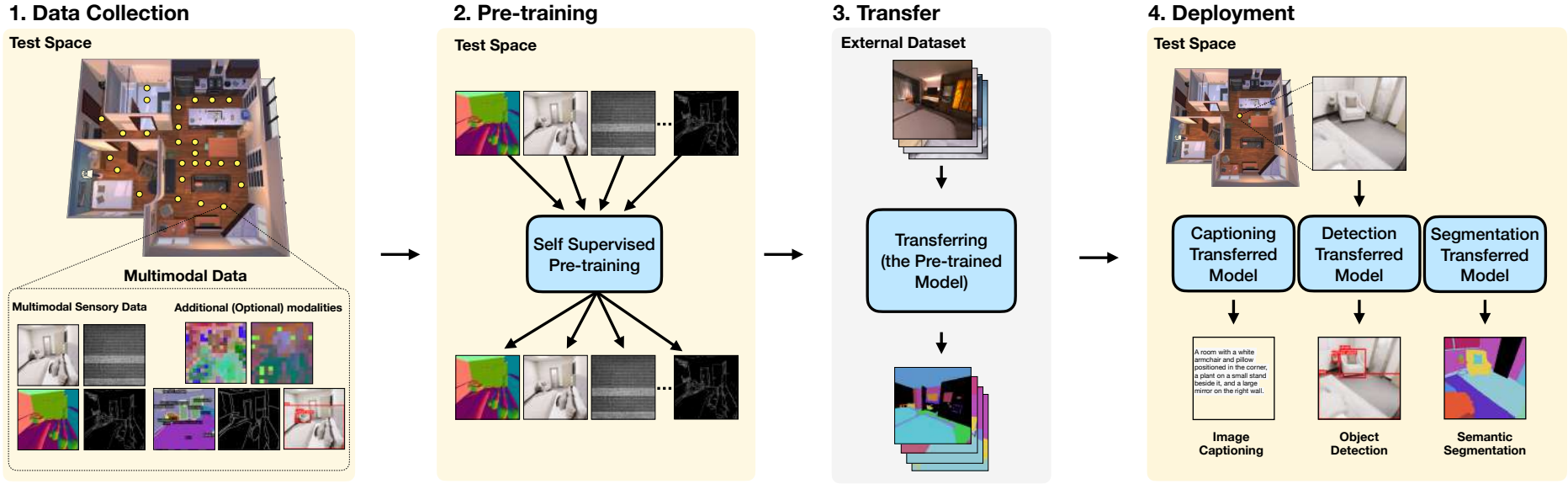
Multimodality as self-Supervision

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





Test-Space Training

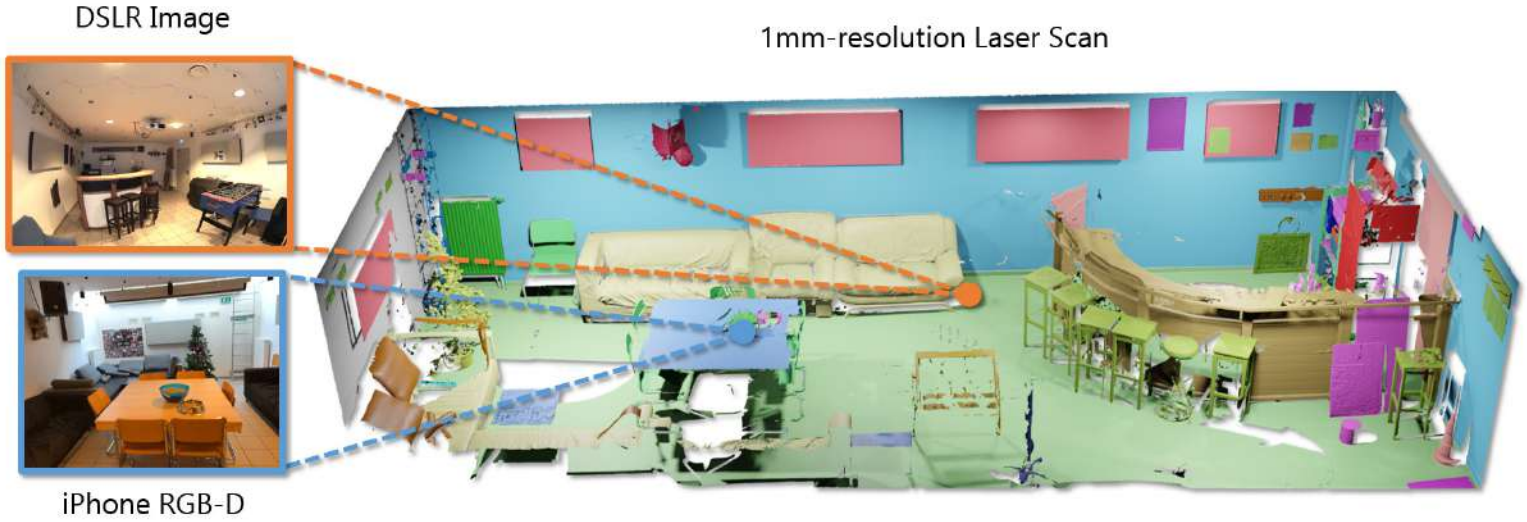


Experimental results

In Scannet++, Replica, THOR

On semantic segmentation, detection, captioning.

Scannet++[1]



1. Yeshwanth et al. 2023

Semantic Segmentation
Vs. internet-based generalists

Input

Ground Truth

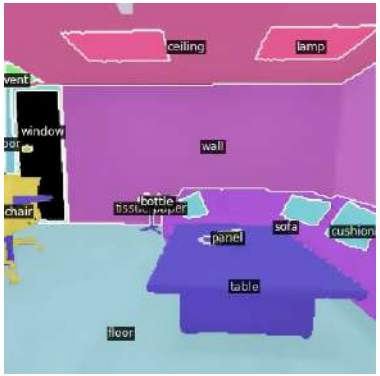
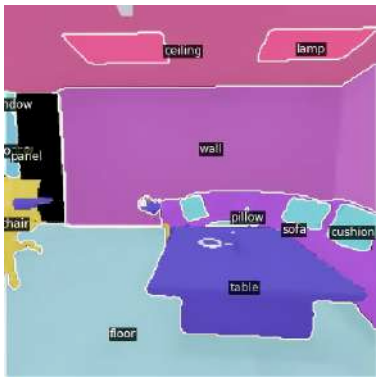
Test-Space Training (Ours)

DINOv2

CLIP

4M-21

Semantic Segmentation
Vs. task specialists

Input	Ground Truth	**Test-Space Training (Ours)**	Mask2Former
			
			

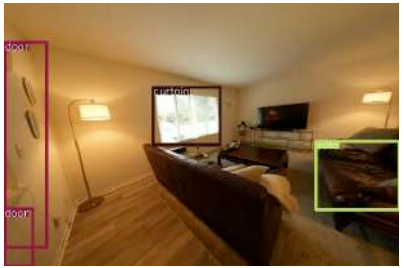
Detection

Input

Ground Truth

Test-Space Training (Ours)

CLIP



4M-21

ViTDet

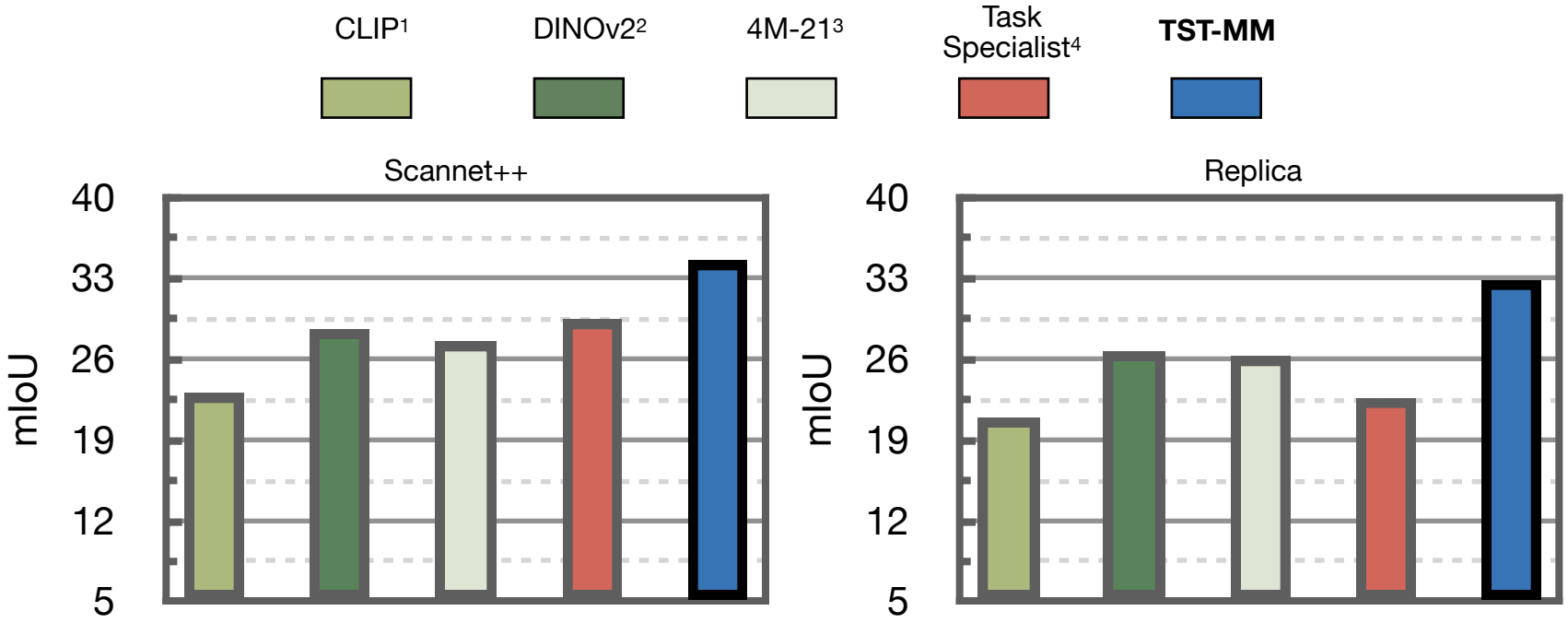
Scratch

DINOv2

Quantitive comparison

Semantic Segmentation

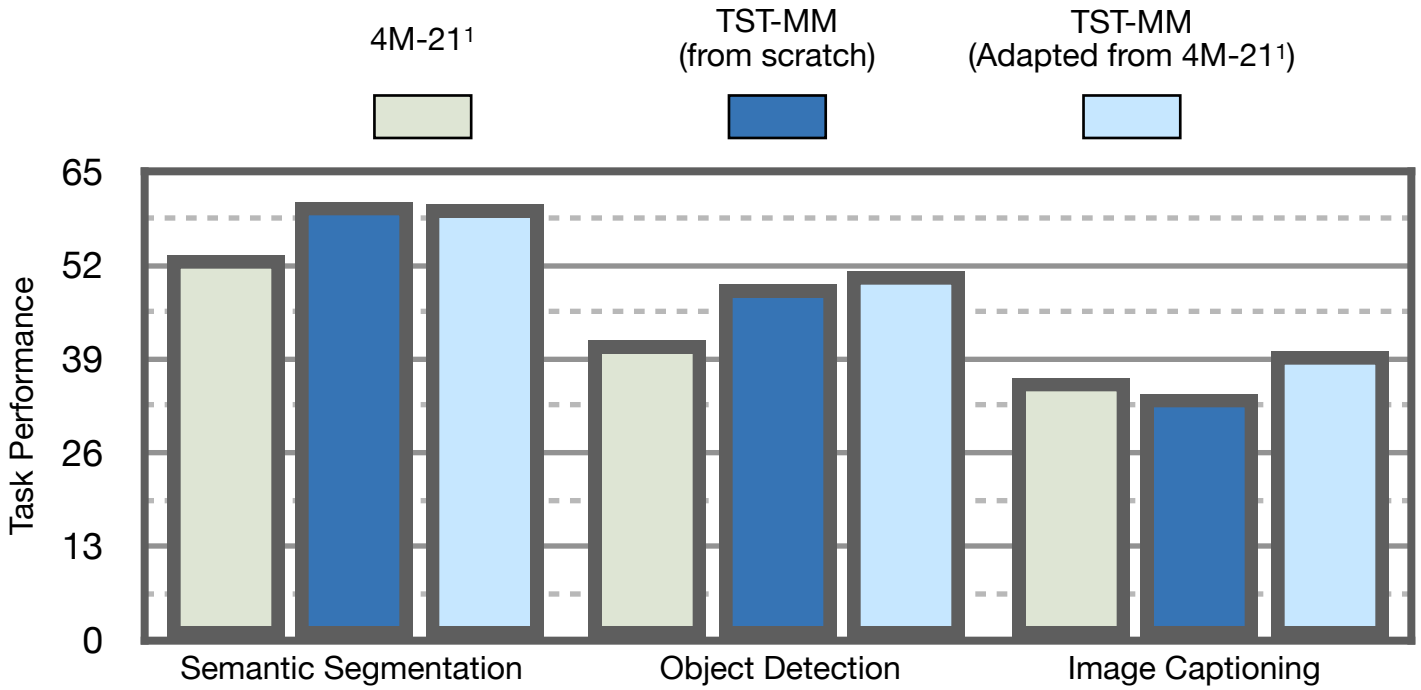
- TST outperforms internet based generalists[1,2,3], and task specialists[4,5].



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



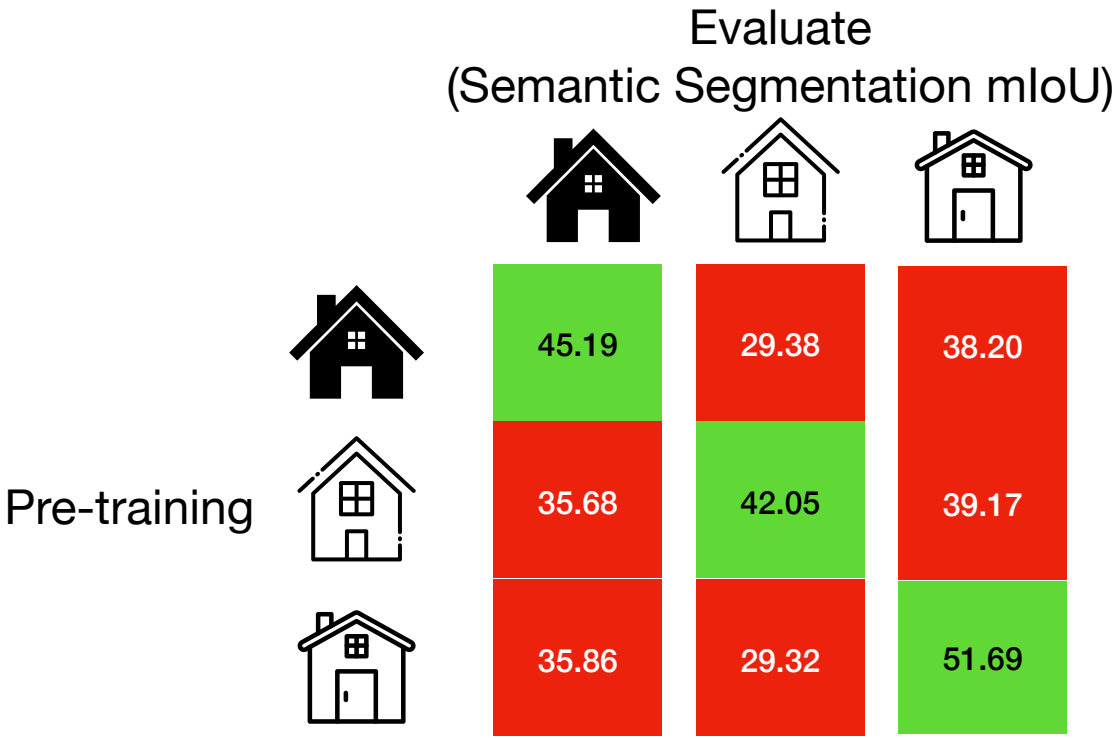
1. Bachmann*, Kar*, Mizrahi* et al. 2024

Analysis

Analysis 1. Is “specialization” actually happening?

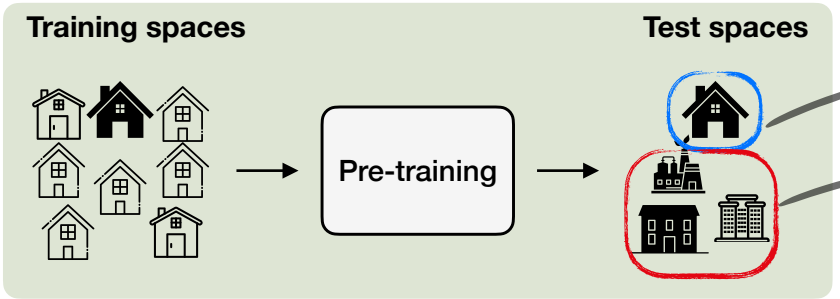
Analysis

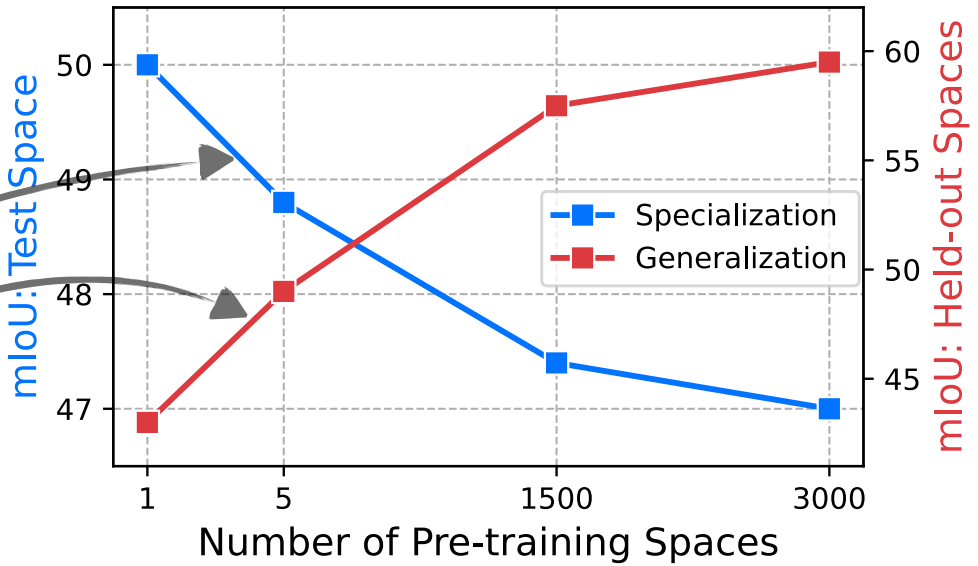
Analysis 1. Is “specialization” actually happening?



Analysis

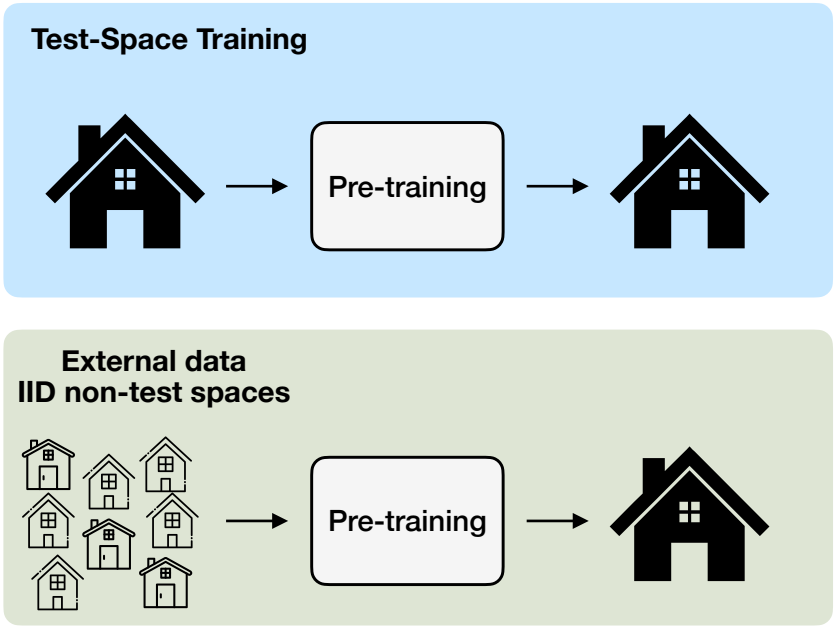
Analysis 2. Specialization-generalization tradeoff

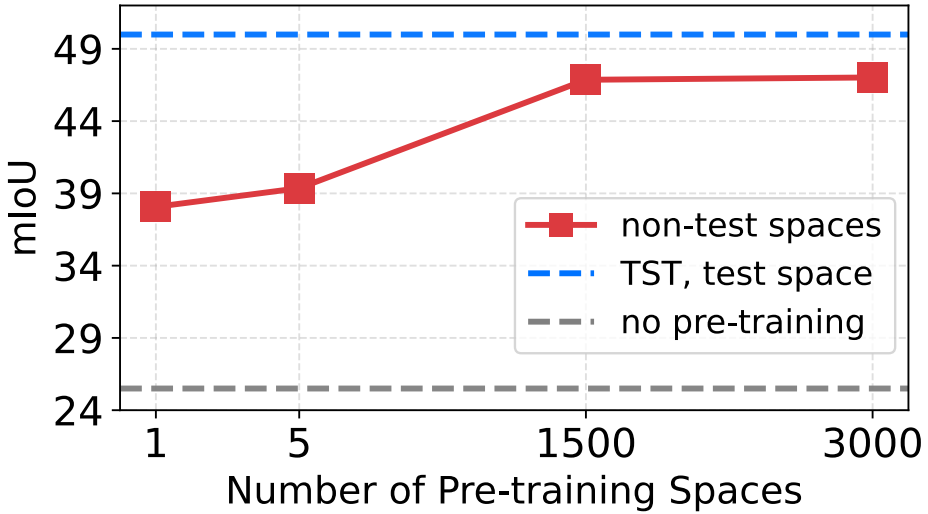




Analysis

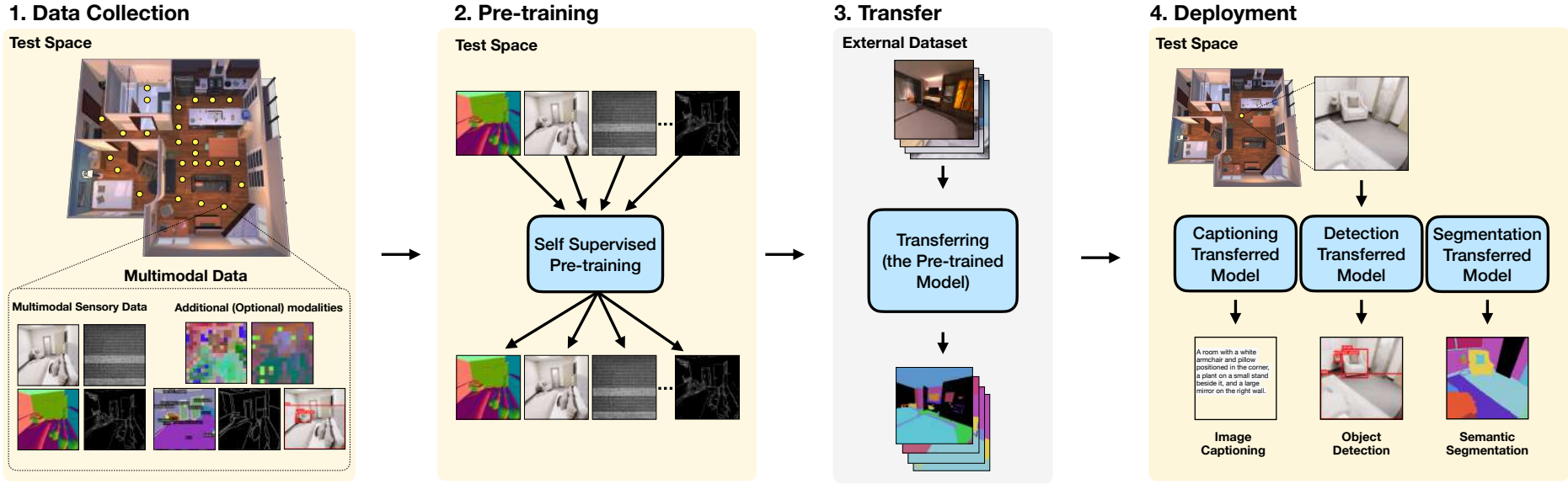
Analysis 3. How much external data is the test-space data worth?





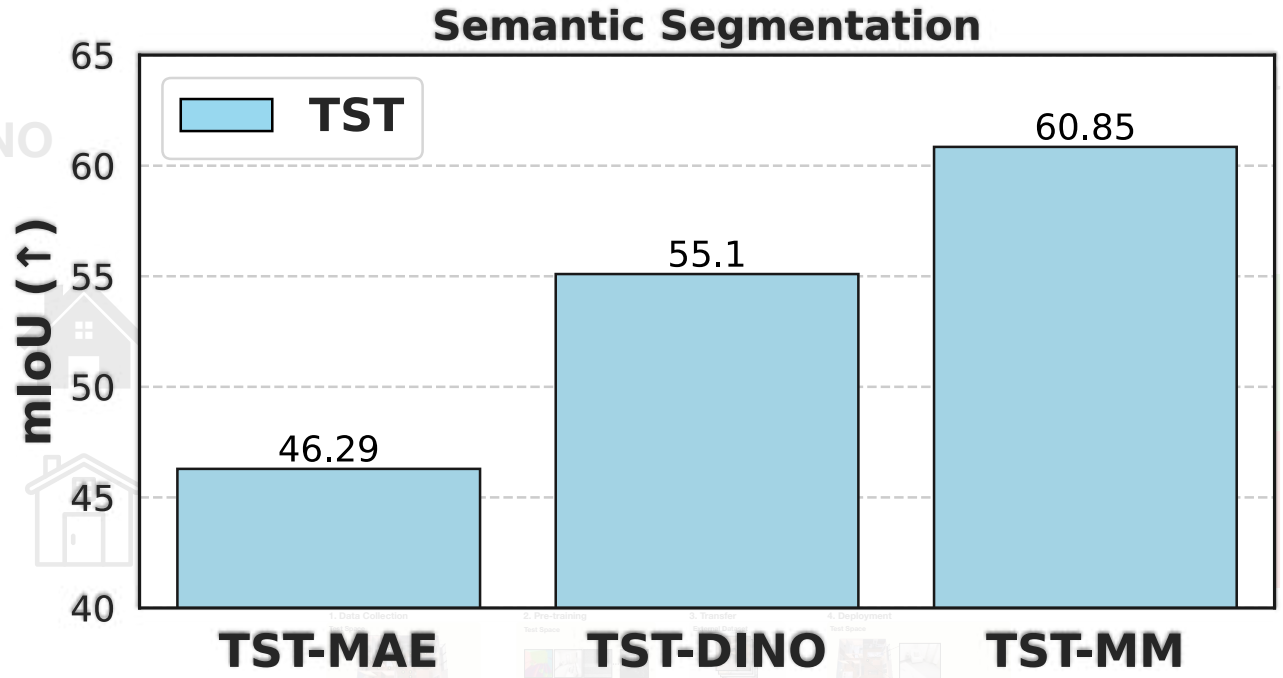
Analysis

Analysis 4. What about other self-supervised objectives?



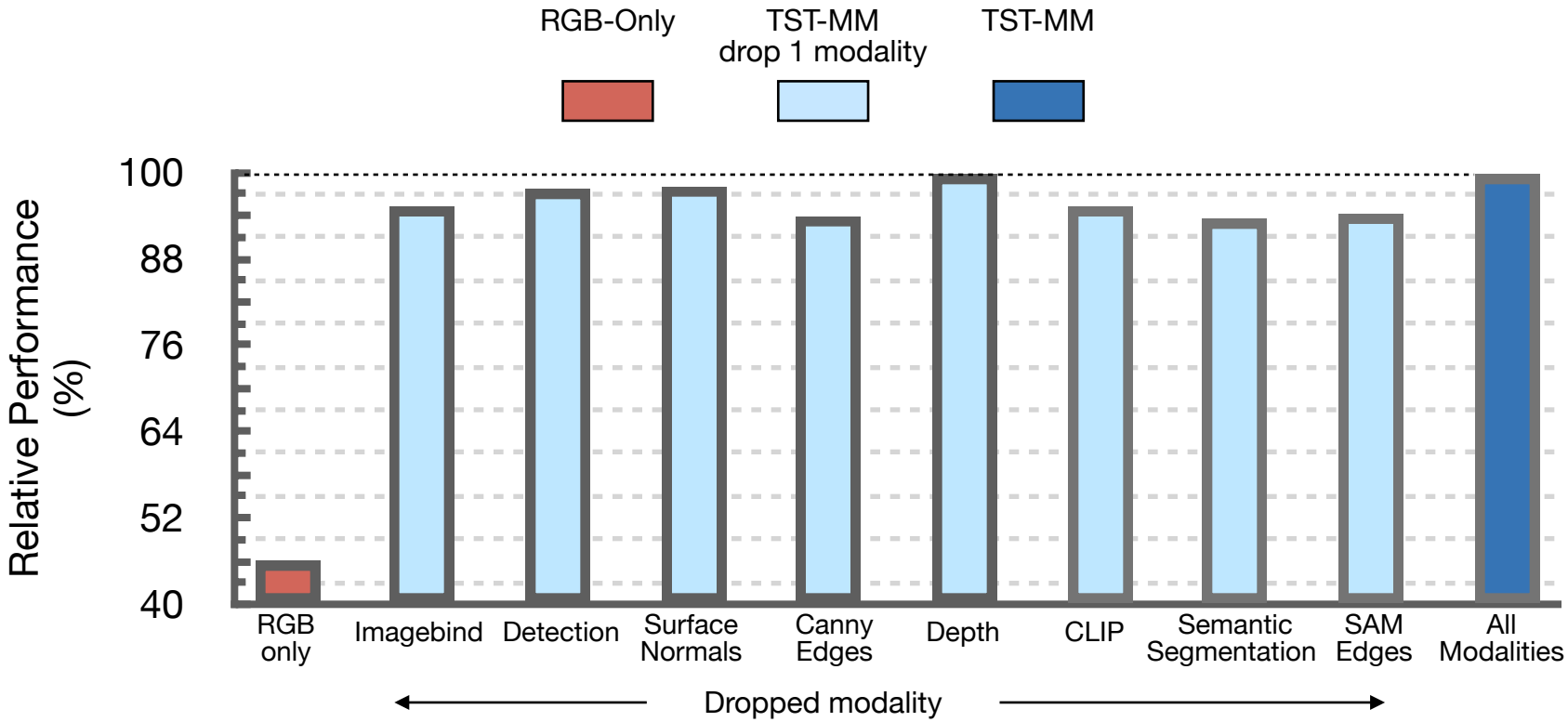
Analysis

Analysis 4. What about other self-supervised objectives?

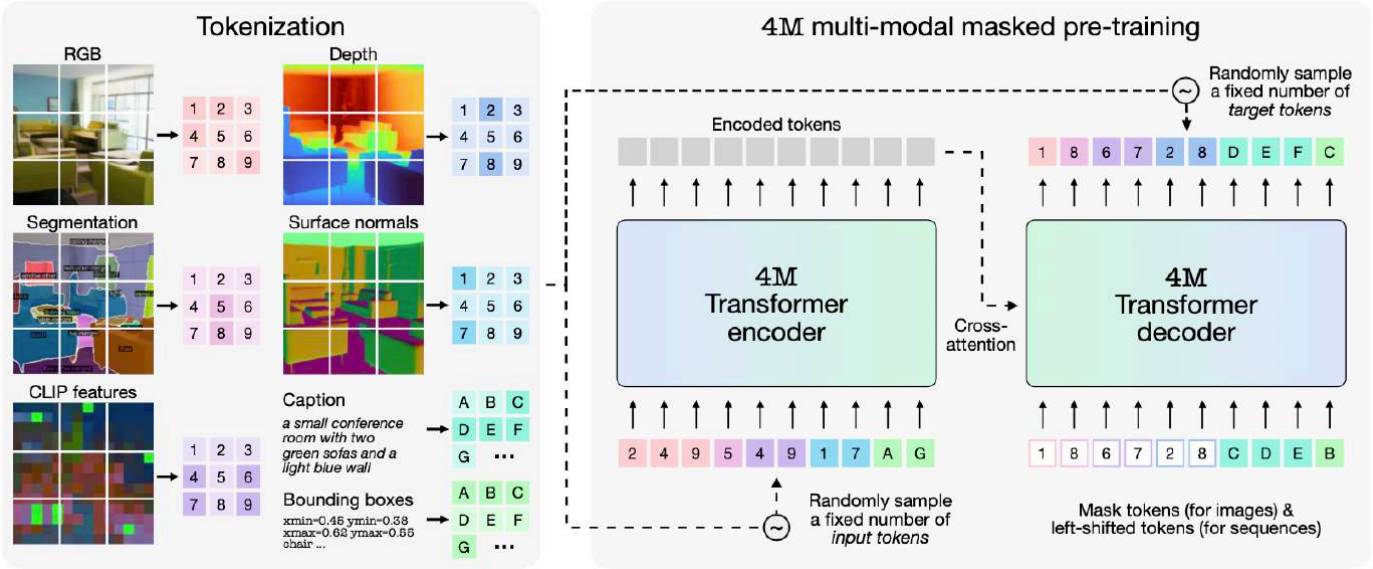


Analysis

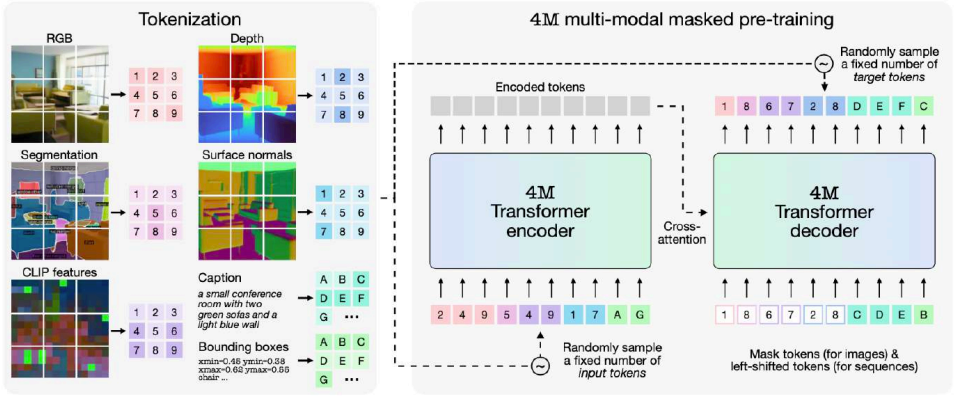
Analysis 5. Is one modality doing most of the job?



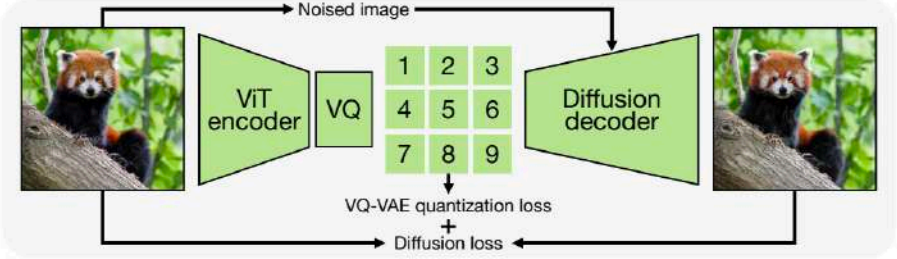
Tokenization



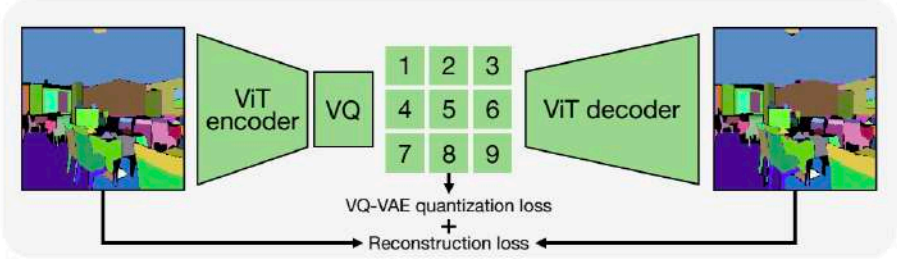
Tokenization



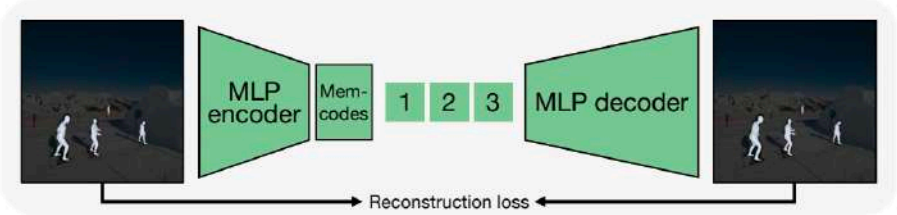
Spatial discrete VAE with diffusion decoder: RGB, normal, depth, edges



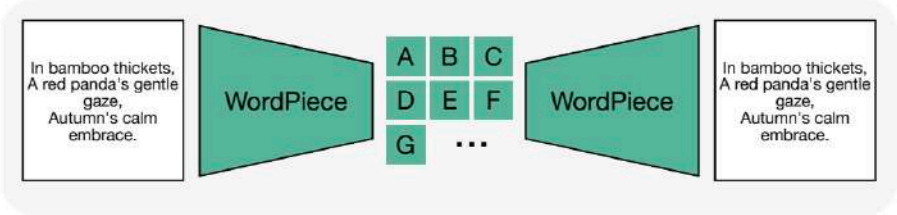
Spatial discrete VAE: Segmentation, CLIP, DINOv2, ImageBind, SAM inst.



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



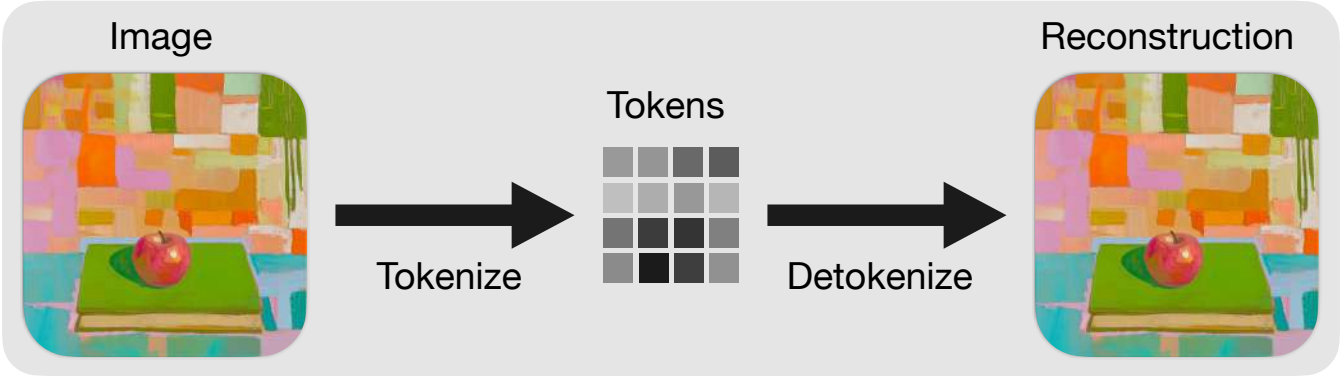
Sequence tokenizer: Text, bounding boxes, metadata, color palette



Token-based generation

Common way to perform generation:

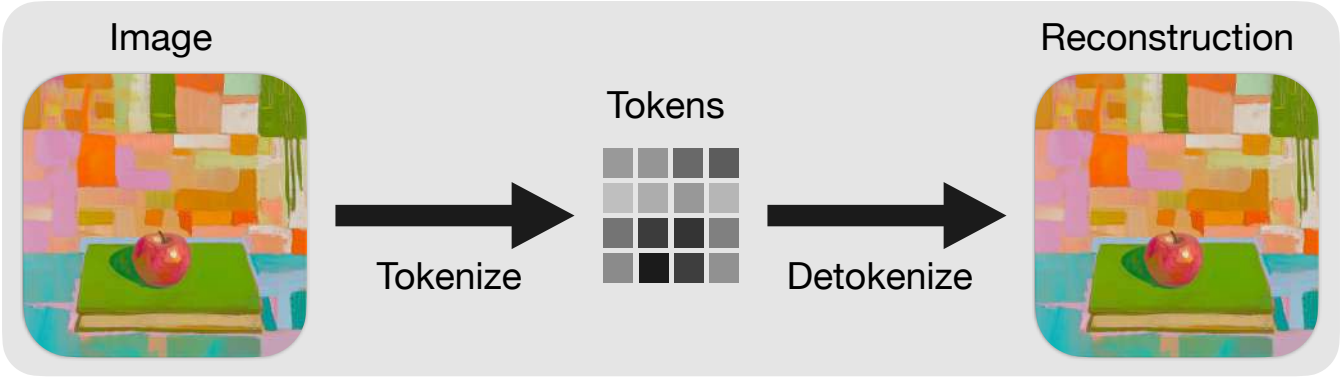
Stage 1: Train *tokenizer* with autoencoding objective



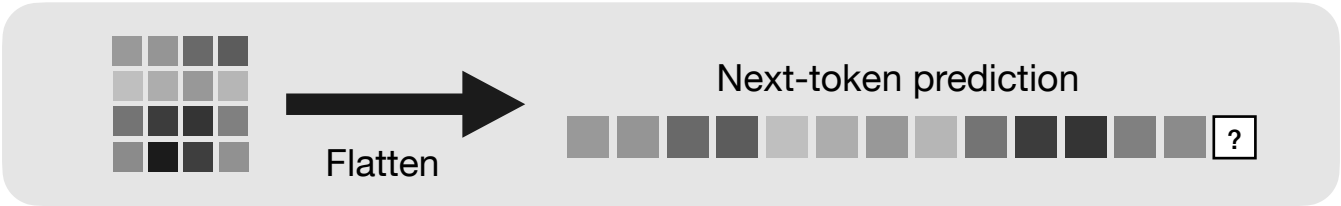
Token-based generation

Common way to perform generation:

Stage 1: Train *tokenizer* with autoencoding objective



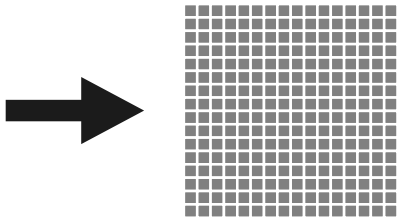
Stage 2: Perform *next-token prediction* on image tokens



Token-based generation

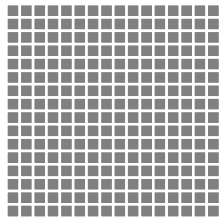
Common 2D grid tokenizers images represented with **a *fixed* number of tokens, regardless of complexity**.

e.g. 256x256 pixels



e.g. 16x16 tokens

e.g. 256x256 pixels

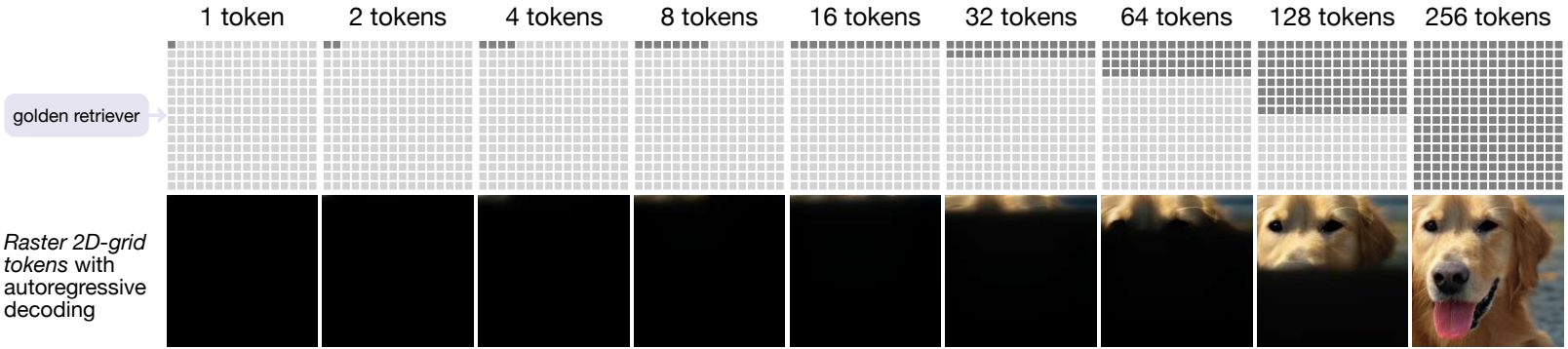


e.g. 16x16 tokens

Token-based generation

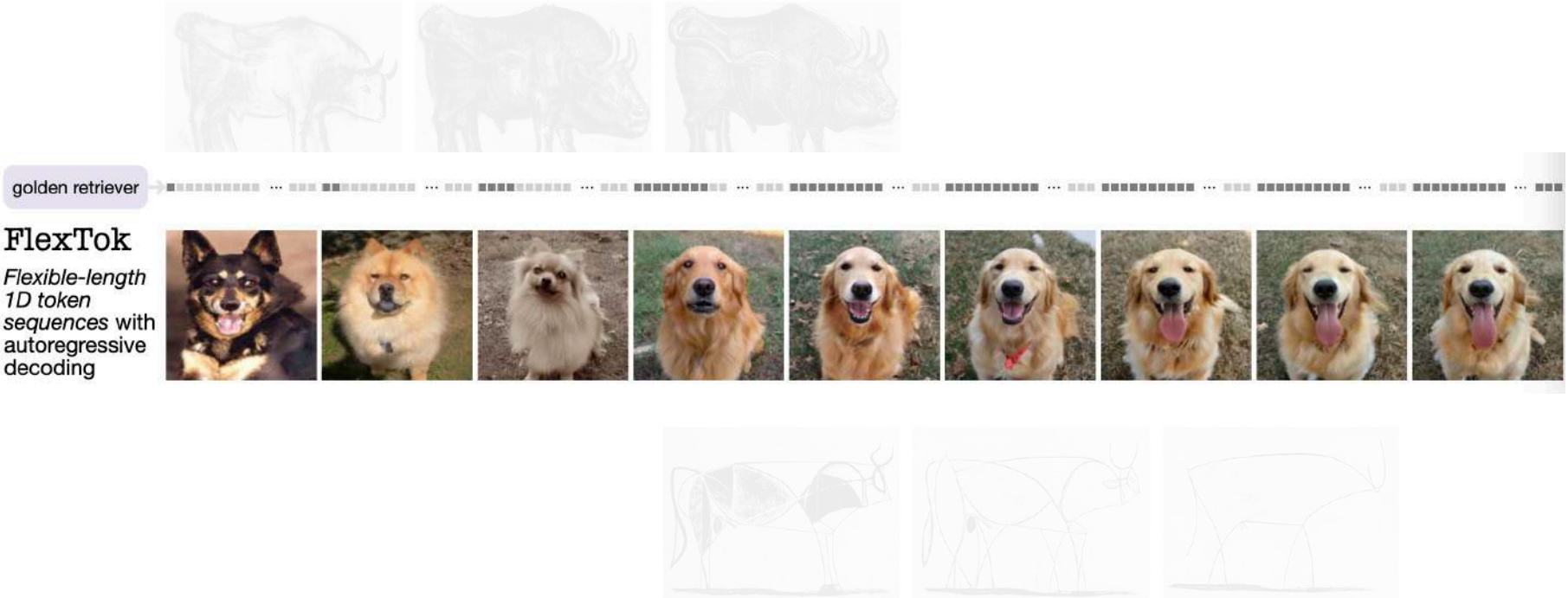
Common 2D grid tokenizers images represented with **a *fixed* number of tokens, regardless of complexity**.

Autoregressive generation is performed ~patch-by-patch.



Abstraction — Compression

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



[Le Taureau, Pablo Picasso]

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

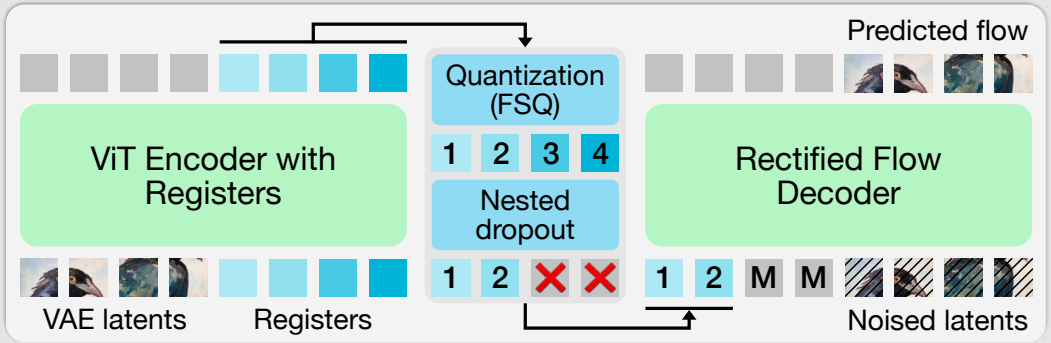
Details

FlexTok method

Overview

Stage 1

FlexTok tokenizer training

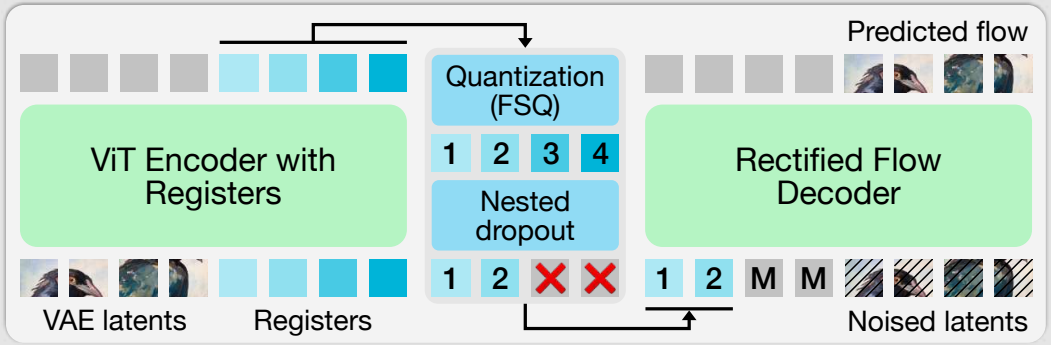


FlexTok method

Overview

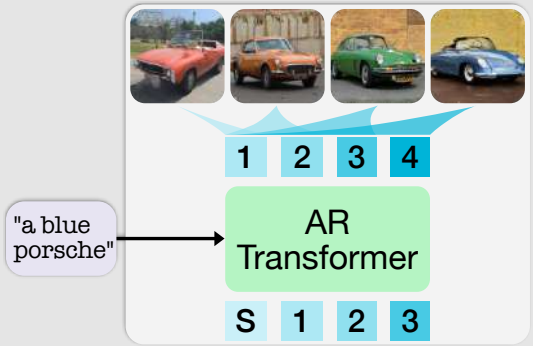
Stage 1

FlexTok tokenizer training



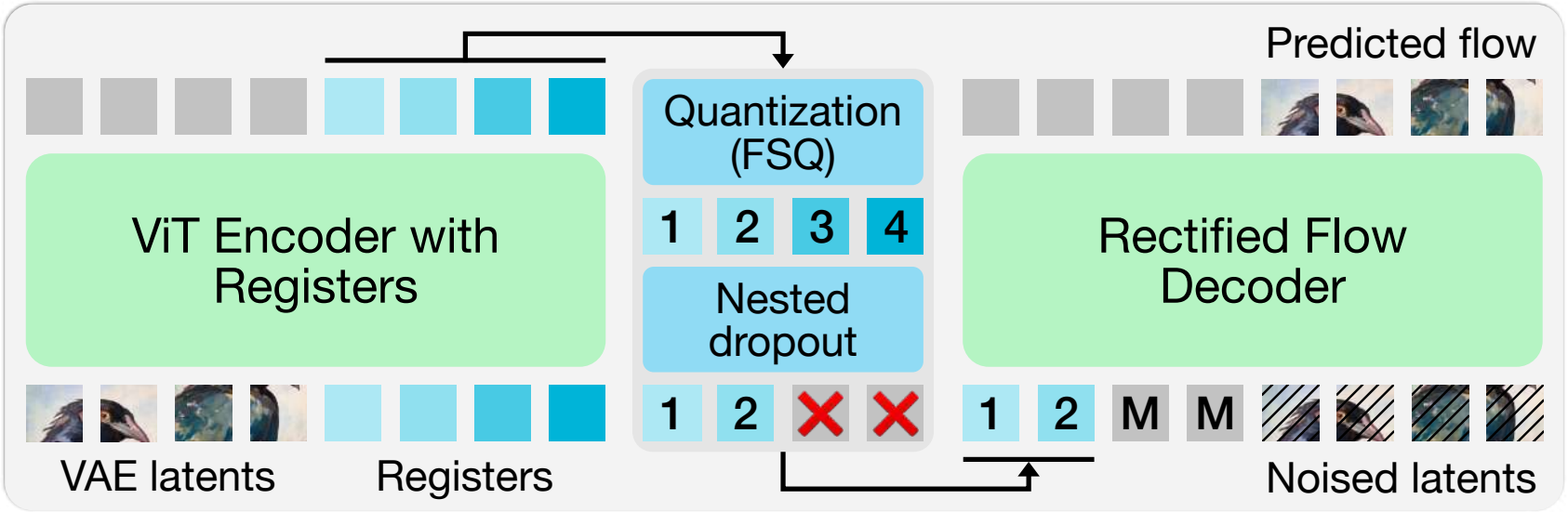
Stage 2

Autoregressive generation using FlexTok tokens



FlexTok method

Stage 1: Tokenizer training



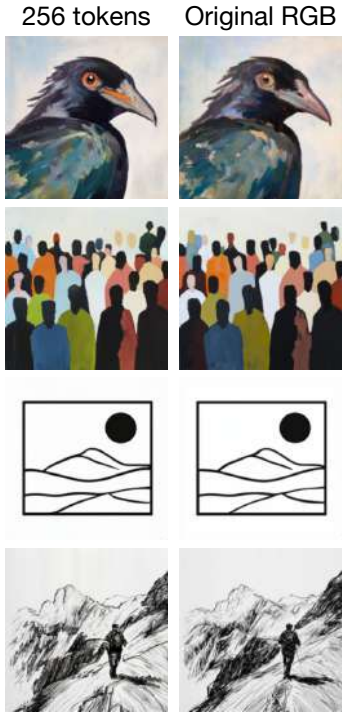
FlexTok reconstruction

Specify a coarse-to-fine "visual vocabulary"

Original RGB

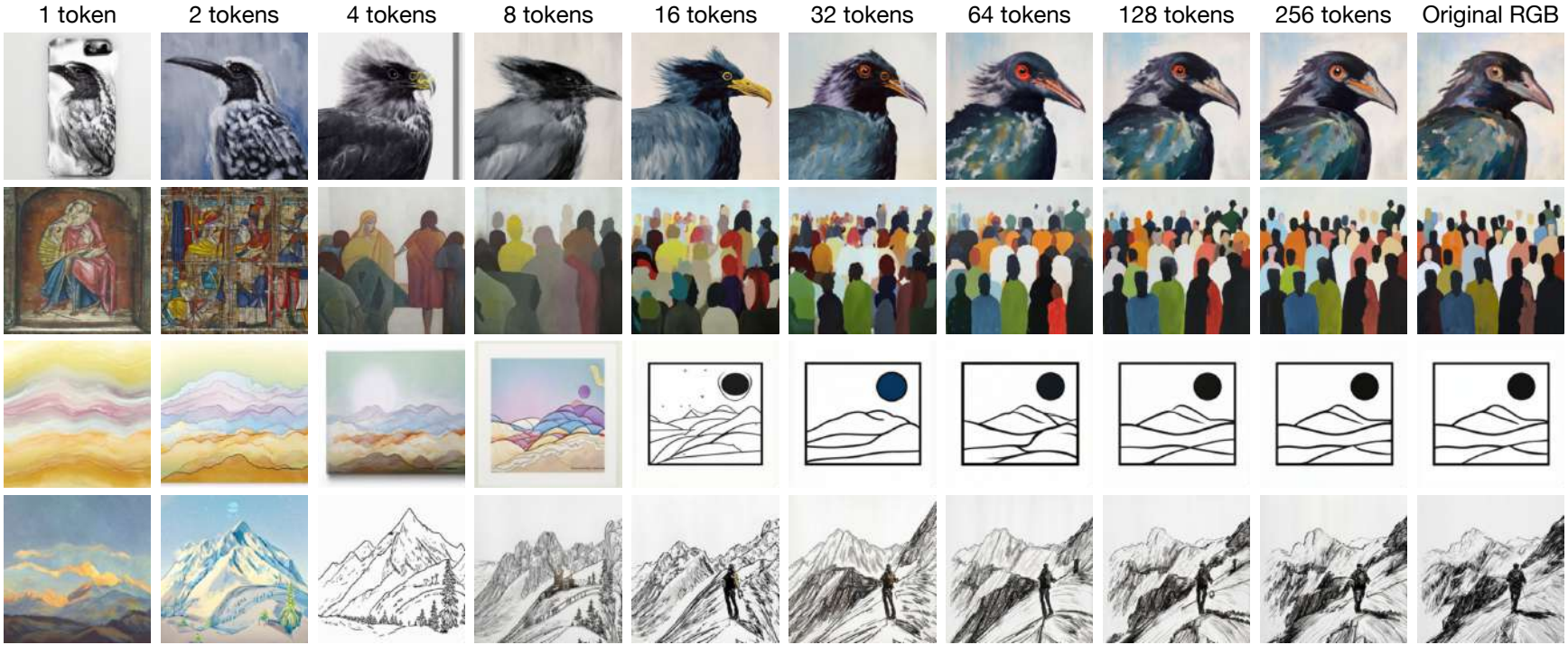
FlexTok reconstruction

Specify a coarse-to-fine "visual vocabulary"



FlexTok reconstruction

Specify a coarse-to-fine "visual vocabulary"

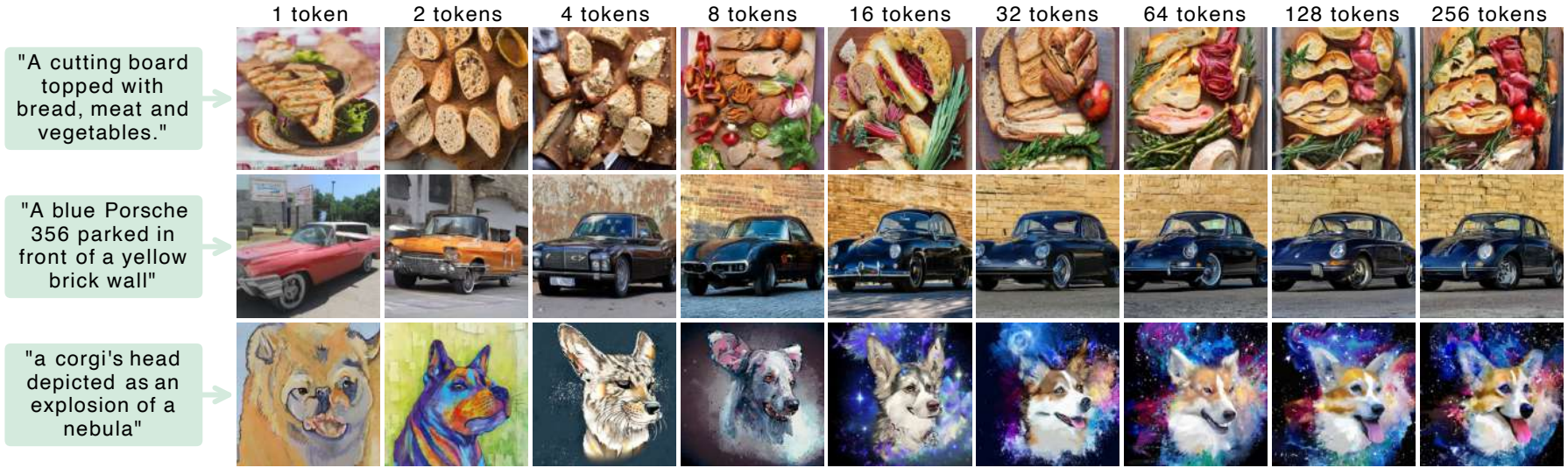


Autoregressive generation

Class-to-image

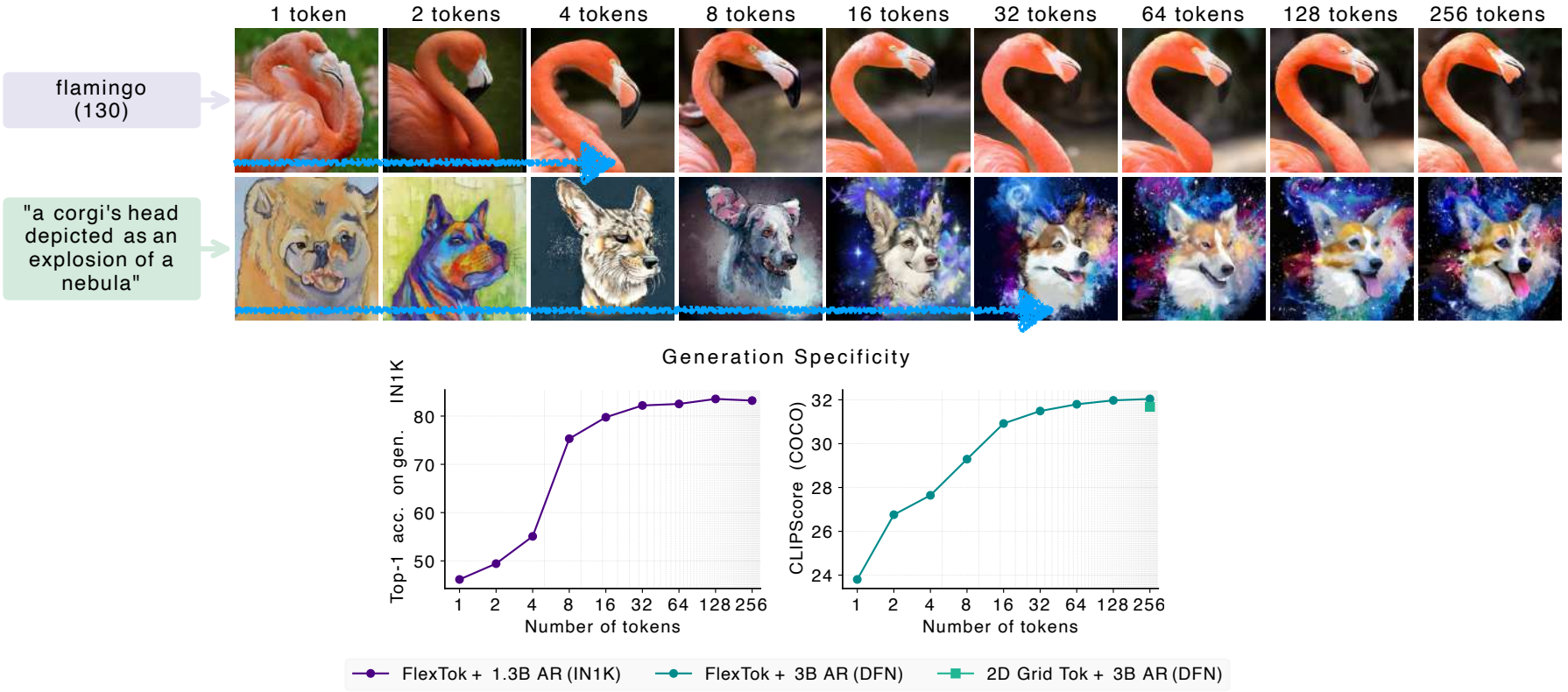
Autoregressive generation

Text-to-image



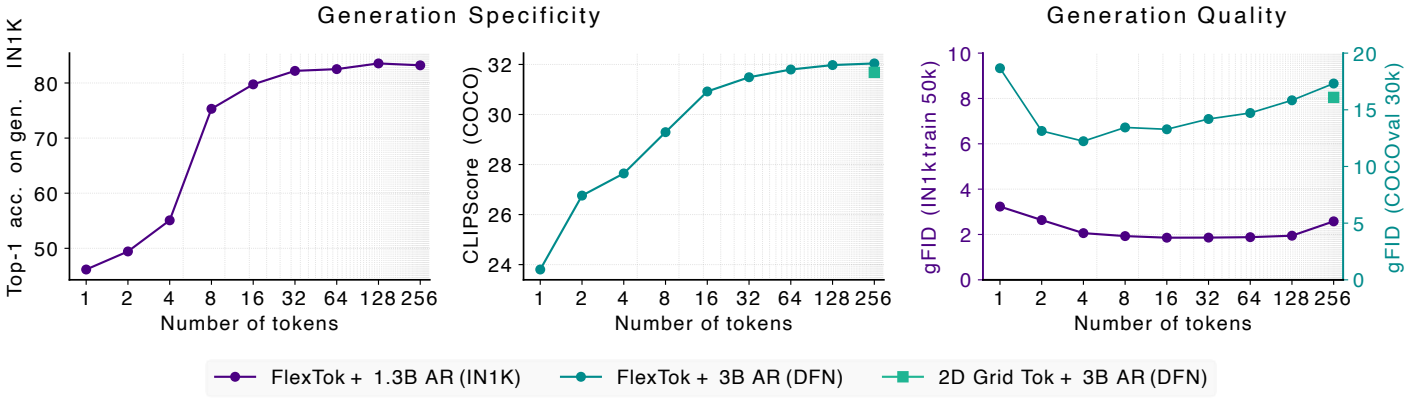
Autoregressive generation

Adaptive conditioning alignment



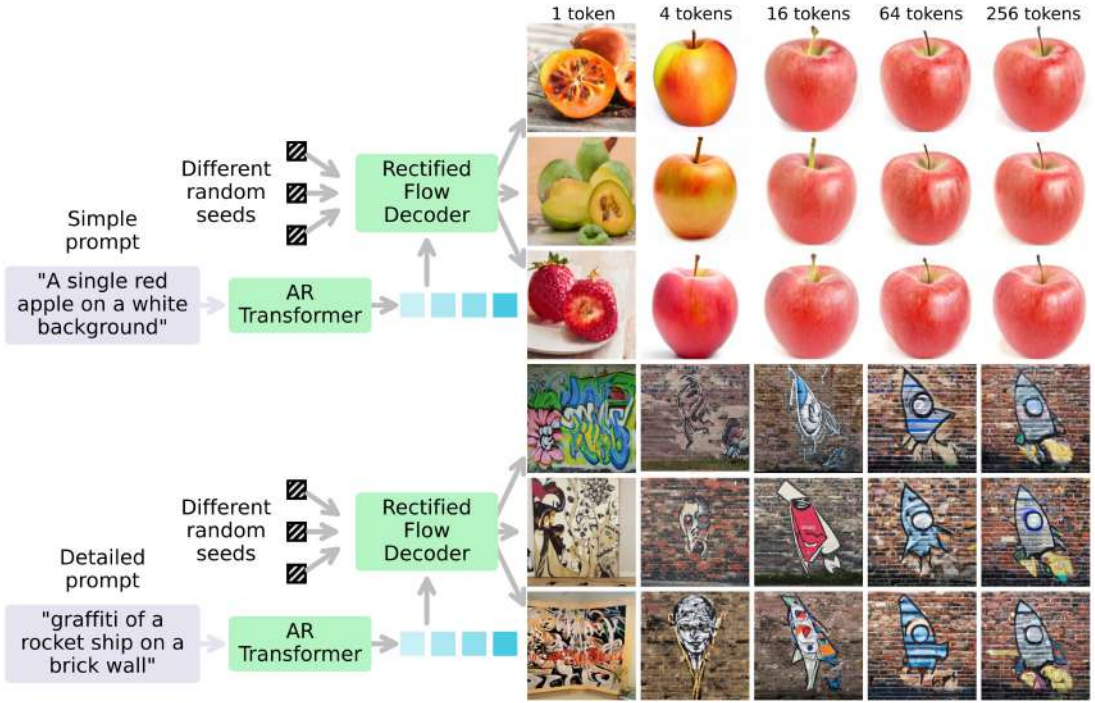
Autoregressive generation

Adaptive conditioning alignment



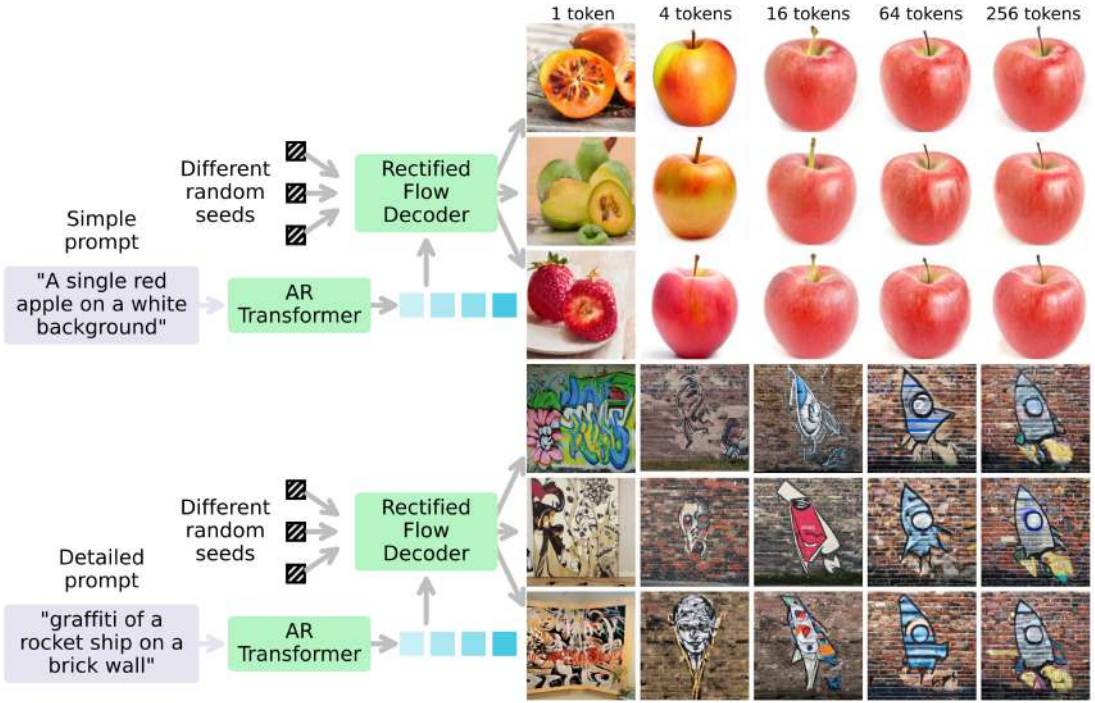
Autoregressive generation

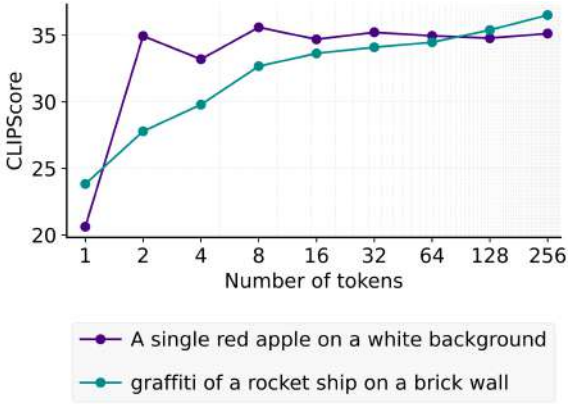
Image generation with simple and complex prompts



Autoregressive generation

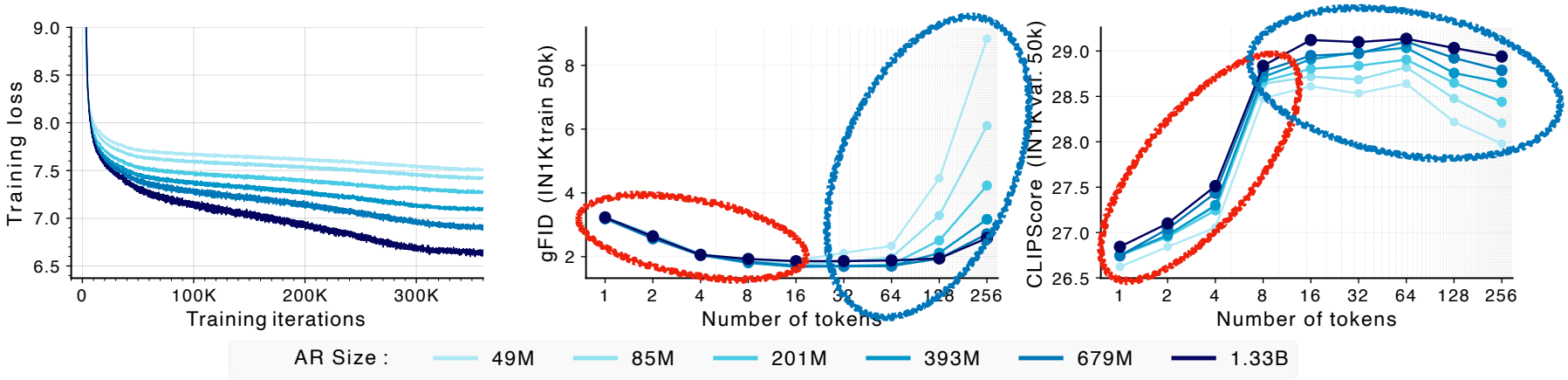
Image generation with simple and complex prompts





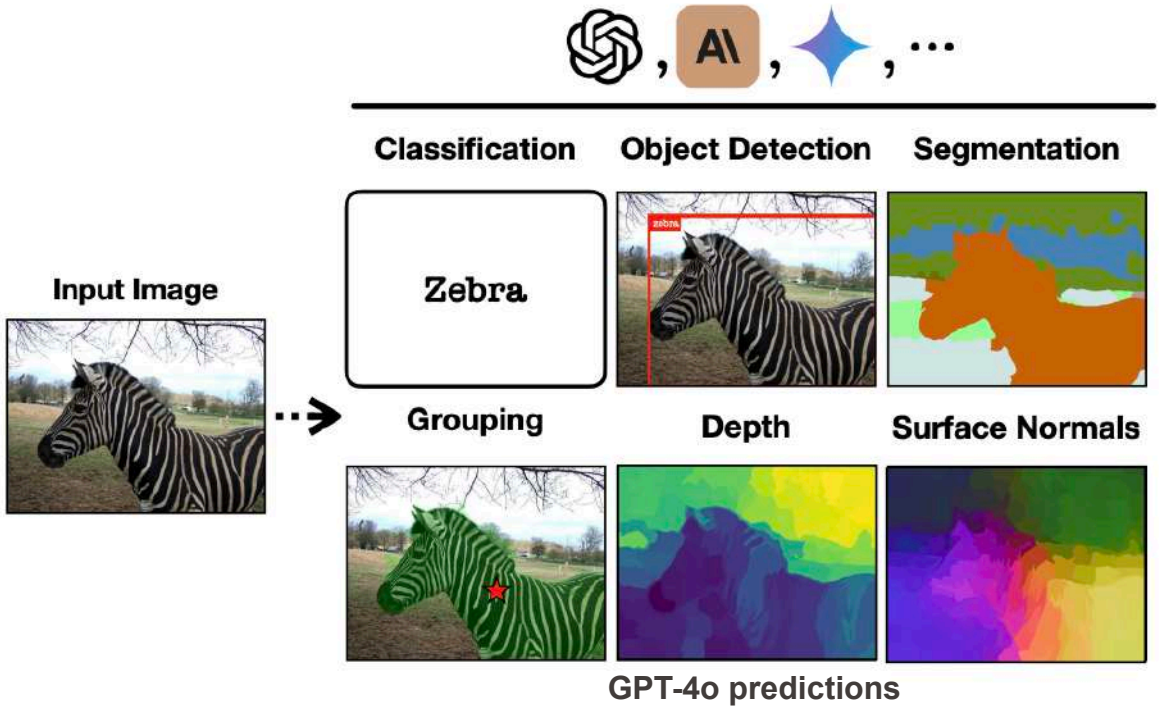
Scaling autoregressive generation

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



Benchmarking popular multimodal FMs

How well does GPT-4o understand vision?

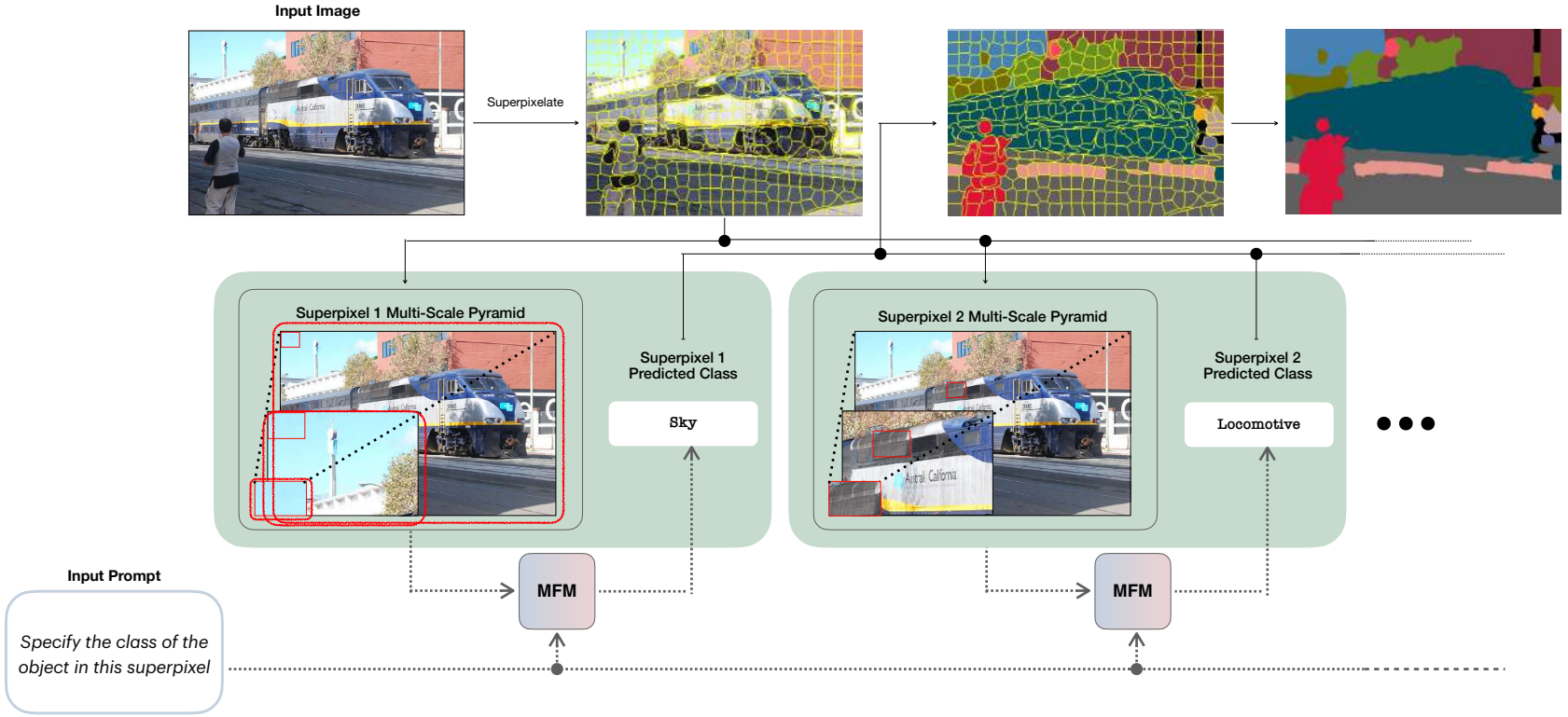


https://fm-vision-evals.epfl.ch/

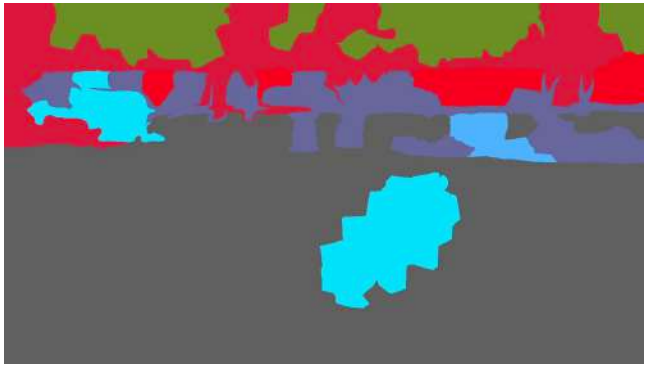
How Well Does GPT-4o Understand Vision? Evaluating Multimodal Foundation Models on Standard Computer Vision Tasks, Ramachandran, Garjani, Bachmann, Atanov *, Kar *, Zamir *. arxiv 2025.

How to extract a non-textual task from chatbots?

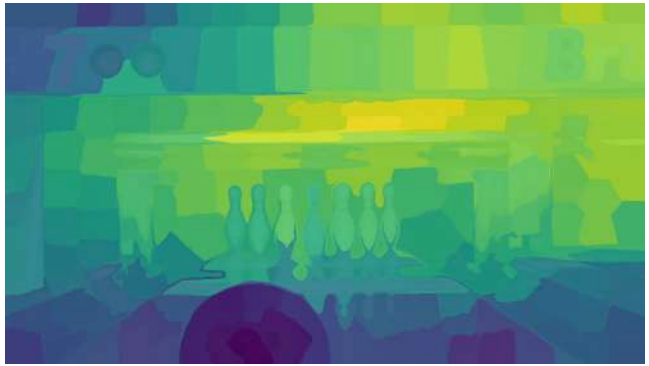
e.g., semantic segmentation from chatGPT?

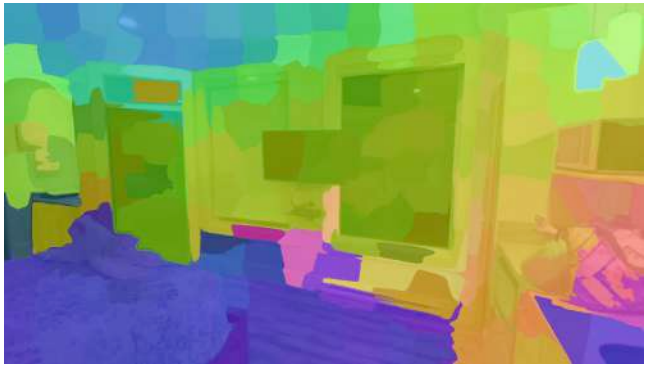


Predictions
(GPT-4o)

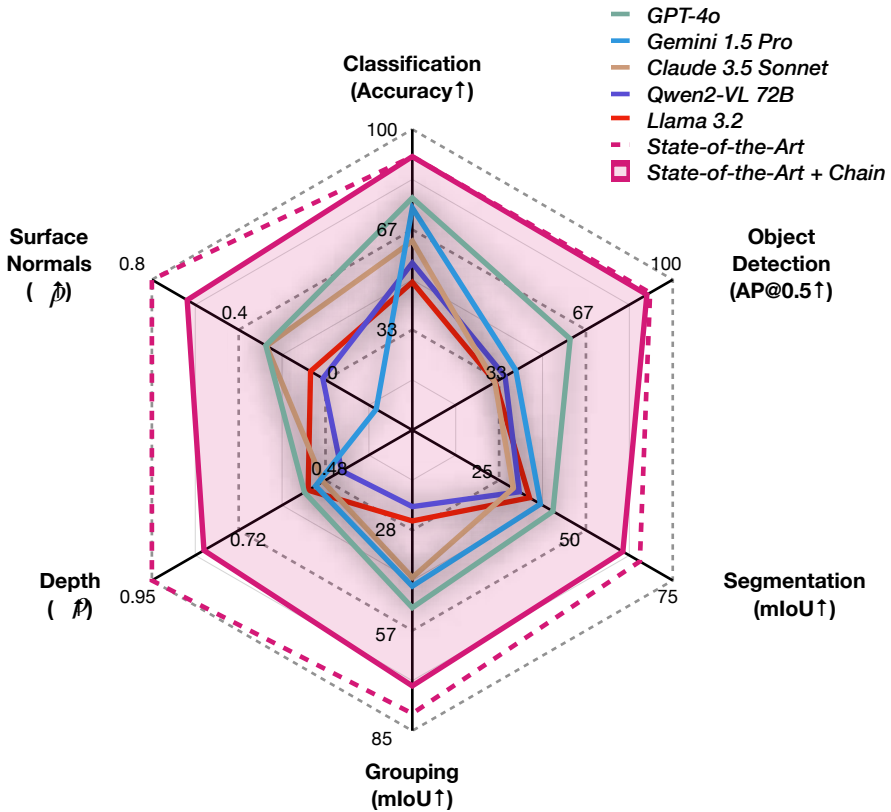


Predictions
(GPT-4o)





The results



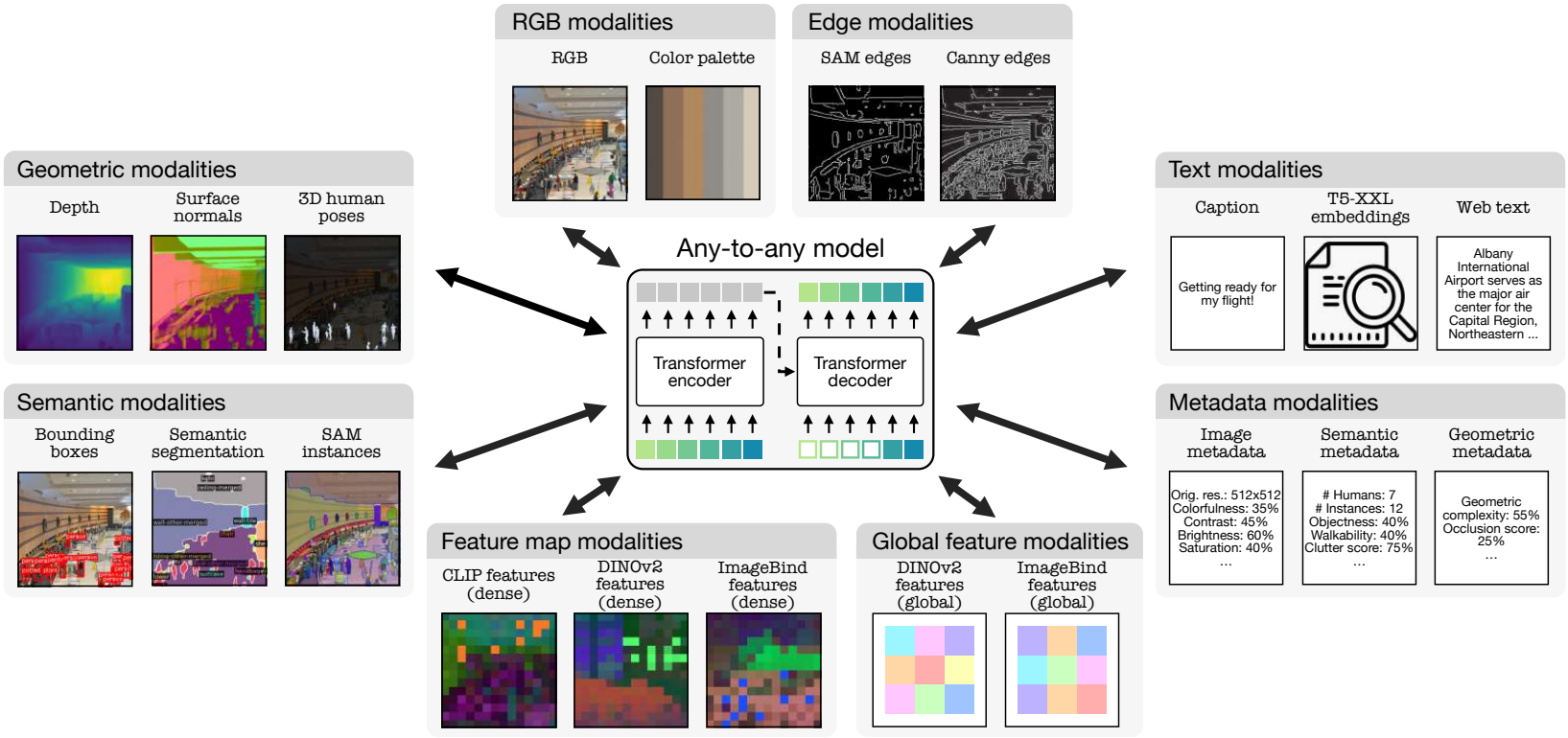
Key Takeaways

- **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.

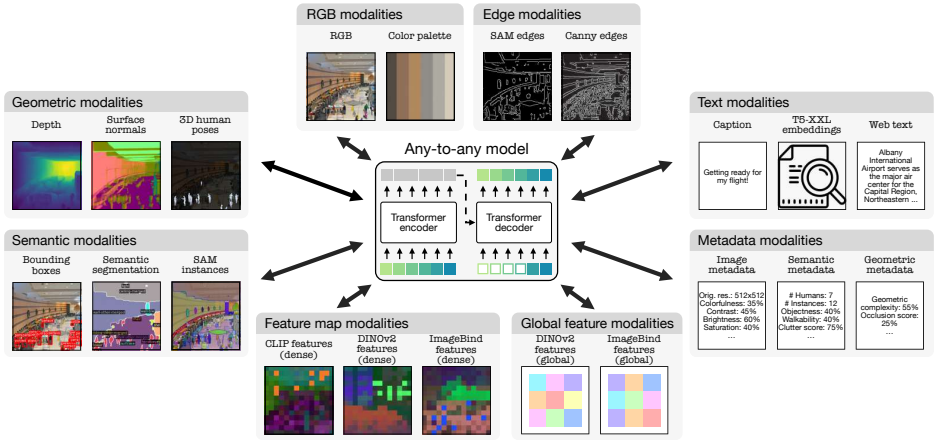
Summary



- 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.

Summary

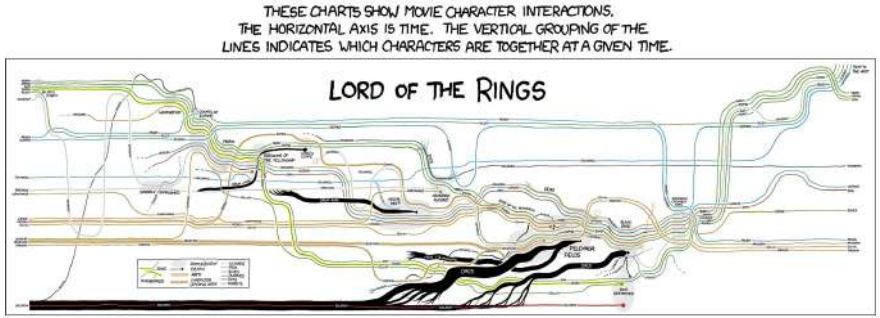
- A scalable versatile **multi-modal/Multi-task foundation model**
- Ultimate goal: **a grounded world model. A “foundation”.**

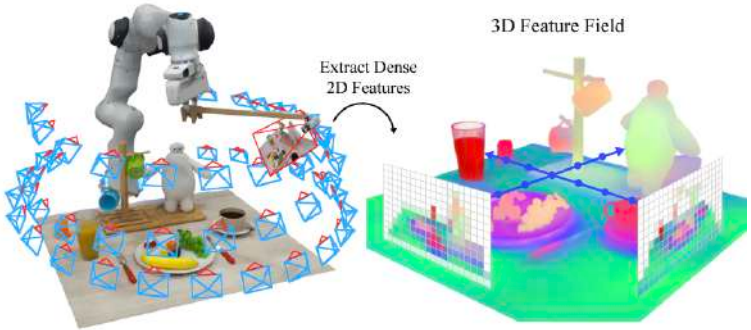


- 4M: Massively Multimodal Masked Modeling, Mizrahi, Bachmann, Kar, Yeo, Gao, Dehghan, Zamir. NeurIPS 2023.
- MultiMAE: Multi-Modal Multi-Task Masked Autoencoders, Bachmann, Mizrahi, Atanov, Zamir. ECCV 2022
- Craik, Kenneth. The nature of explanation. Vol. 445. CUP Archive, 1967.

Summary

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



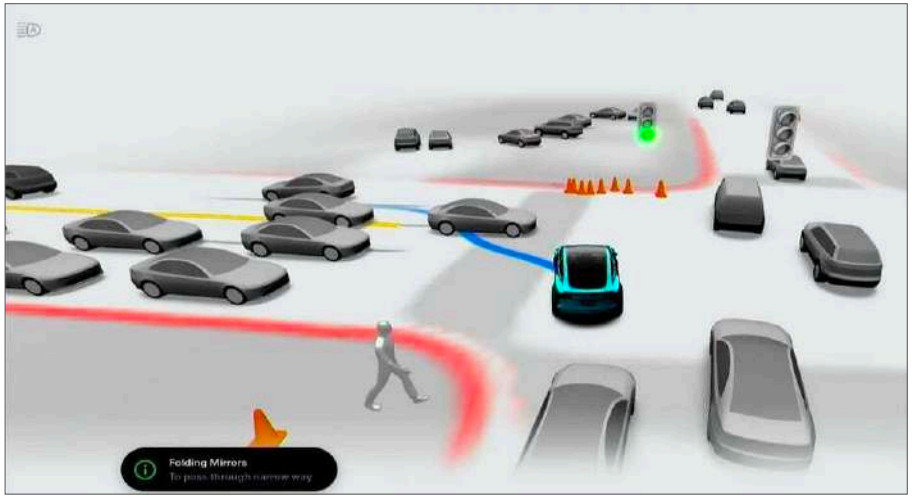


Shen et al., 2023

- 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.

Summary

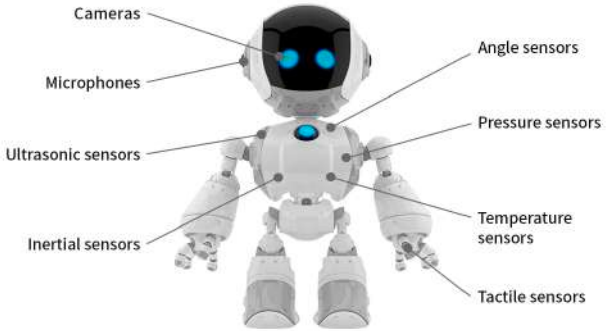
- A scalable versatile **multi-modal/Multi-task foundation model**
- Ultimate goal: **a grounded world model. A “foundation”.**
- (Long-form) Video understanding
- Learning in **higher-level spaces**

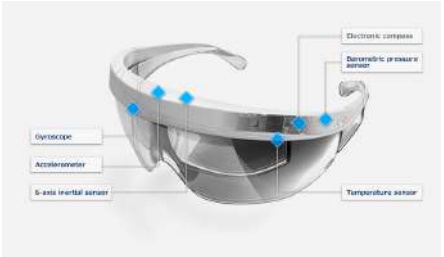


• 4M: Massively Multimodal Masked Modeling, Mizrahi, Bachmann, Kar, Yeo, Gao, Dehghan, Zamir. NeurIPS 2023.
• MultiMAE: Multi-Modal Multi-Task Masked Autoencoders, Bachmann, Mizrahi, Atanov, Zamir. ECCV 2022
• 4M-21: An Any-to-Any Vision Model for Tens of Tasks and Modalities, Bachmann, Kar, Mizrahi,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

Summary

- 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.

Summary

- 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.

Questions?


Roman Bachmann


David Mizrahi


Oguzhan Kar

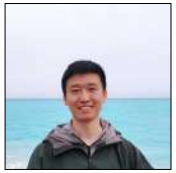

Ali Garjani


Mingfei Gao

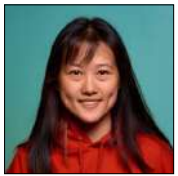

David Griffiths


Sogand Salehi


Andrei Atanov


Jiawei Fu


Rishubh Singh


Isabella Yu


Andrew Spielberg


Jiming Hu


Teresa Yeo


Afshin Dehghan


Amir Zamir

Multimodal Learning

https://4m.epfl.ch/

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

https://viper.epfl.ch/

https://amirzamir.com/