Scaling the Modalities in Multimodal Foundation Models

Oğuzhan Fatih KAR

Visual Intelligence and Learning Lab (VILAB)
Advisor: Prof. Amir Zamir

12 May 2025
PhD Public Defense

EPFL Swiss Federal Institute of Technology

Our experience of the world is multimodal

Our experience of the world is multimodal

Our experience of the world is multimodal

Our experience of the world is multimodal

Our experience of the world is multimodal

Our experience of the world is multimodal

Multimodal Learning

Fundamental to human perceptual cognition

Credits: Amir Zamir, CS 503: Visual Intelligence.
Test-Space Training: Self-Supervised Specialization of Vision Models to the Test Environment, Sing et al., 2025

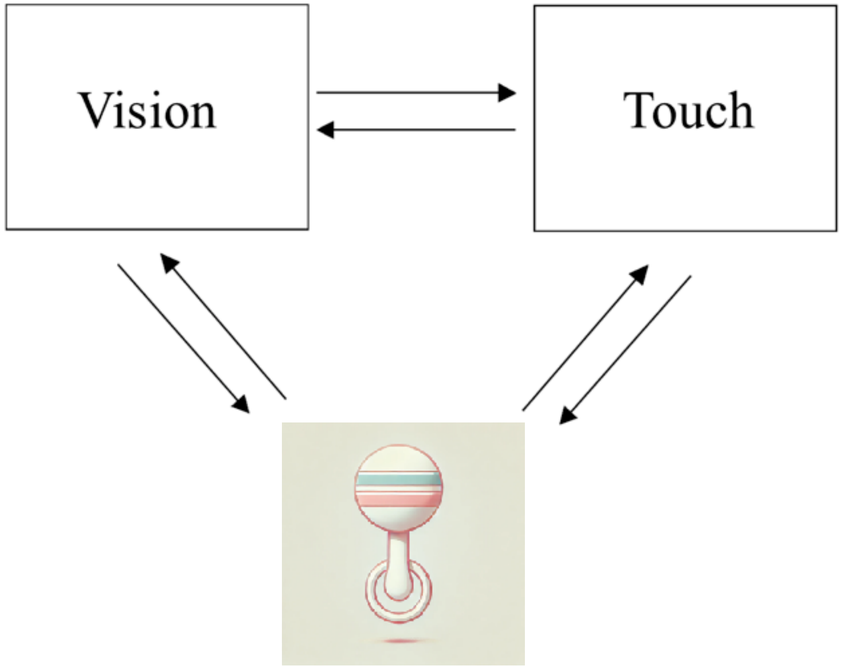
Multimodal Learning

The Development of Embodied Cognition:
Six Lessons from Babies, Smith & Gasser 2005:

- *Multiple overlapping and time-locked sensory systems enable the developing system to* ***educate itself.***

Credits: Amir Zamir, CS 503: Visual Intelligence.
Test-Space Training: Self-Supervised Specialization of Vision Models to the Test Environment, Sing et al., 2025

Multimodal Learning



The Development of Embodied Cognition:
Six Lessons from Babies, Smith & Gasser 2005:

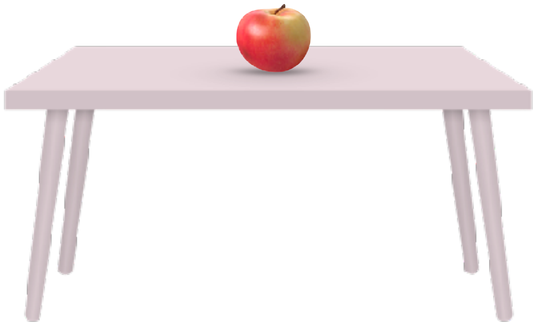
- *Multiple overlapping and time-locked sensory systems enable the developing system to* ***educate itself.***

- *Cross-modal learning*
- *Sensory fusion*

Credits: Amir Zamir, CS 503: Visual Intelligence.
Test-Space Training: Self-Supervised Specialization of Vision Models to the Test Environment, Sing et al., 2025

Multimodal Foundation Model (MFM)

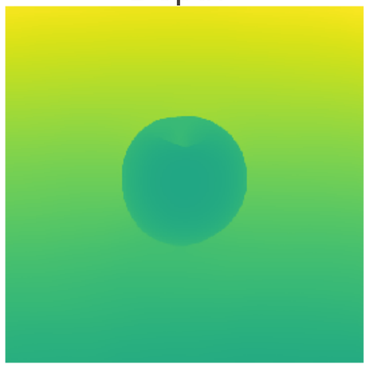
Multimodal Foundation Model (MFM)



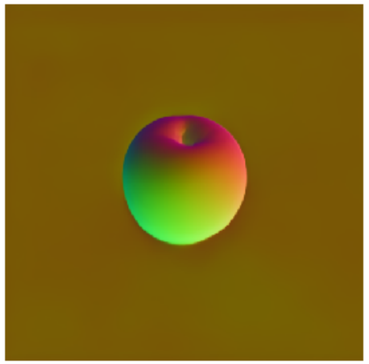
The world

RGB Image

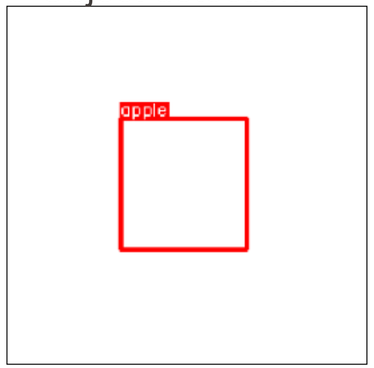
Depth



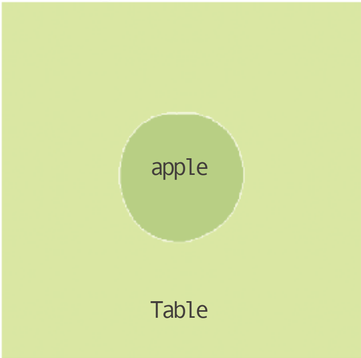
3D Surface normals



Object detection

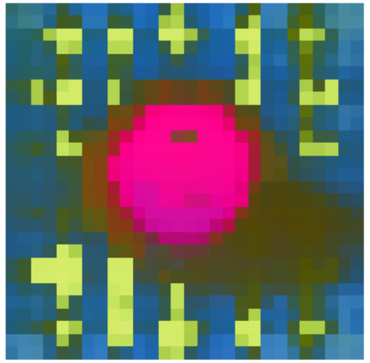


Segmentation



Text

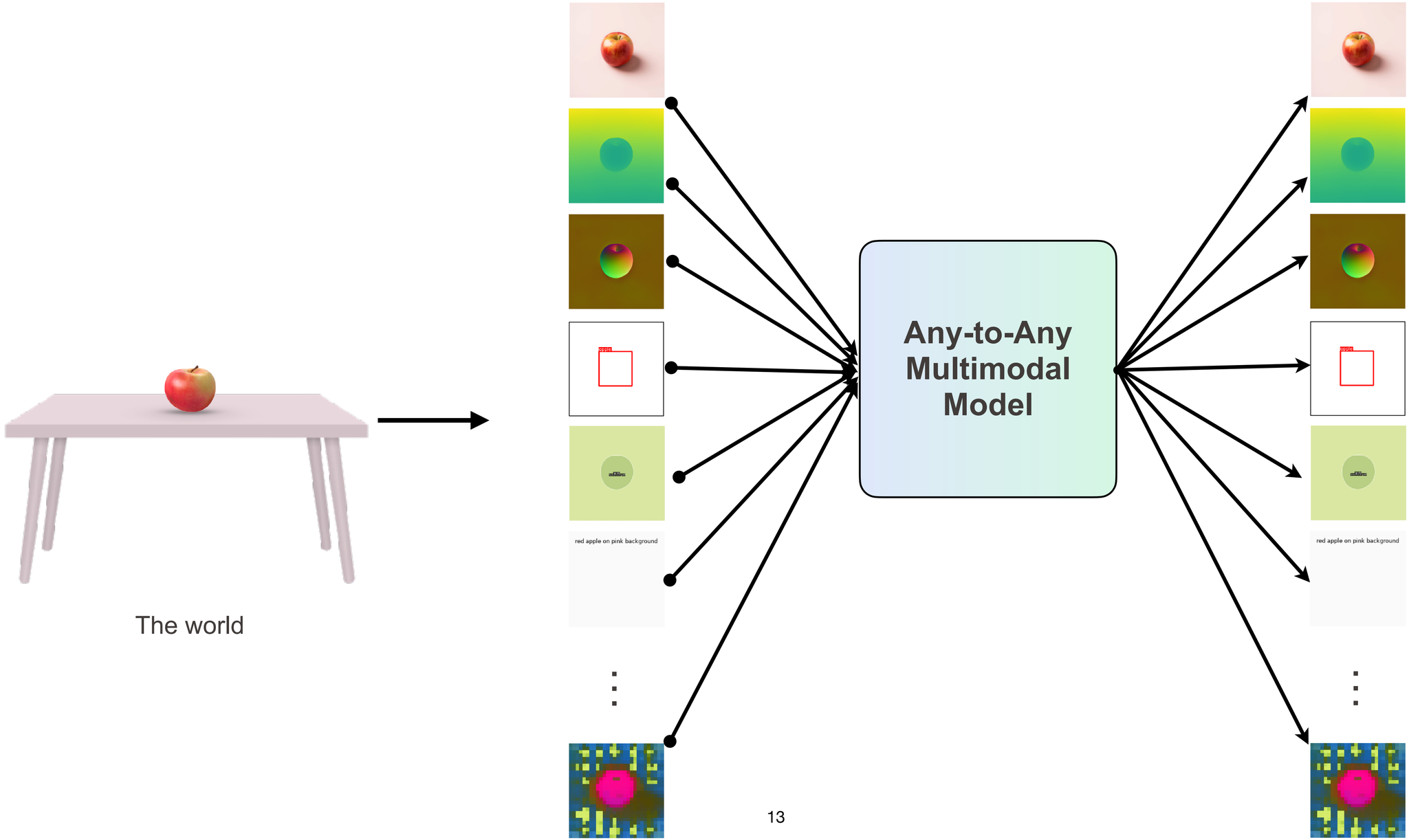
CLIP features



more

. . .

Multimodal Foundation Model (MFM)



But it is not trivial…

Loss balancing

- *[Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics, Kendall et al., 2017]*
- *[GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks, Chen et al., 2017]*
- *…*

Which tasks and how?

- *[Taskonomy: Disentangling Task Transfer Learning, Zamir et al., 2018]*
- *[Which Tasks Should Be Learned Together in Multi-task Learning?, Standley et al., 2020]*
- *…*

Multimodal fusion

- *[Efficient Low-rank Multimodal Fusion with Modality-Specific Factors, Liu et al. 2018]*
- *[Attention Bottlenecks for Multimodal Fusion, Nagrani et al., 2021]*
- *…*

Task diversity & consistency

- *[UberNet : Training a 'Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory, Kokkinos, 2016]*
- *[Robust Learning Through Cross-Task Consistency, Zamir et al. 2020]*
- *…*

Core challenges addressed in this thesis

Core challenges addressed in this thesis

- How to handle the data requirements of different modalities?
  - **Handling Modalities**
- How to build a scalable training framework?
  - **Multimodal Training**
- How to evaluate and track the progress?
  - **Analysis**

Core challenges addressed in this thesis

- How to handle the data requirements of different modalities?
  - **Handling Modalities**
- How to build a scalable training framework?
  - **Multimodal Training**
- How to evaluate and track the progress?
  - **Analysis**

Core challenges addressed in this thesis

- How to handle the data requirements of different modalities?
  - **Handling Modalities**
- How to build a scalable training framework?
  - **Multimodal Training**
- How to evaluate and track the progress?
  - **Analysis**

Publication map & Thesis outline

Part I: Handling Modalities | Part II: Multimodal Training | Part III: Analysis

Robustness Mechanisms

- Cross-task Consistency (CVPR 2020, Best Paper Nominee)
- Cross-domain Ensembles (ICCV 2021, Oral)
- Rapid Network Adaptation (ICCV 2023)
- OOD Generalization via Diversity (ICLR 2024)

Pretraining Framework

- 4M: Multimodal Modeling (NeurIPS 2023, Spotlight)

Internet-based Generalists

- Evaluating MFMs on Vision (arXiv 2025)

Architecture

Status check

Training the Unified Model

- 4M-21: Any-to-Any Models (NeurIPS 2024) [Demo]

Real-world Specialization

- Test-space Multimodal Training (arXiv 2025)

Data

Model

Robust Pseudolabelers & Vision Encoders

- 3D Common Corruptions (CVPR 2022, Oral) [Demo]
- Broadening Vision in VLMs (ECCV 2024, Oral)

Variable Tokenization

- Resampling Images via FlexTok (arXiv 2025)

Publication map & Thesis outline

Part I: Handling Modalities

Robustness Mechanisms

- Cross-task Consistency (CVPR 2020, Best Paper Nominee)
- Cross-domain Ensembles (ICCV 2021, Oral)
- Rapid Network Adaptation (ICCV 2023)
- OOD Generalization via Diversity (ICLR 2024)

Robust Pseudolabelers & Vision Encoders

- 3D Common Corruptions (CVPR 2022, Oral) [Demo]
- Broadening Vision in VLMs (ECCV 2024, Oral)

Part II: Multimodal Training

Pretraining Framework

- 4M: Multimodal Modeling (NeurIPS 2023, Spotlight)

Architecture

Training the Unified Model

- 4M-21: Any-to-Any Models (NeurIPS 2024) [Demo]

Data

Model

Part III: Analysis

Internet-based Generalists

- Evaluating MFMs on Vision (arXiv 2025)

Status check

Real-world Specialization

- Test-space Multimodal Training (arXiv 2025)

Variable Tokenization

- Resampling Images via FlexTok (arXiv 2025)

Publication map & Thesis outline

Part I: Handling Modalities

Robustness Mechanisms

- Cross-task Consistency (CVPR 2020, Best Paper Nominee)
- Cross-domain Ensembles (ICCV 2021, Oral)
- Rapid Network Adaptation (ICCV 2023)
- OOD Generalization via Diversity (ICLR 2024)

Robust Pseudolabelers & Vision Encoders

- 3D Common Corruptions (CVPR 2022, Oral) [Demo]
- Broadening Vision in VLMs (ECCV 2024, Oral)

Part II: Multimodal Training

Pretraining Framework

- 4M: Multimodal Modeling (NeurIPS 2023, Spotlight)

Architecture

Training the Unified Model

- 4M-21: Any-to-Any Models (NeurIPS 2024) [Demo]

Data

Model

Part III: Analysis

Internet-based Generalists

- Evaluating MFMs on Vision (arXiv 2025)

Status check

Real-world Specialization

- Test-space Multimodal Training (arXiv 2025)

Variable Tokenization

- Resampling Images via FlexTok (arXiv 2025)

Publication map & Thesis outline

Part I: Handling Modalities

Robustness Mechanisms

- Cross-task Consistency (CVPR 2020, Best Paper Nominee)
- Cross-domain Ensembles (ICCV 2021, Oral)
- Rapid Network Adaptation (ICCV 2023)
- OOD Generalization via Diversity (ICLR 2024)

Robust Pseudolabelers & Vision Encoders

- 3D Common Corruptions (CVPR 2022, Oral) [Demo]
- Broadening Vision in VLMs (ECCV 2024, Oral)

Part II: Multimodal Training

Pretraining Framework

- 4M: Multimodal Modeling (NeurIPS 2023, Spotlight)

Architecture

Status check

Training the Unified Model

- 4M-21: Any-to-Any Models (NeurIPS 2024) [Demo]

Data

Model

Part III: Analysis

Internet-based Generalists

- Evaluating MFMs on Vision (arXiv 2025)

Real-world Specialization

- Test-space Multimodal Training (arXiv 2025)

Variable Tokenization

- Resampling Images via FlexTok (arXiv 2025)

Publication map & Thesis outline

Part I: Handling Modalities

Robustness Mechanisms

- Cross-task Consistency (CVPR 2020, Best Paper Nominee)
- Cross-domain Ensembles (ICCV 2021, Oral)
- Rapid Network Adaptation (ICCV 2023)
- OOD Generalization via Diversity (ICLR 2024)

Robust Pseudolabelers & Vision Encoders

- 3D Common Corruptions (CVPR 2022, Oral) [Demo]
- Broadening Vision in VLMs (ECCV 2024, Oral)

Part II: Multimodal Training

Pretraining Framework

- 4M: Multimodal Modeling (NeurIPS 2023, Spotlight)

Architecture

Training the Unified Model

- 4M-21: Any-to-Any Models (NeurIPS 2024) [Demo]

Data

Part III: Analysis

Internet-based Generalists

- Evaluating MFMs on Vision (arXiv 2025)

Status check

Real-world Specialization

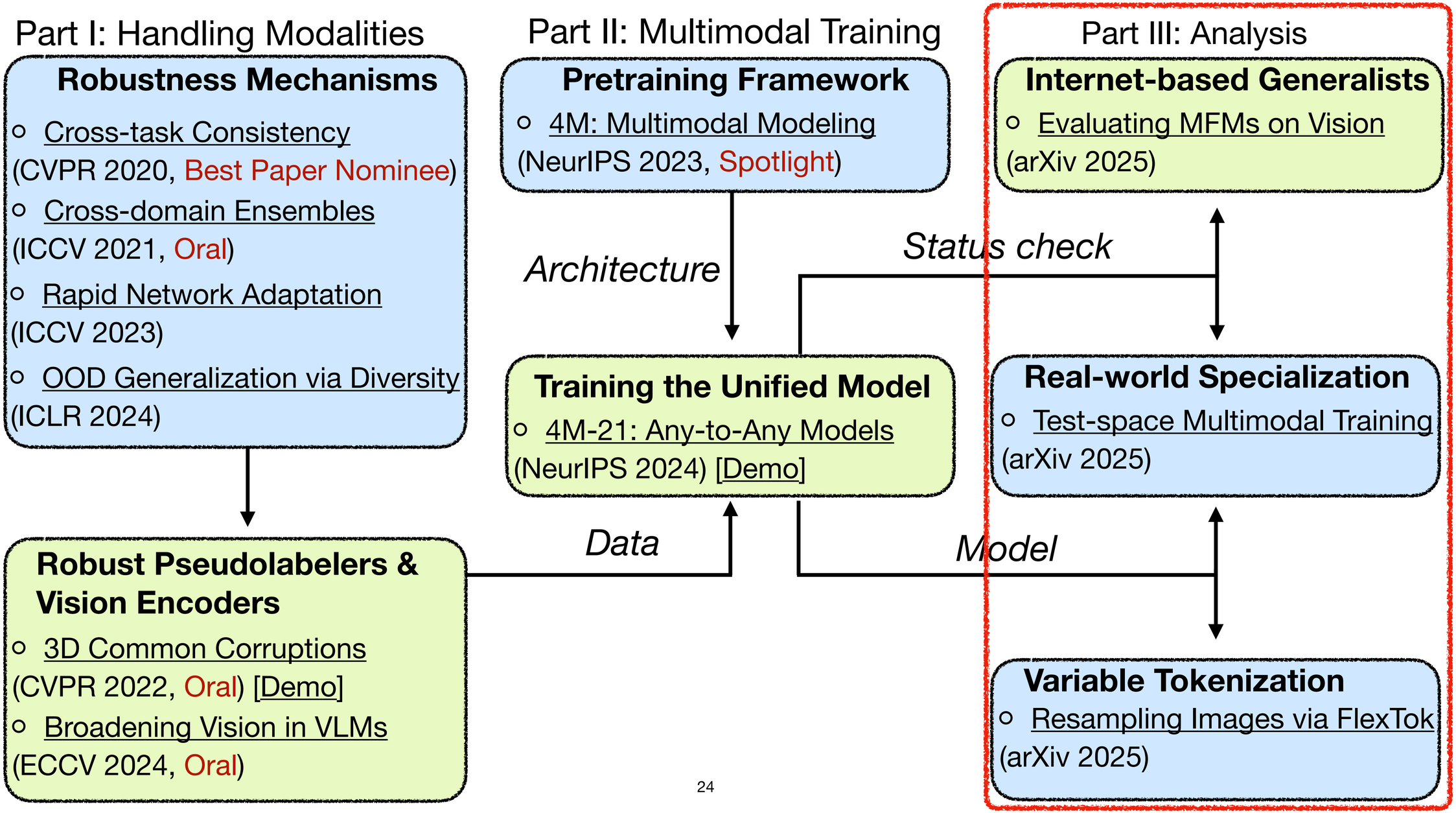
- Test-space Multimodal Training (arXiv 2025)

Model

Variable Tokenization

- Resampling Images via FlexTok (arXiv 2025)

Publication map & Thesis outline



Publication map & Thesis outline

Part I: Handling Modalities

Robustness Mechanisms

- Cross-task Consistency (CVPR 2020, Best Paper Nominee)
- Cross-domain Ensembles (ICCV 2021, Oral)
- Rapid Network Adaptation (ICCV 2023)
- OOD Generalization via Diversity (ICLR 2024)

Robust Pseudolabelers & Vision Encoders

- 3D Common Corruptions (CVPR 2022, Oral) [Demo]
- Broadening Vision in VLMs (ECCV 2024, Oral)

Part II: Multimodal Training

Pretraining Framework

- 4M: Multimodal Modeling (NeurIPS 2023, Spotlight)

Architecture

Training the Unified Model

- 4M-21: Any-to-Any Models (NeurIPS 2024) [Demo]

Data

Model

Status check

Part III: Analysis

Internet-based Generalists

- Evaluating MFMs on Vision (arXiv 2025)

Real-world Specialization

- Test-space Multimodal Training (arXiv 2025)

Variable Tokenization

- Resampling Images via FlexTok (arXiv 2025)

3D Common Corruptions and Data Augmentation

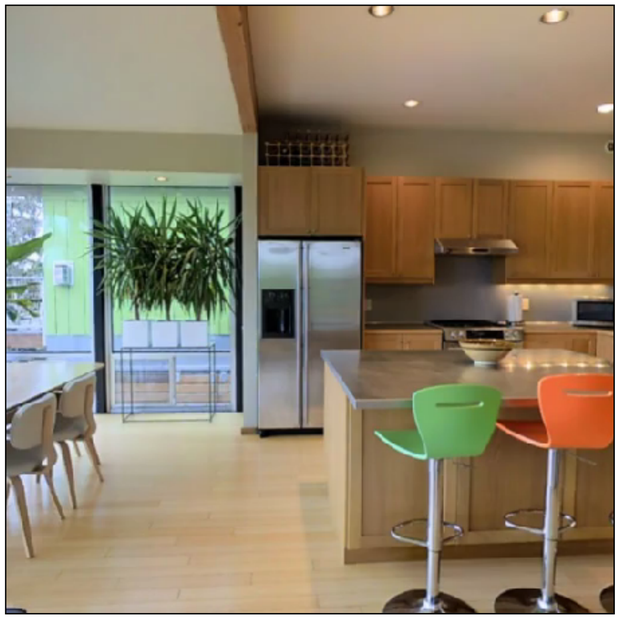
CVPR 2022 (Oral)

3dcommoncorruptions.epfl.ch

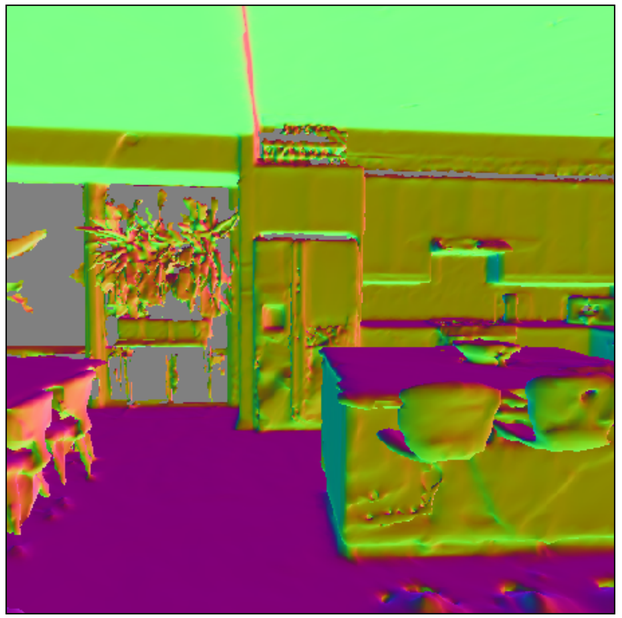
EPFL Swiss Federal Institute of Technology

(Vision) models fail in the real world

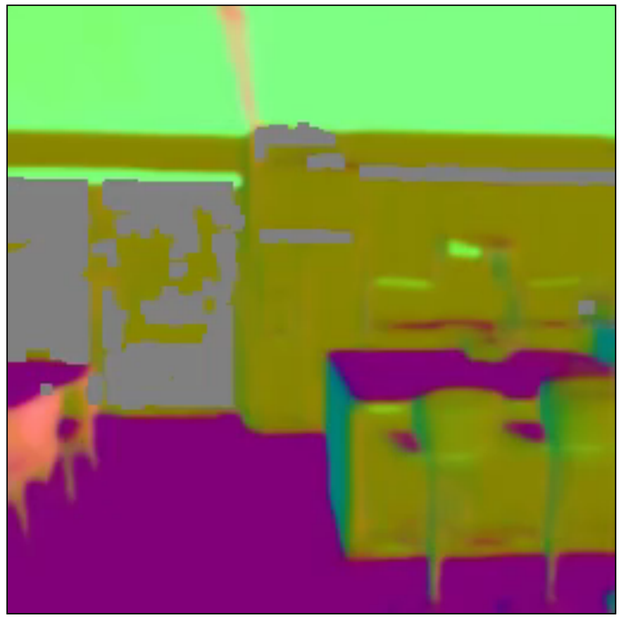
Input

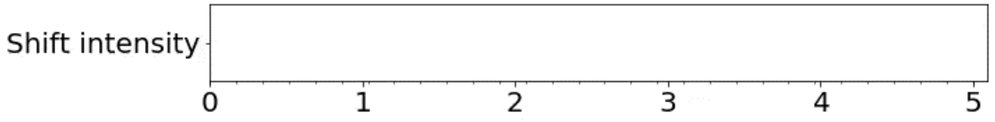


Ground Truth



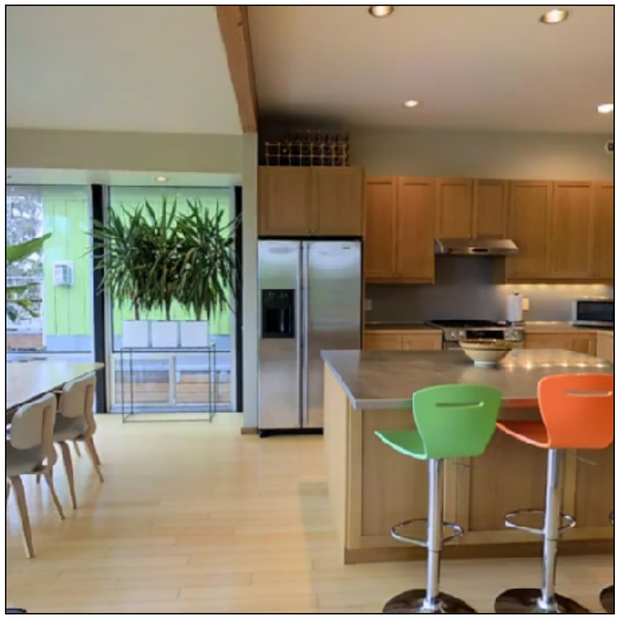
Prediction (normals)



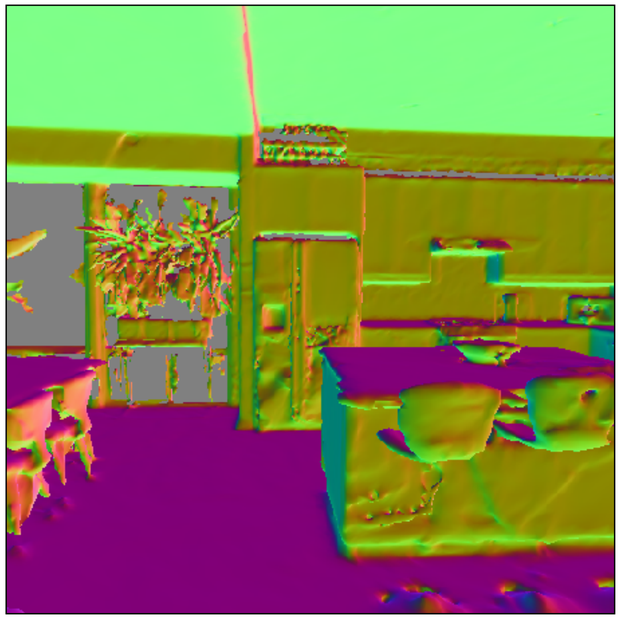


(Vision) models fail in the real world
(not an adversarial attack!)

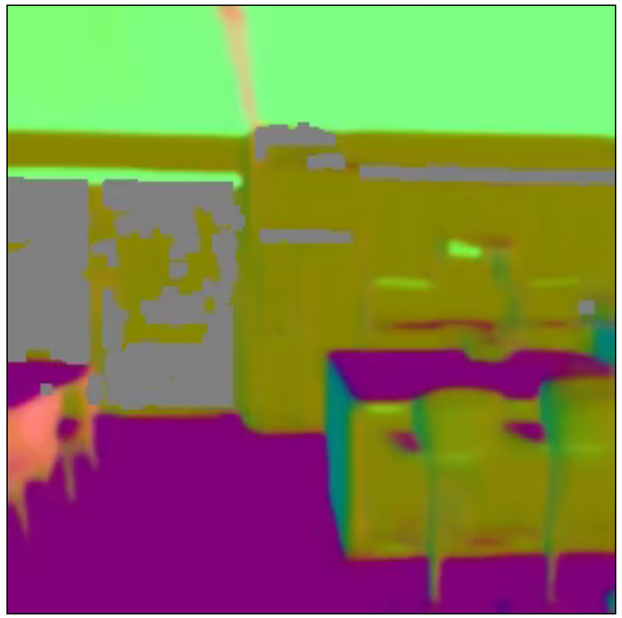
Input

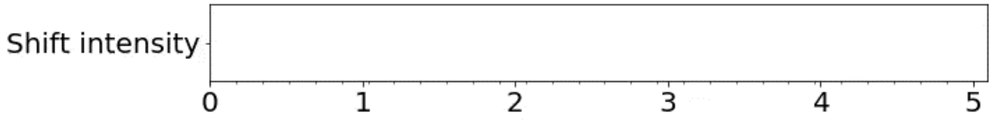


Ground Truth

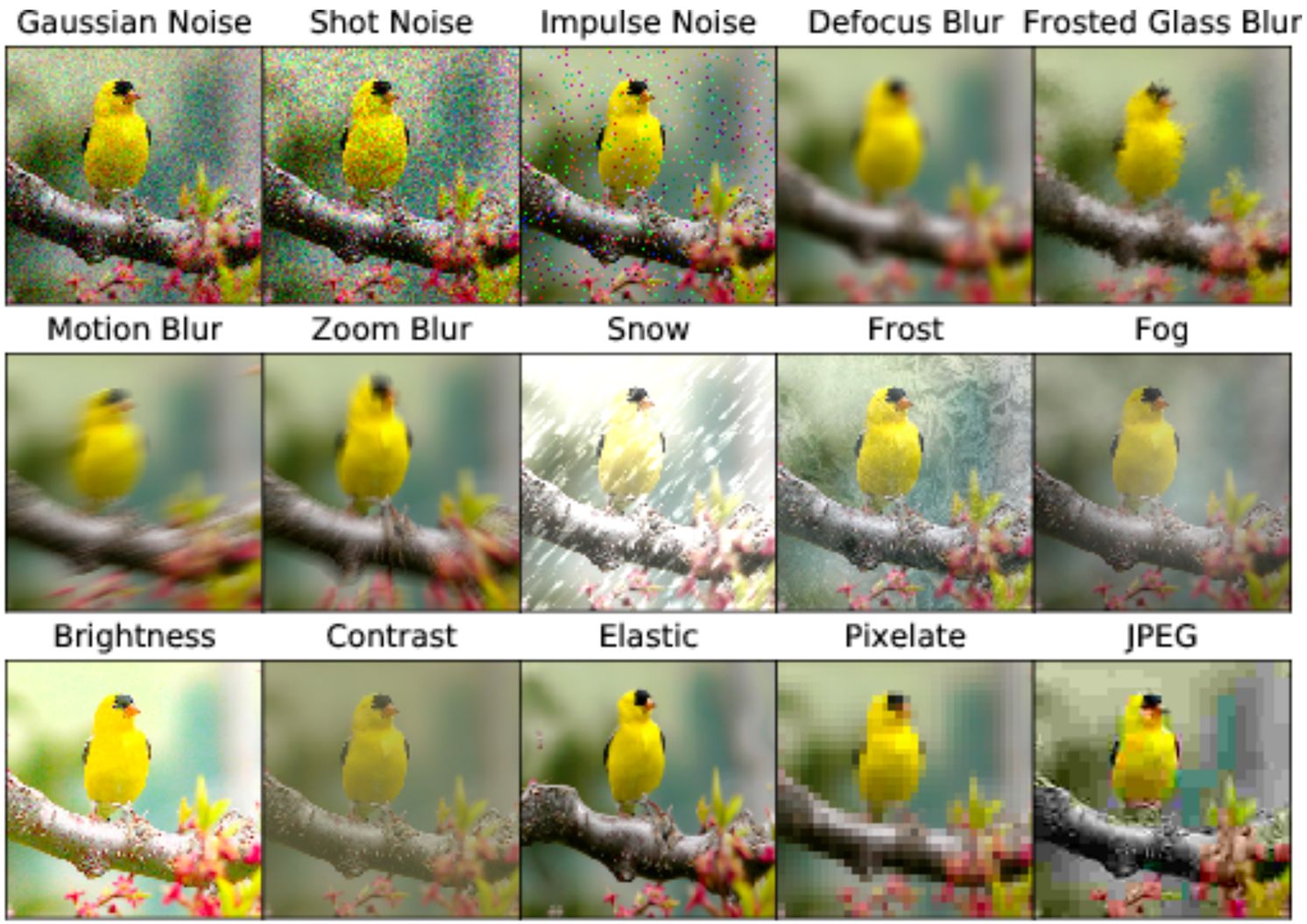


Prediction (normals)





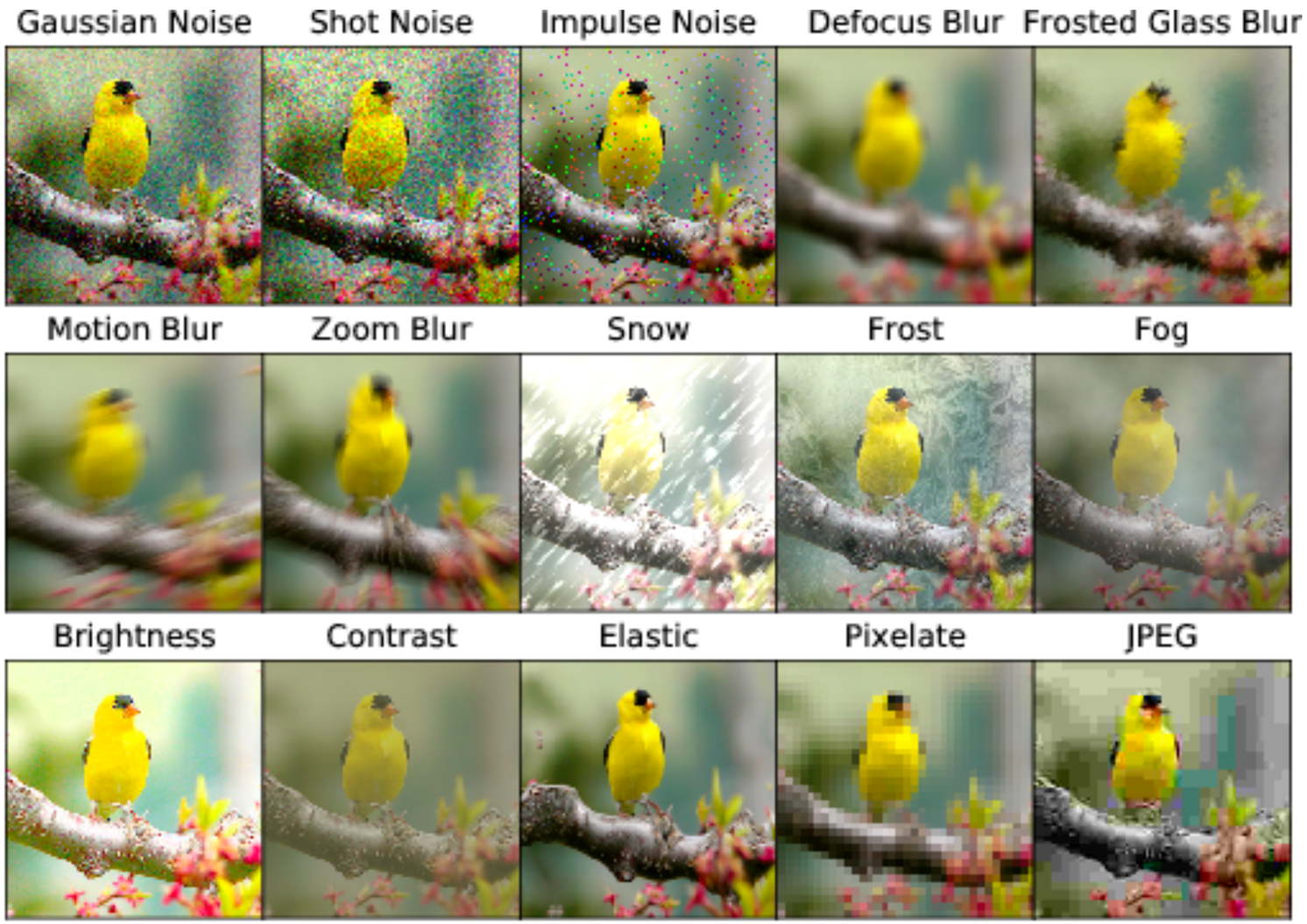
(2D) Common Corruptions[1]



[1] Hendrycks and Dietterich, ICLR 2019

(2D) Common Corruptions[1]

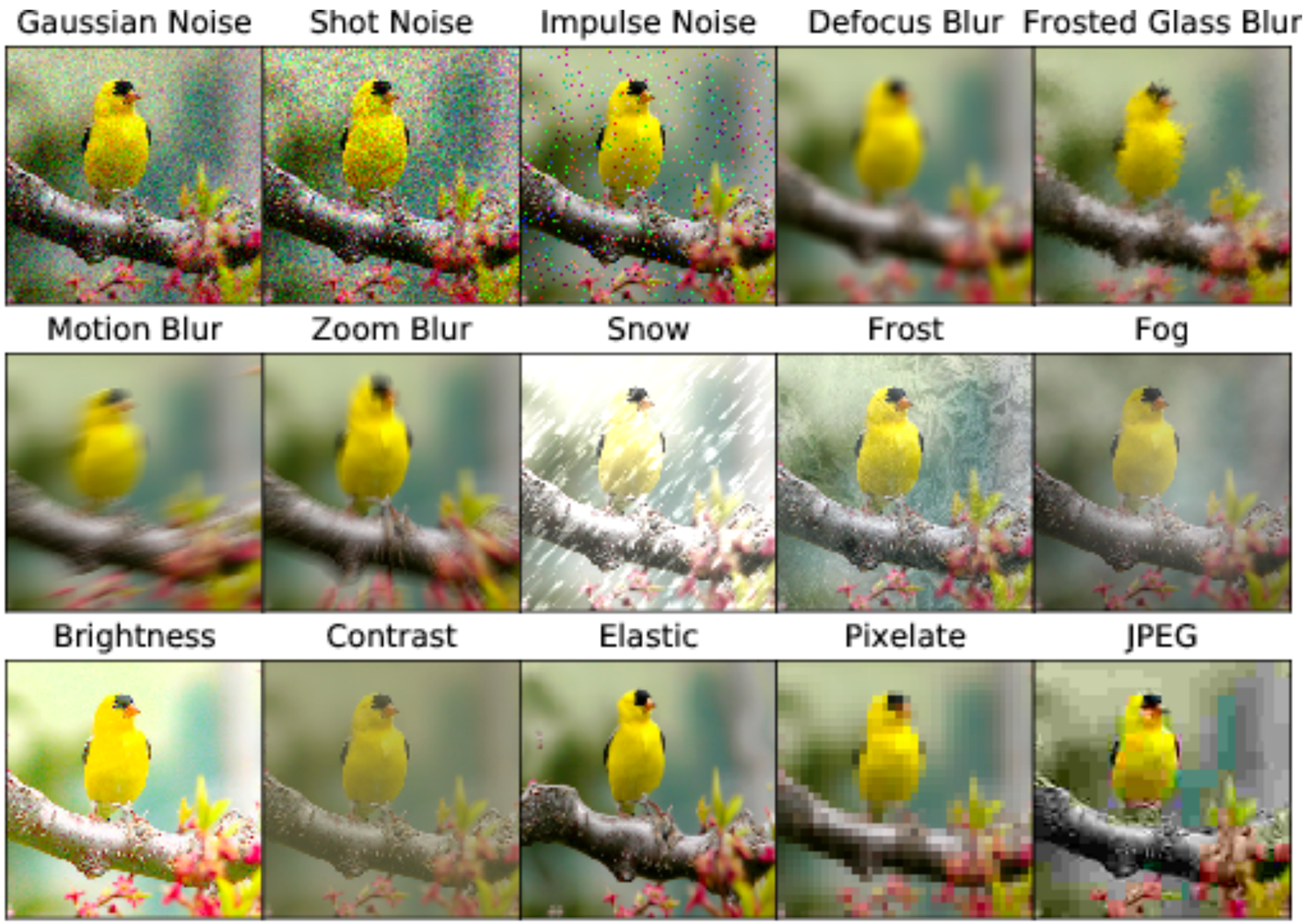
- Uniform modifications



[1] Hendrycks and Dietterich, ICLR 2019

(2D) Common Corruptions[1]

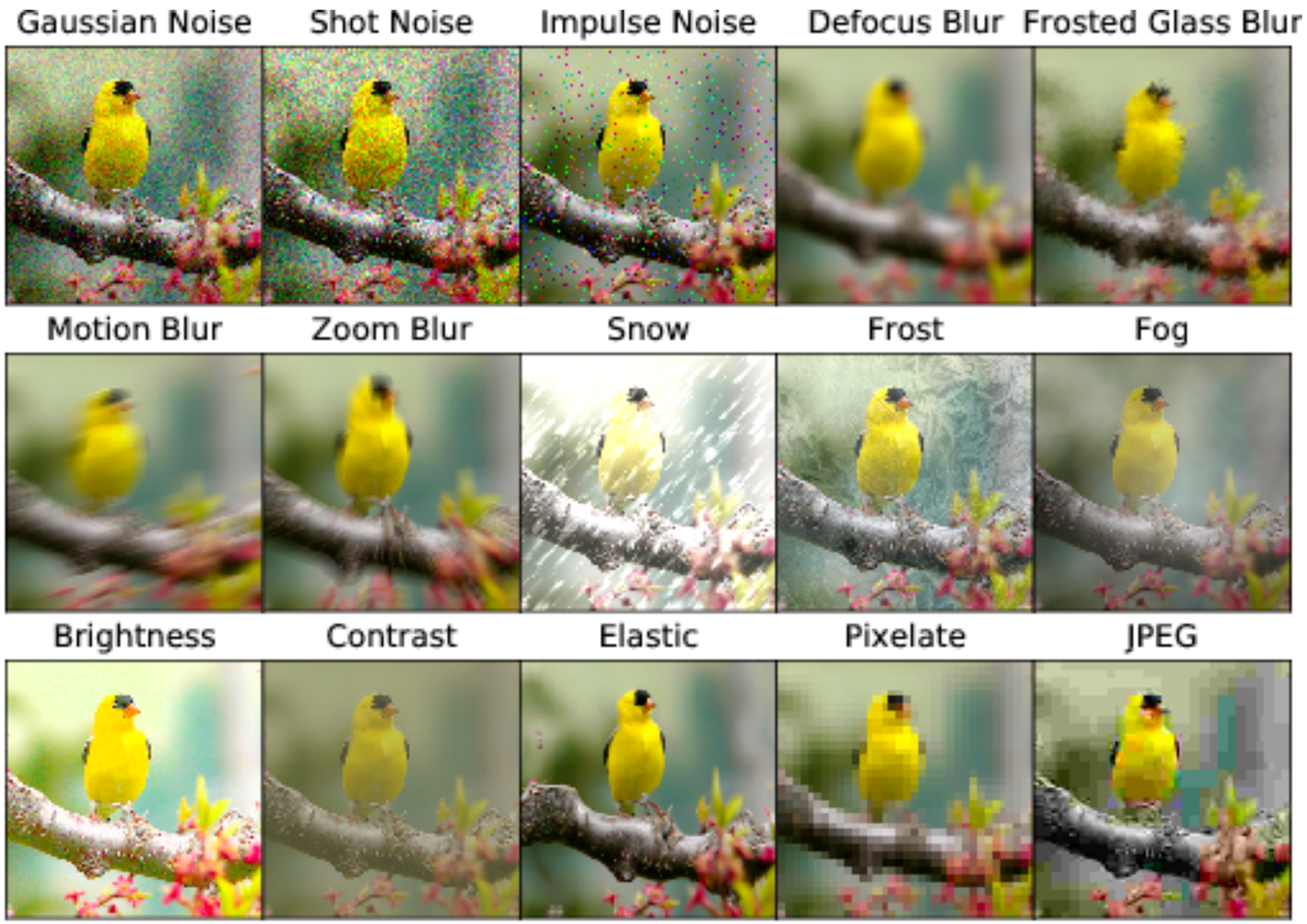
- Uniform modifications
- No use of 3D or semantics



[1] Hendrycks and Dietterich, ICLR 2019

(2D) Common Corruptions[1]

- Uniform modifications
- No use of 3D or semantics
- → **Unrealistic**

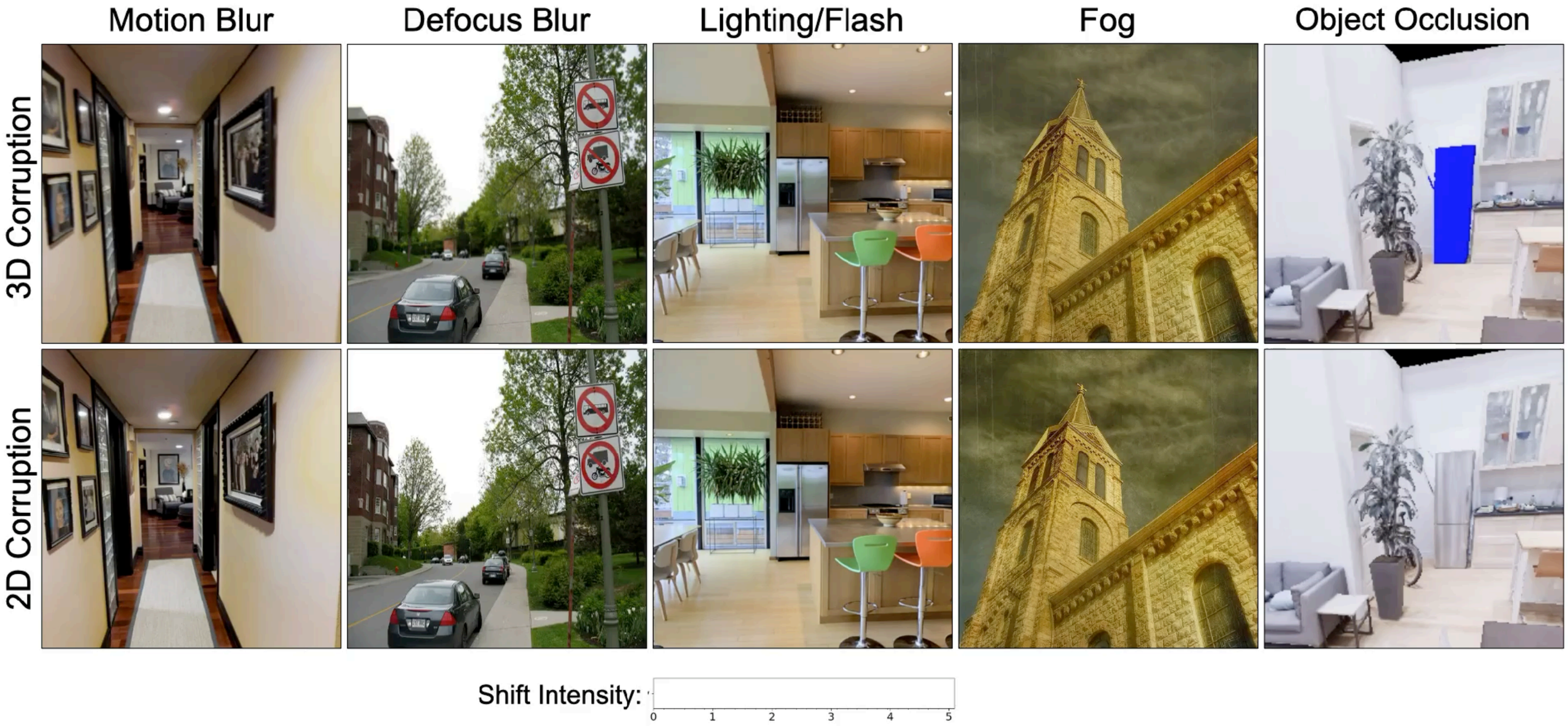


[1] Hendrycks and Dietterich, ICLR 2019

Main idea

Incorporate **3D info + real-world properties** to generate corruptions

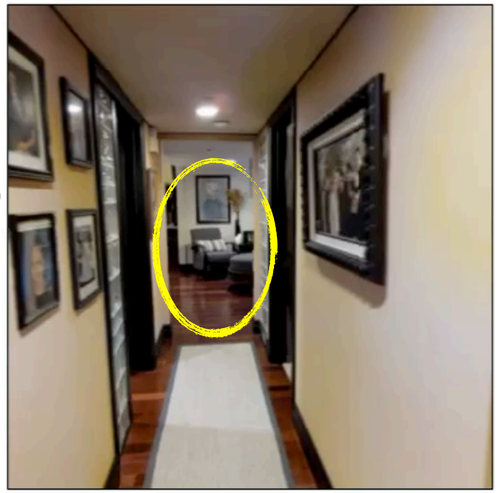
Outcome: A set of more realistic corruptions



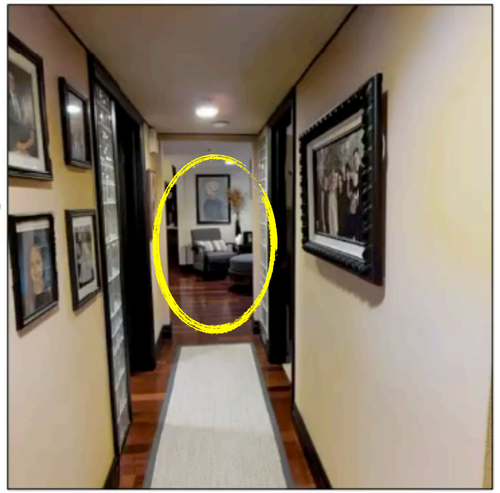
Outcome: A set of more realistic corruptions

Motion Blur

3D Corruption



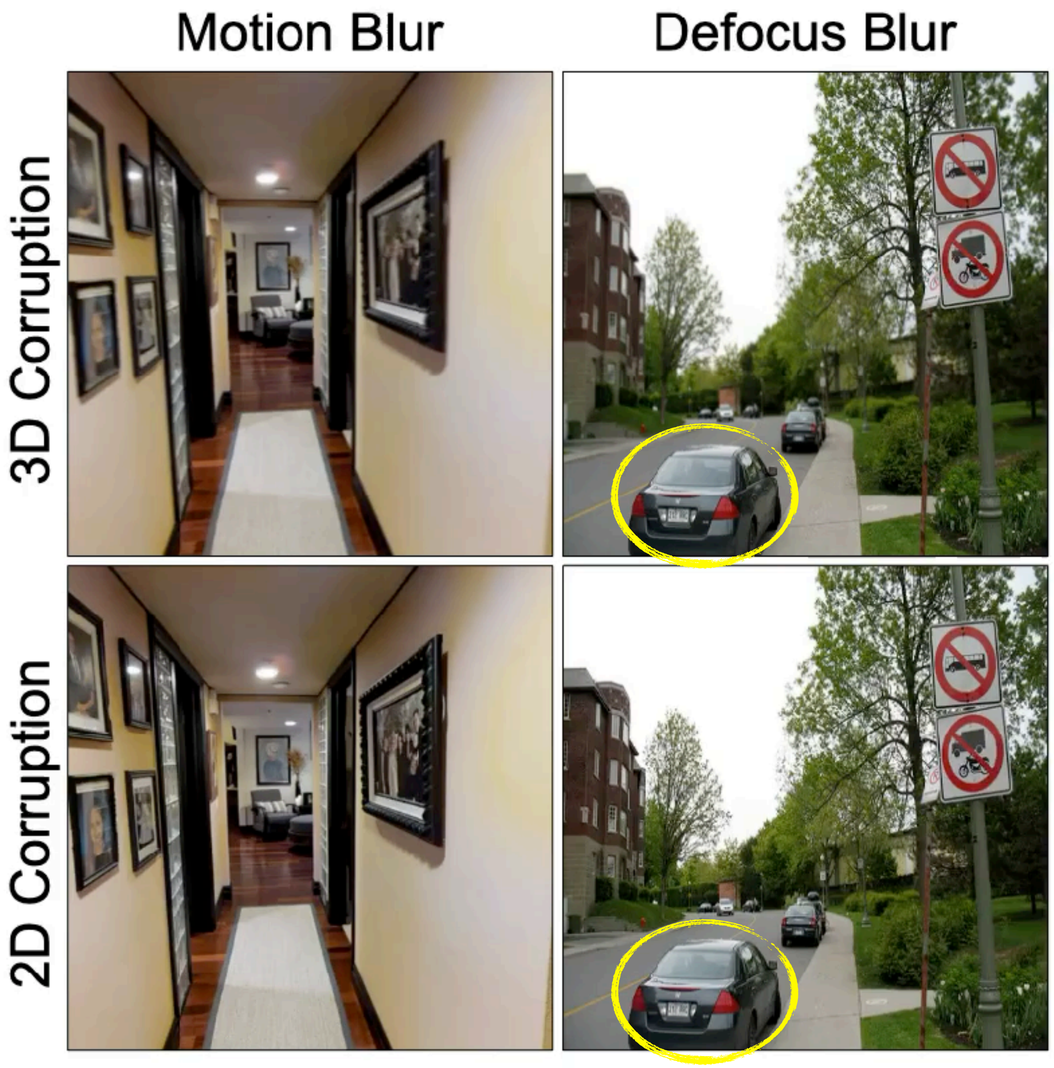
2D Corruption



Shift Intensity: 0 1 2 3 4 5

Parallax-consistent motion blur

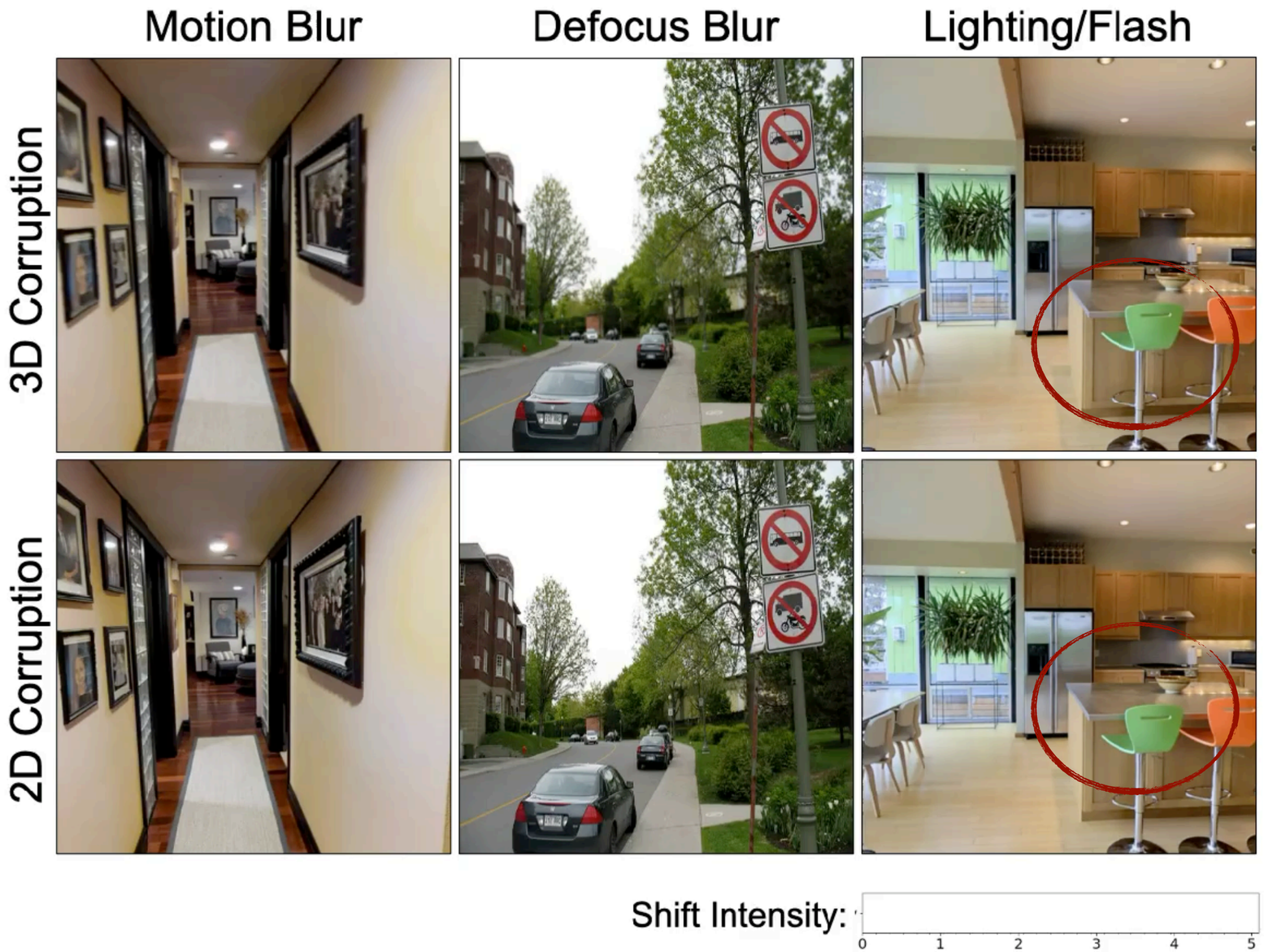
Outcome: A set of more realistic corruptions



Shift Intensity: 0 1 2 3 4 5

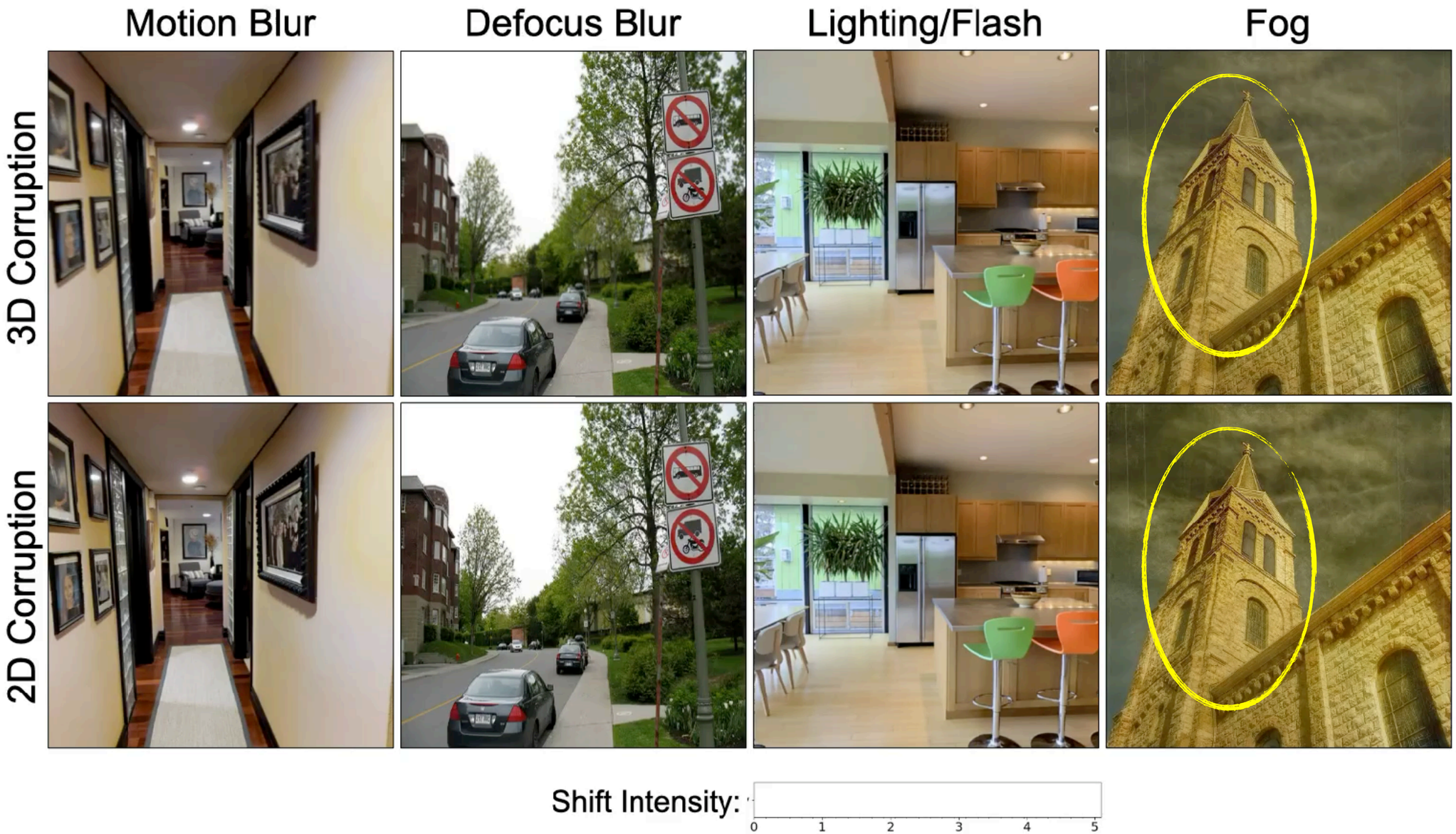
Depth of field effect

Outcome: A set of more realistic corruptions



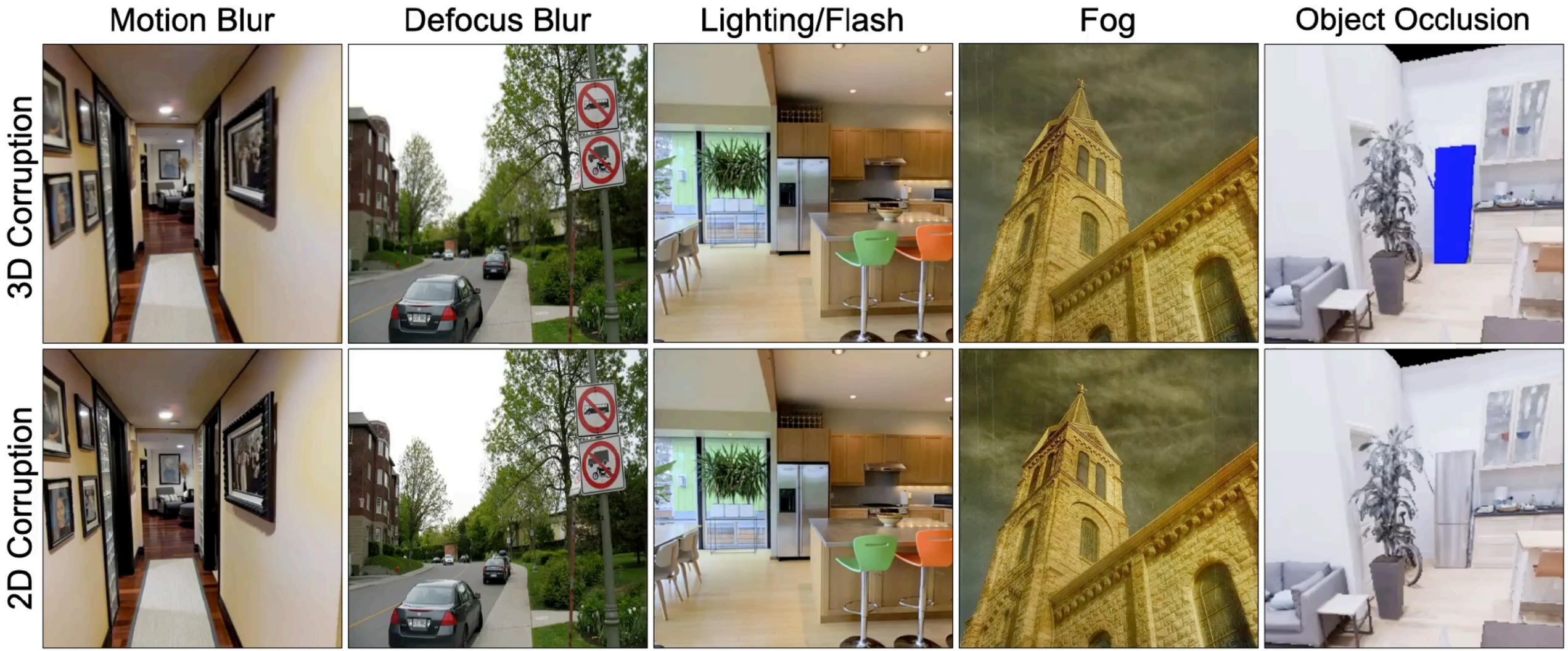
Geometry-informed lighting changes

Outcome: A set of more realistic corruptions



Thickening fog with more distance

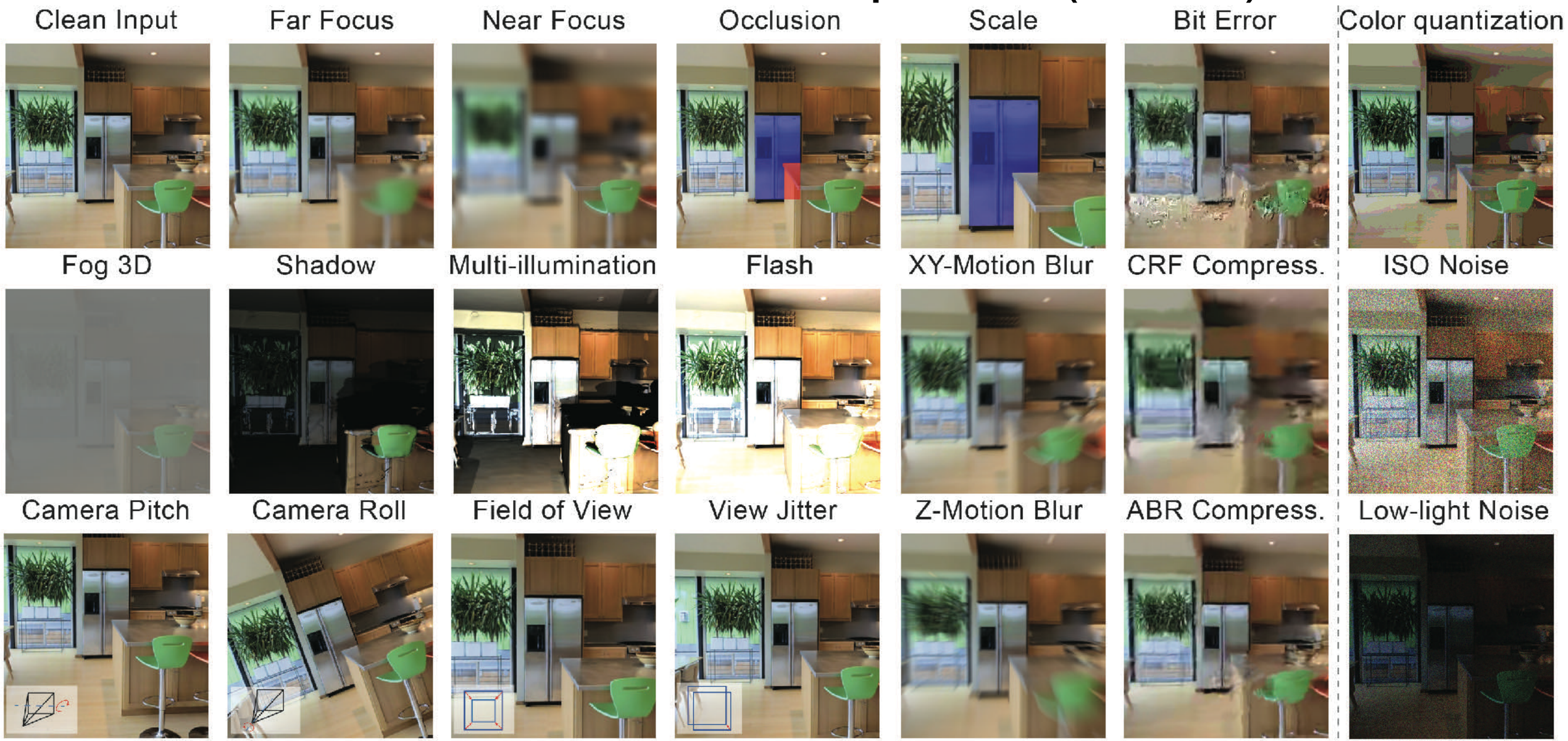
Outcome: A set of more realistic corruptions



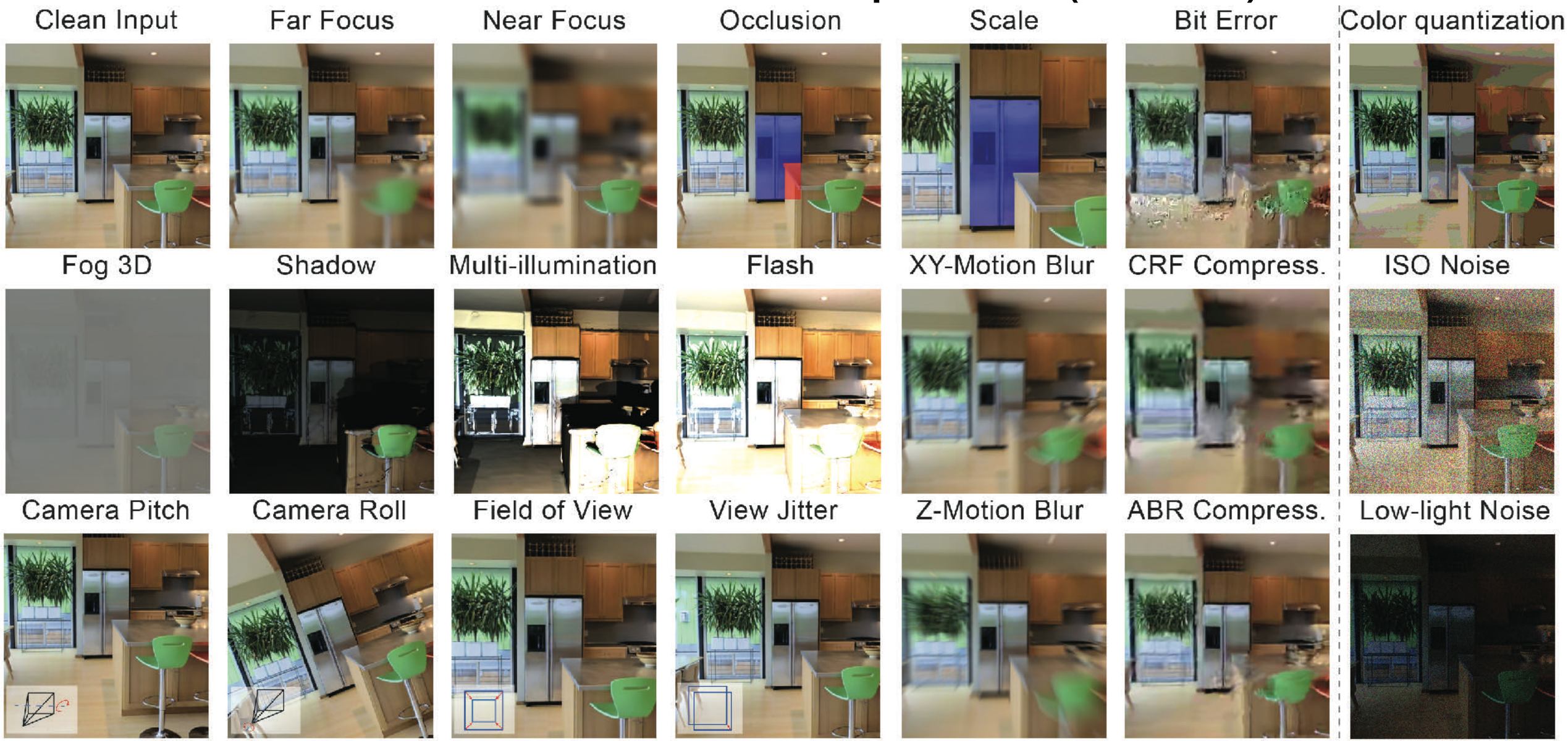
Shift Intensity: 0 1 2 3 4 5

Semantic corruptions

All 3D Common Corruptions (3DCC)

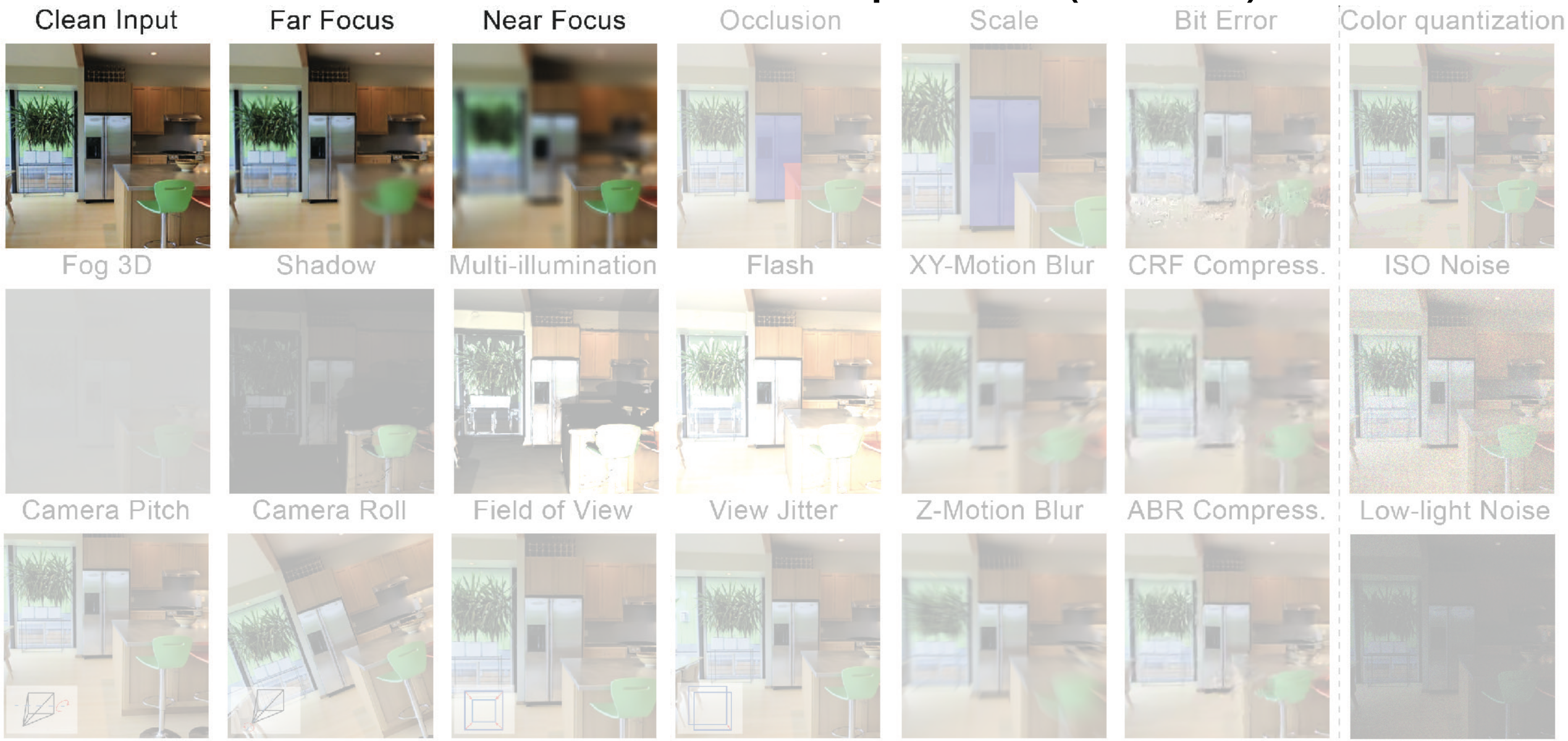


All 3D Common Corruptions (3DCC)



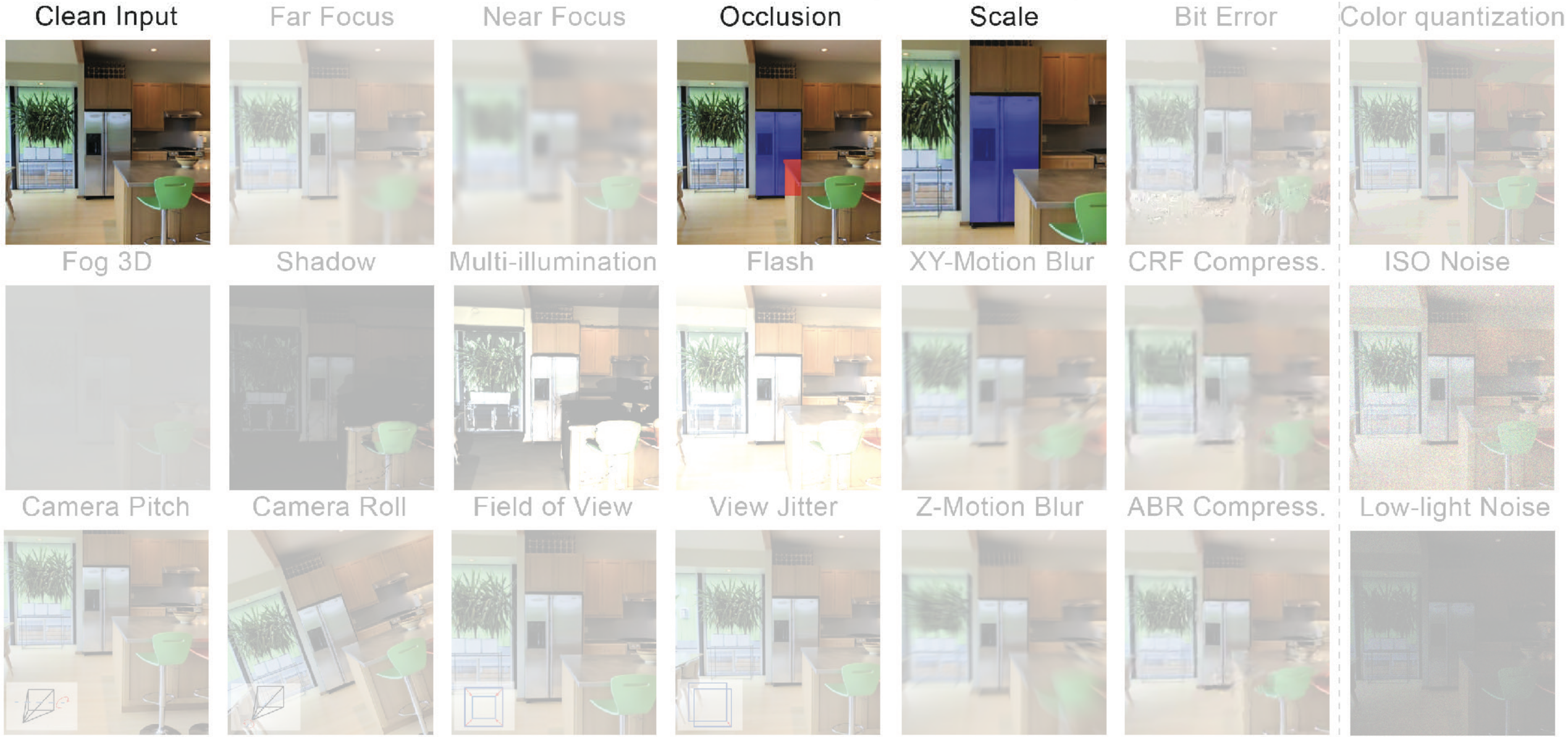
They address several aspects of real world:

All 3D Common Corruptions (3DCC)



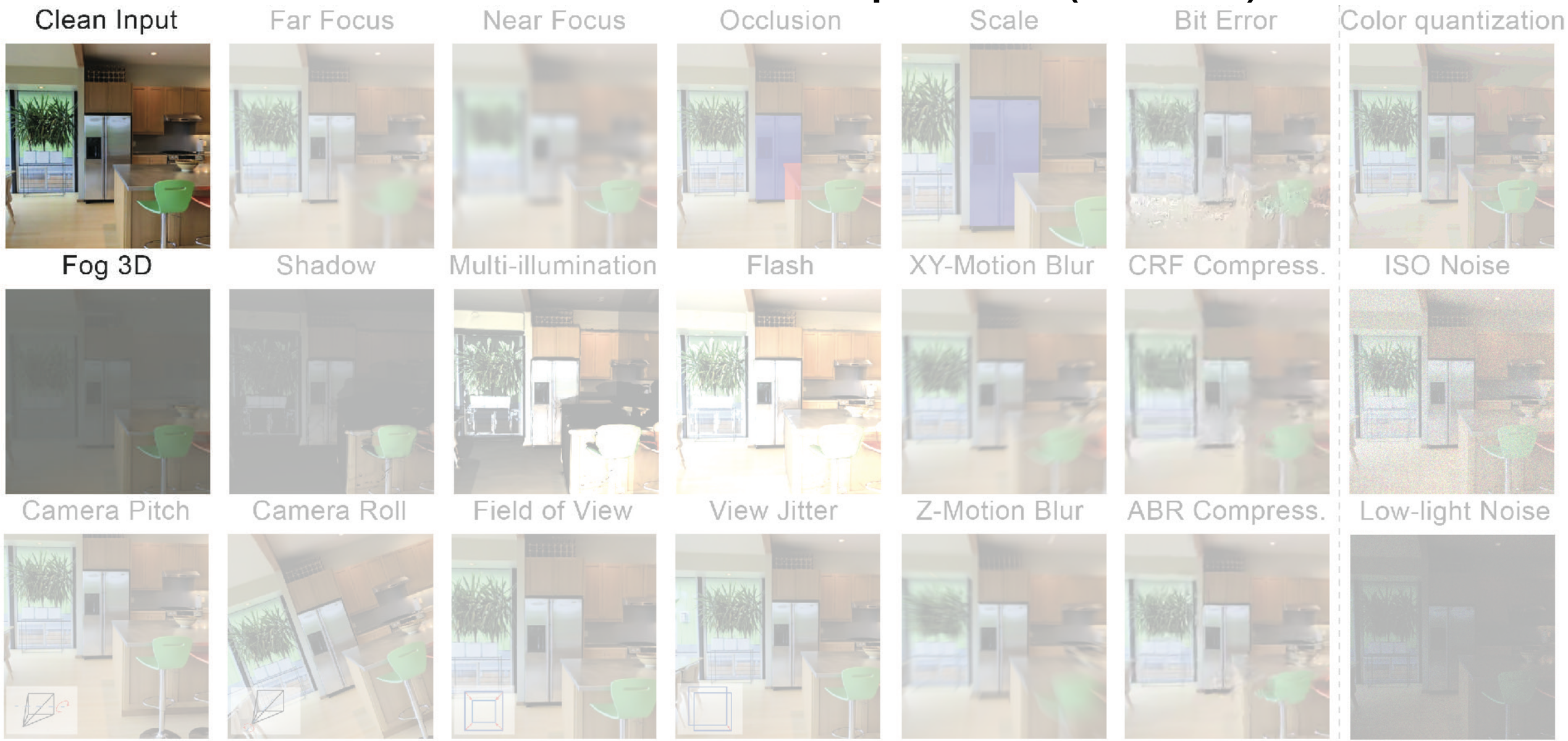
Depth of field

All 3D Common Corruptions (3DCC)



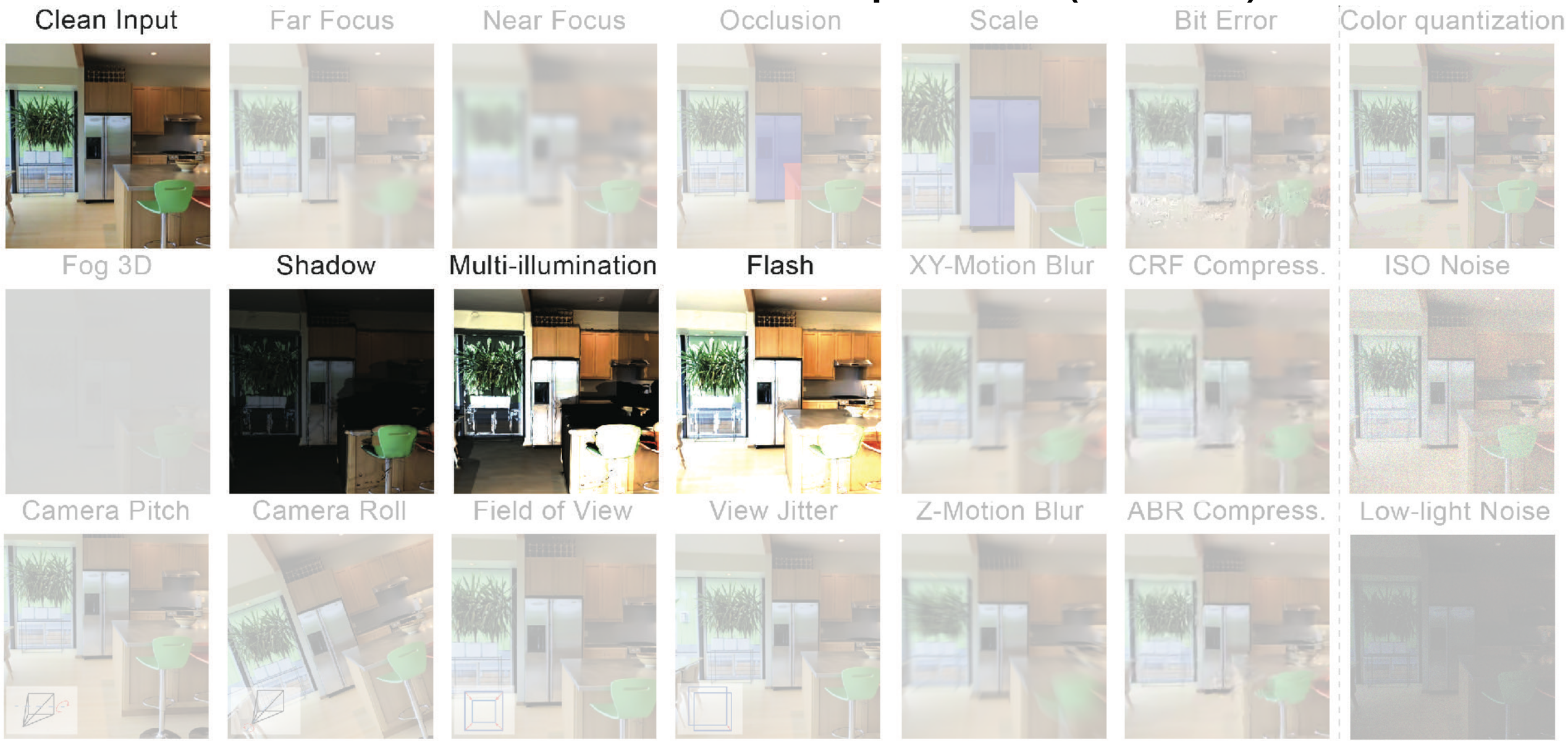
Object occlusions and scale changes

All 3D Common Corruptions (3DCC)



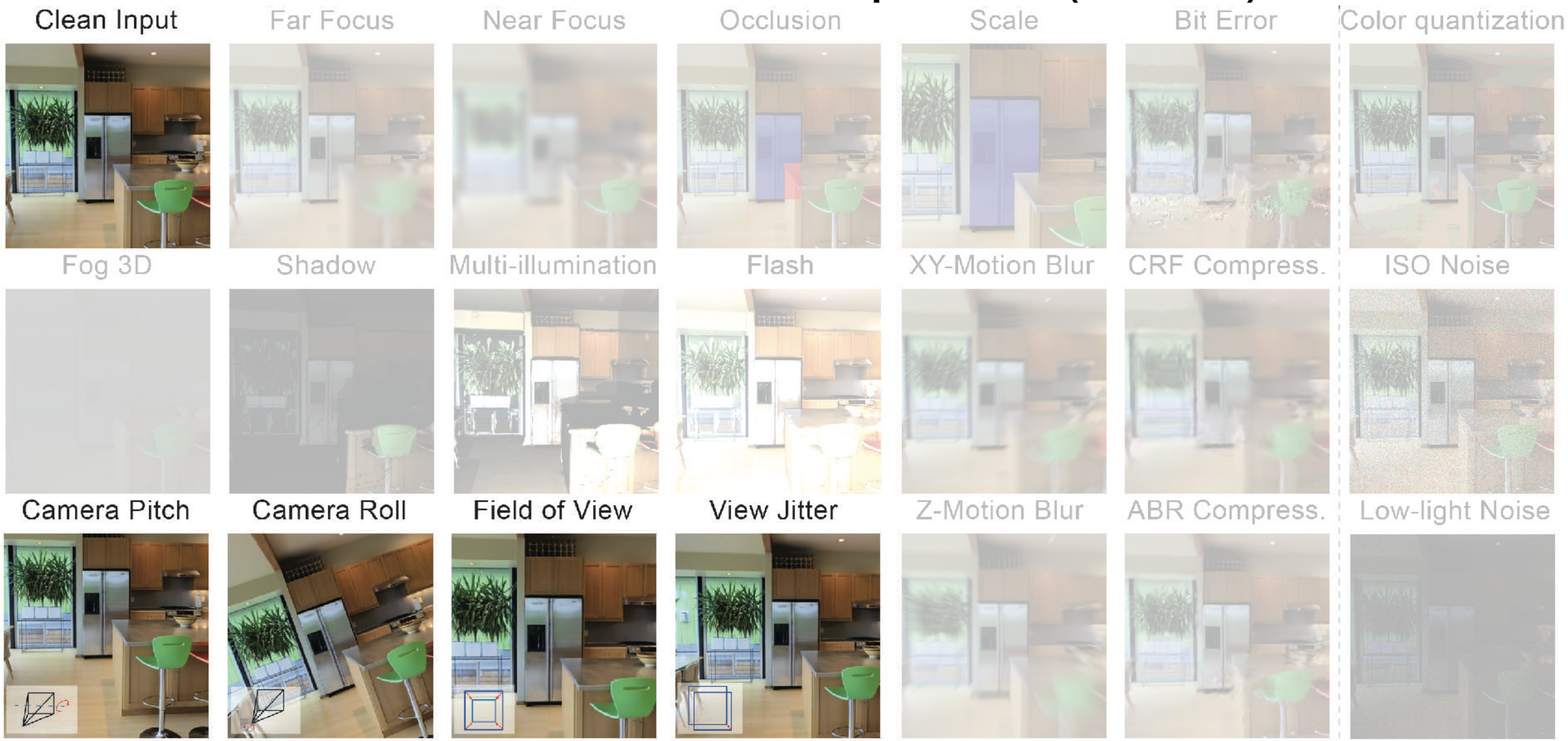
Weather

All 3D Common Corruptions (3DCC)



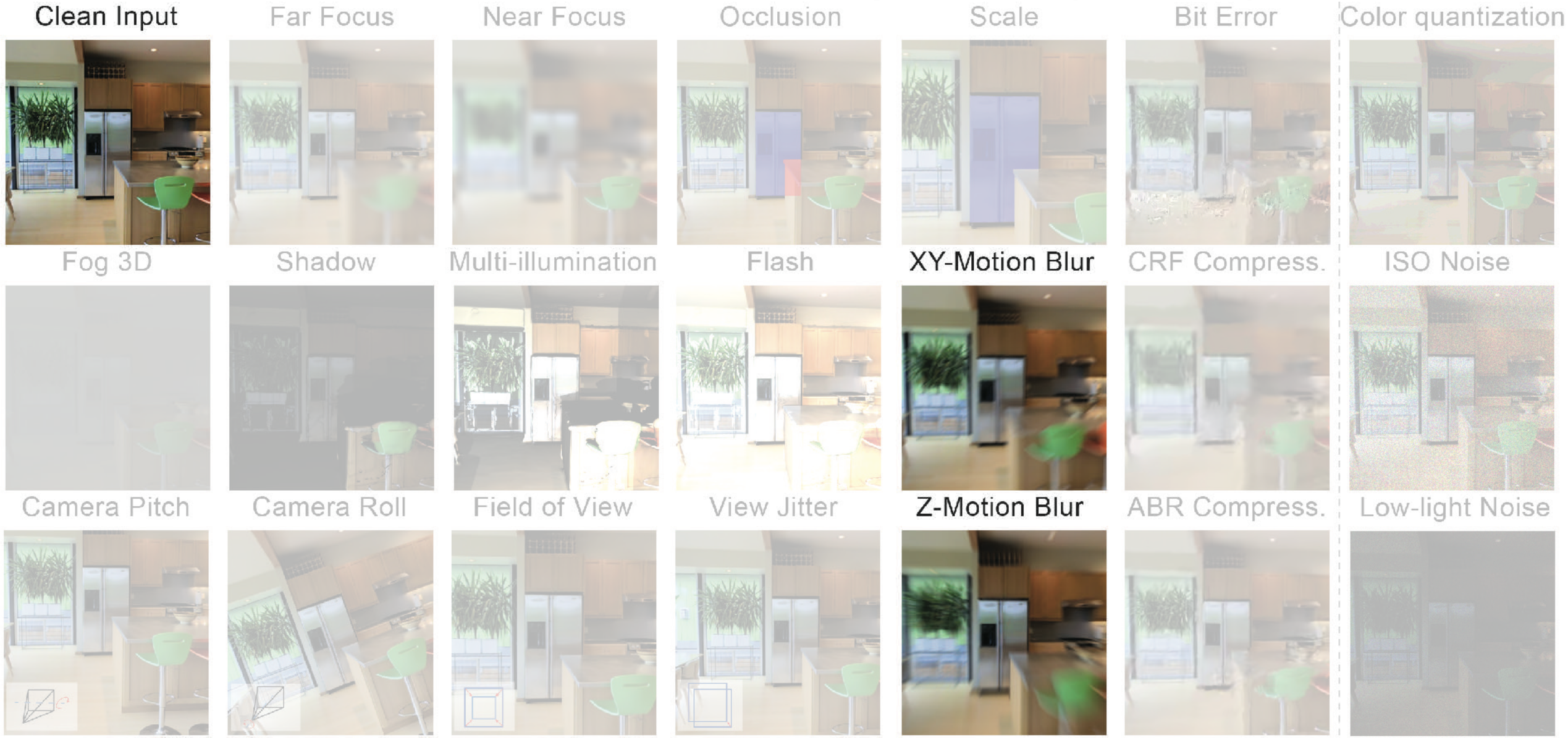
Lighting

All 3D Common Corruptions (3DCC)



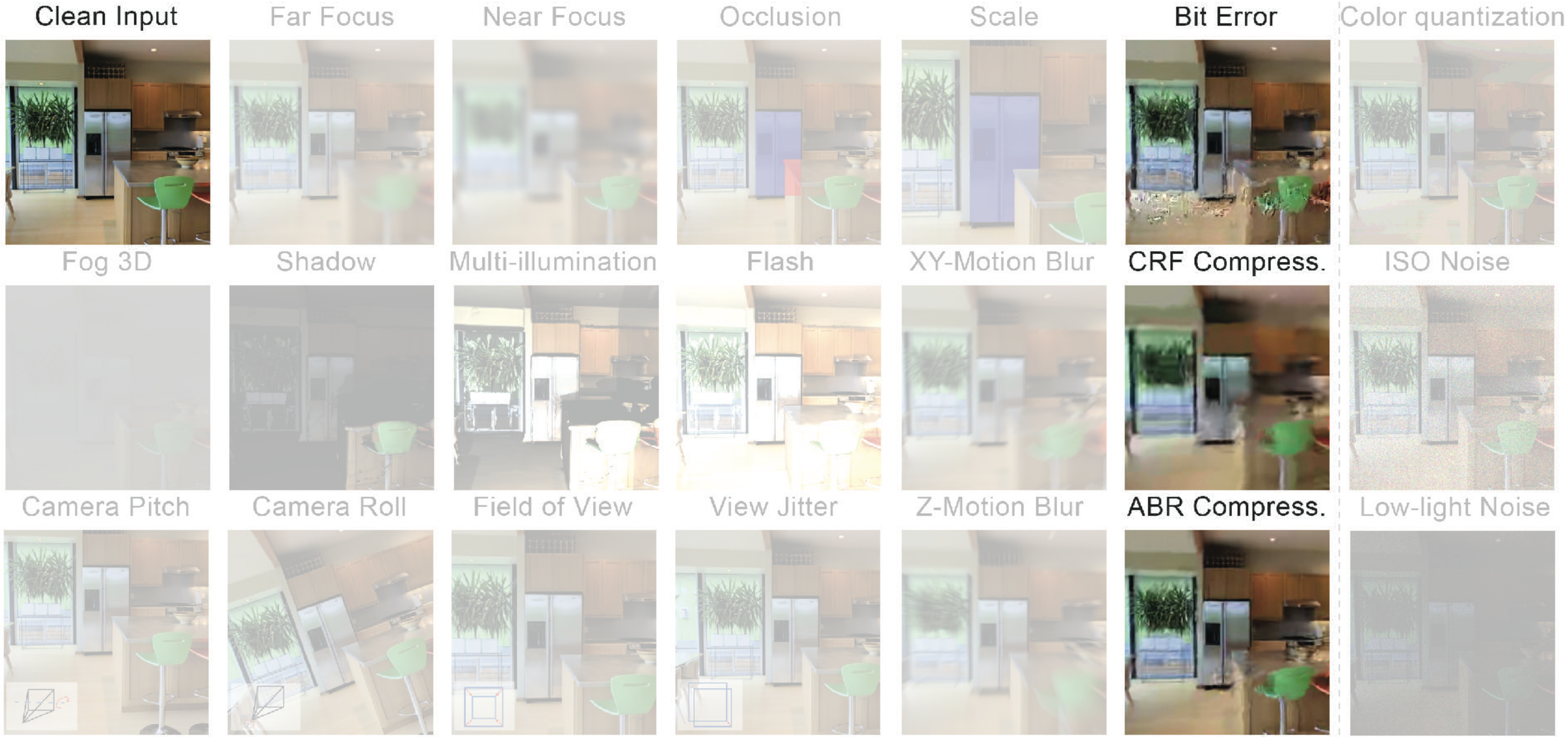
View changes

All 3D Common Corruptions (3DCC)



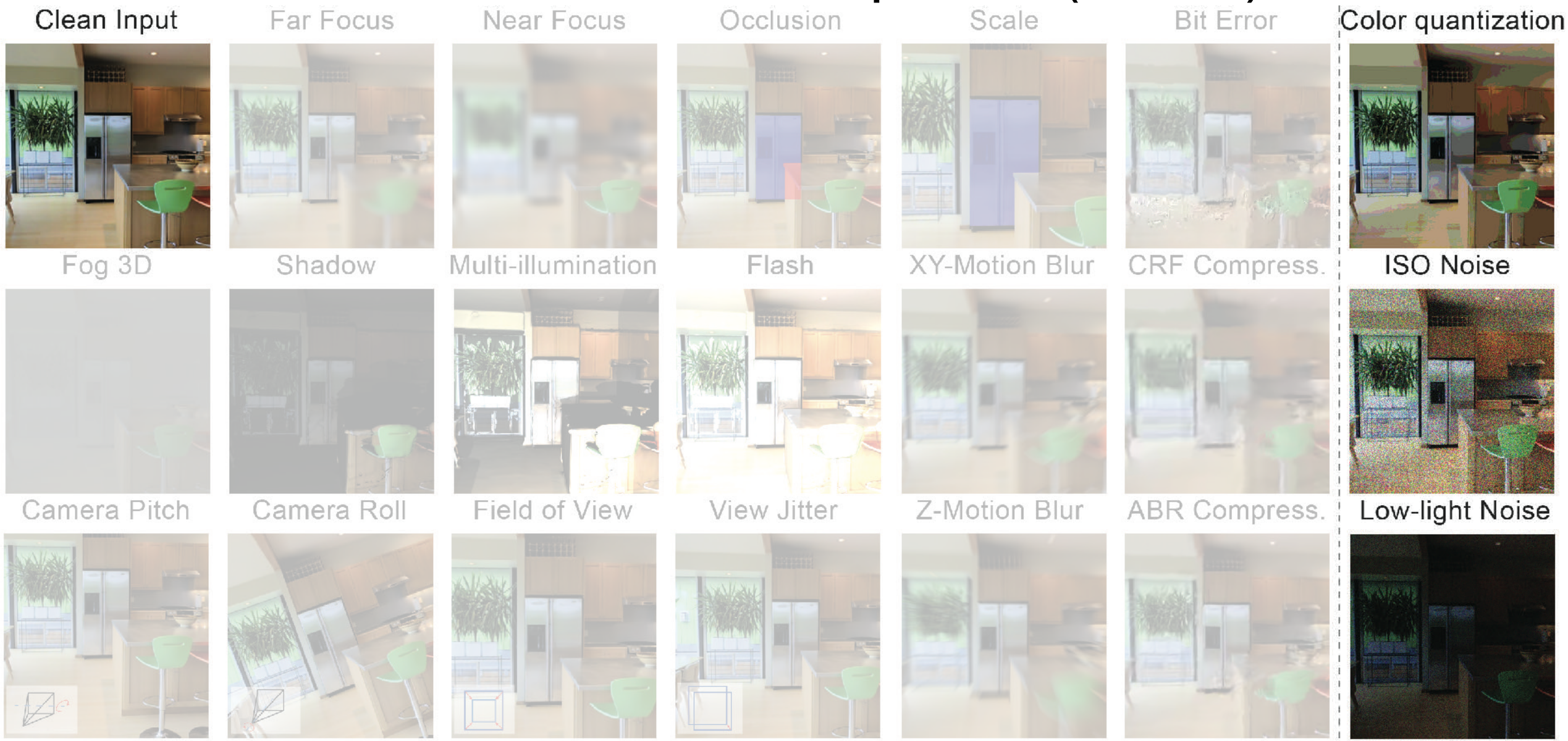
Blur due to camera motion

All 3D Common Corruptions (3DCC)



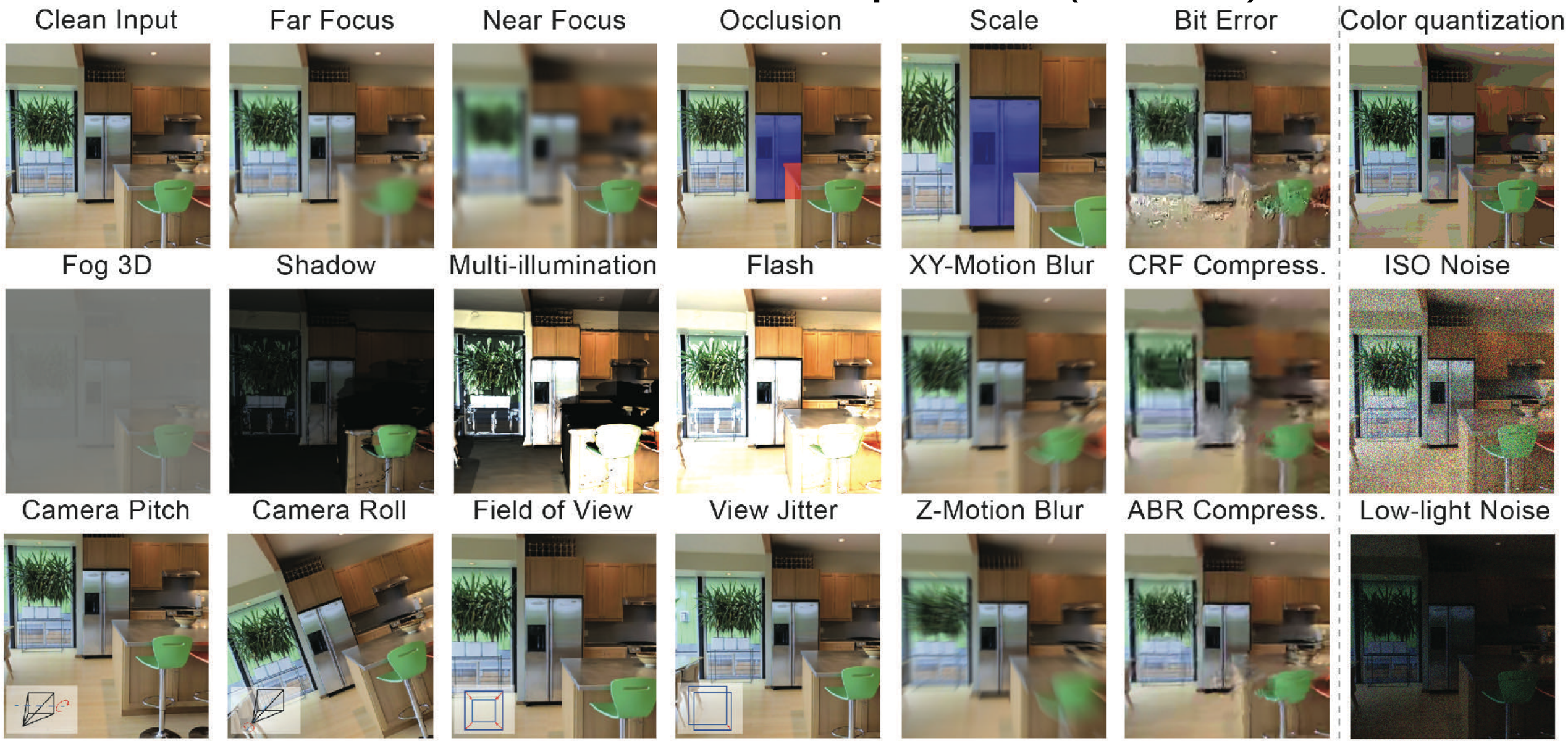
Video processing/transmission artifacts

All 3D Common Corruptions (3DCC)



Noise due to camera hardware (non-3D)

All 3D Common Corruptions (3DCC)



A diverse set of distribution shifts

What is the **significance** of these corruptions?

What is the **significance** of these corruptions?

- A challenging robustness benchmark

What is the **significance** of these corruptions?

- A challenging robustness benchmark
- Extendable to standard datasets

What is the **significance** of these corruptions?

- A challenging robustness benchmark
- Extendable to standard datasets
- Data augmentation

Idea: Use corruptions from 3DCC as data augmentation

Idea: Use corruptions from 3DCC as data augmentation
Outcome: Significantly improved robustness

Robust pseudolabelers!

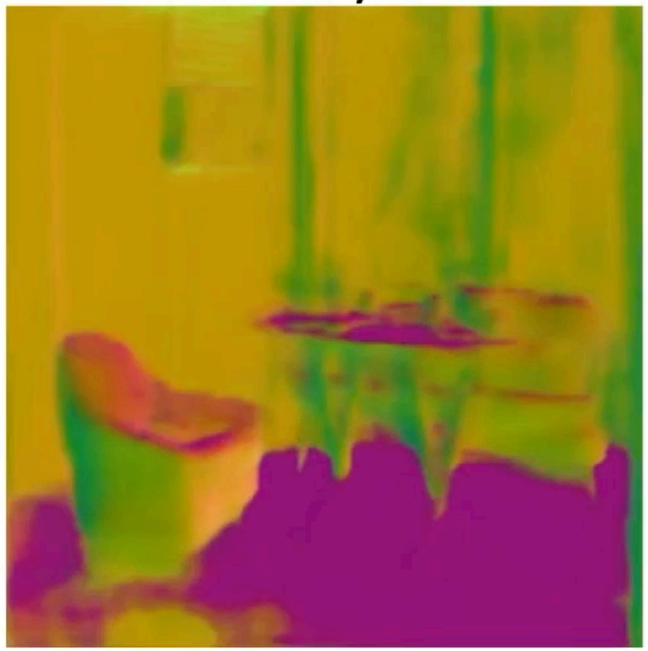
Qualitative results

- Baselines
  - Taskonomy+UNet: Taskonomy[1] trained model with UNet[2] backbone
  - Taskonomy+DPT: Taskonomy[1] trained model with DPT[3] backbone
  - Omnidata+DPT: Omnidata[4] trained model with DPT[3] backbone

[1]Zamir et al., CVPR 2018
[2]Ronneberger et al., MICCAI 2015
[3]Ranftl et al., ICCV 2021
[4]Eftekhar et al., ICCV 2021

Input

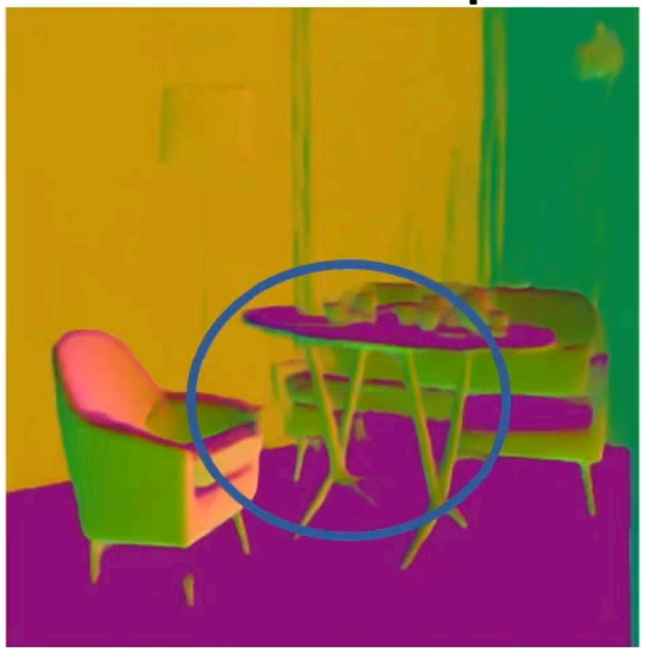
Taskonomy + UNet

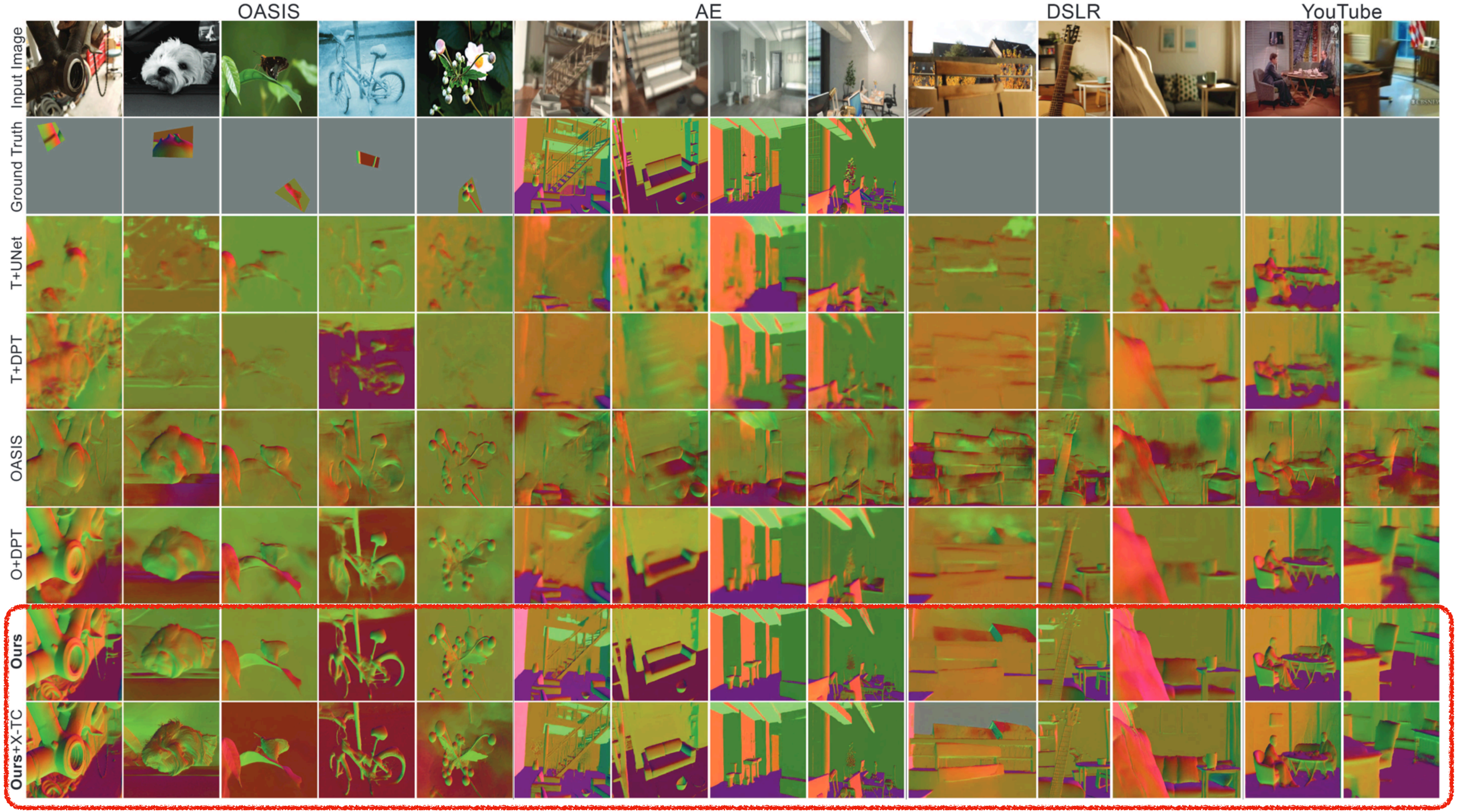


Taskonomy + DPT

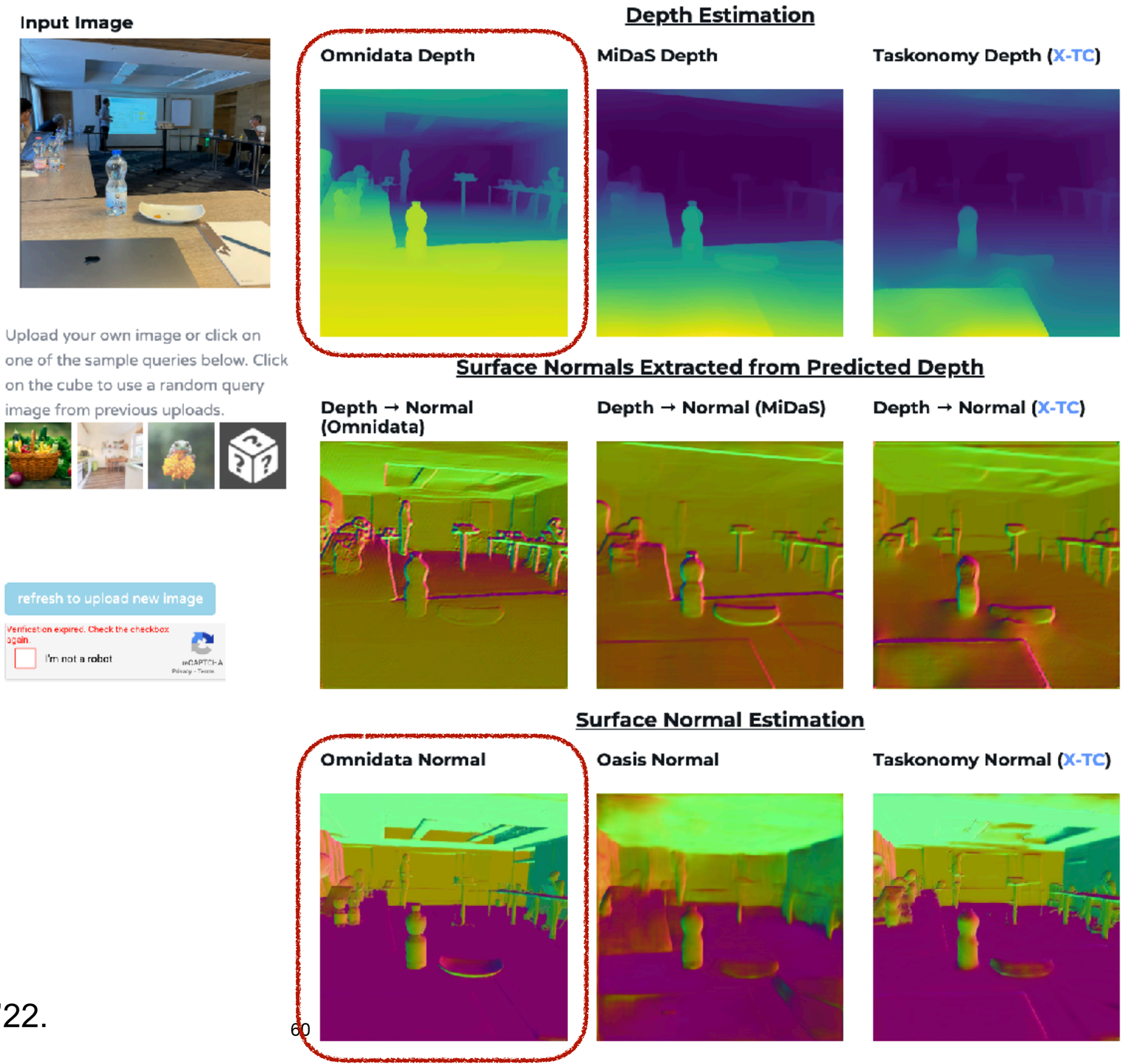
Omnidata + DPT

3D Common Corruptions



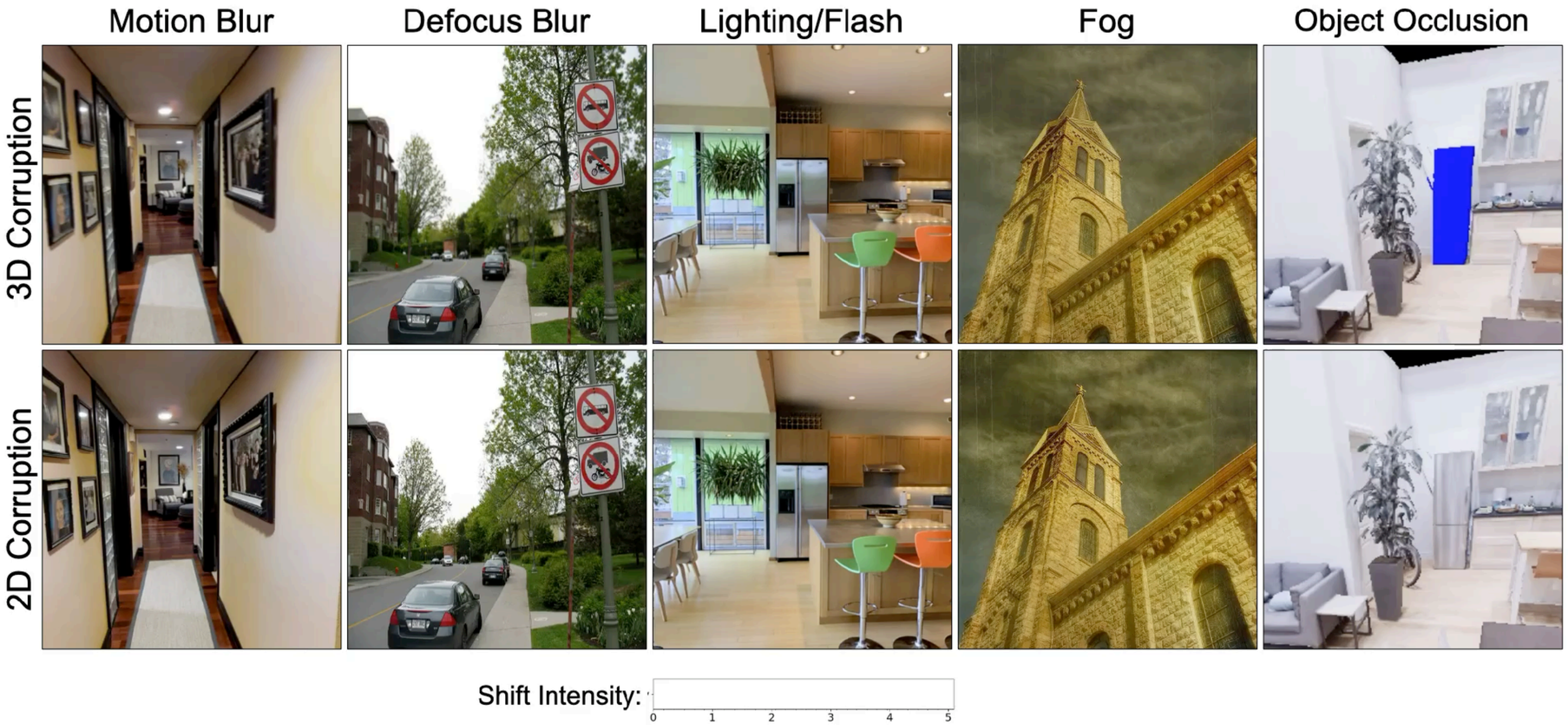

OASIS
AE
DSLR
YouTube
Input Image
Ground Truth
T+UNet
T+DPT
OASIS
O+DPT
Ours
Ours+X-TC

Pseudolabeling



https://omnidata.vision/demo/

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



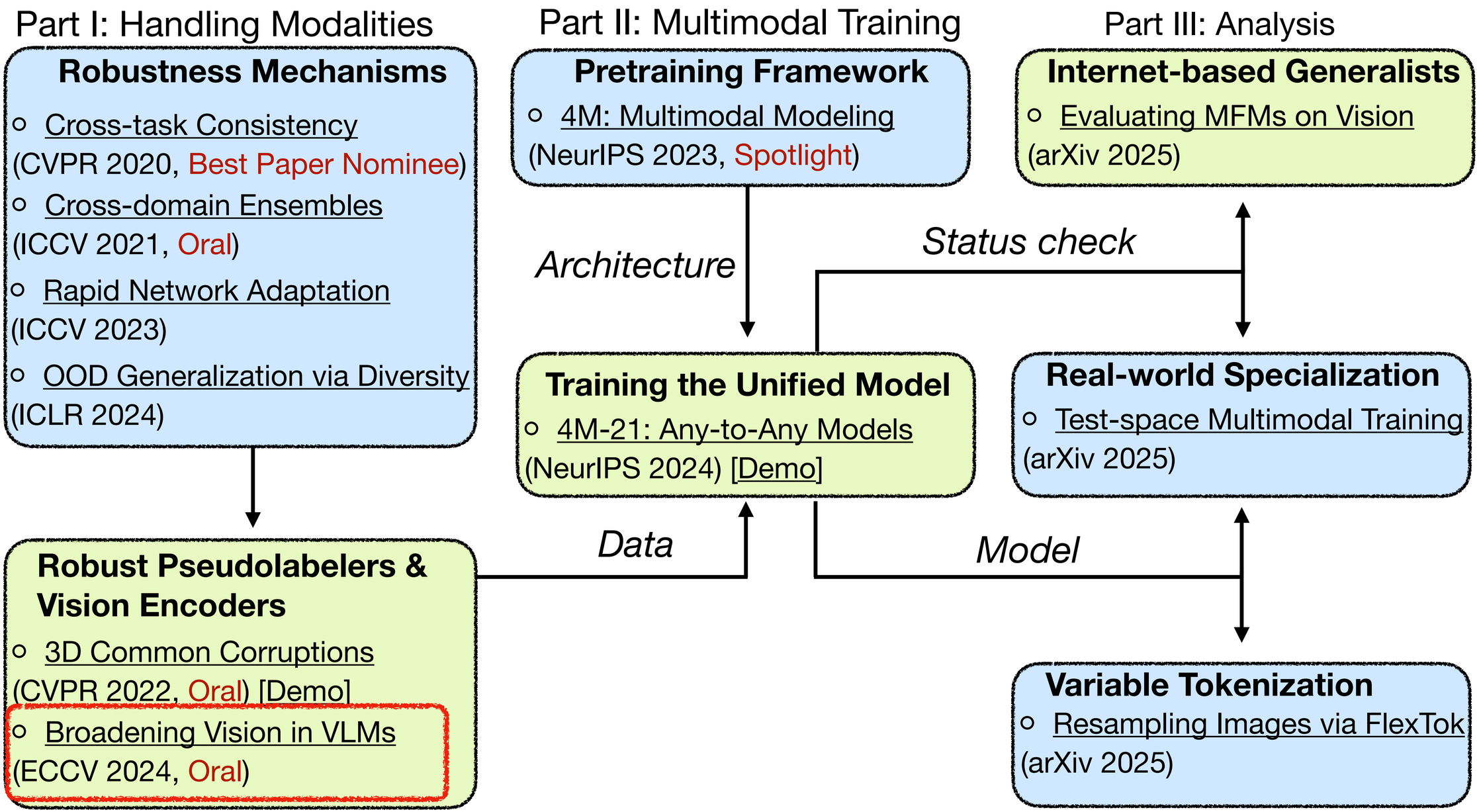
3D Common Corruptions and Data Augmentation

- *A challenging benchmark (in RobustBench)*
- *Extendable to standard datasets (ImageNet-3DCC)*
- *Strong pseudolabelers used by community (and by us)*

CVPR 2022 (Oral)

Code, Data, Models, Live Demo:
https://3dcommoncorruptions.epfl.ch/

Publication map & Thesis outline



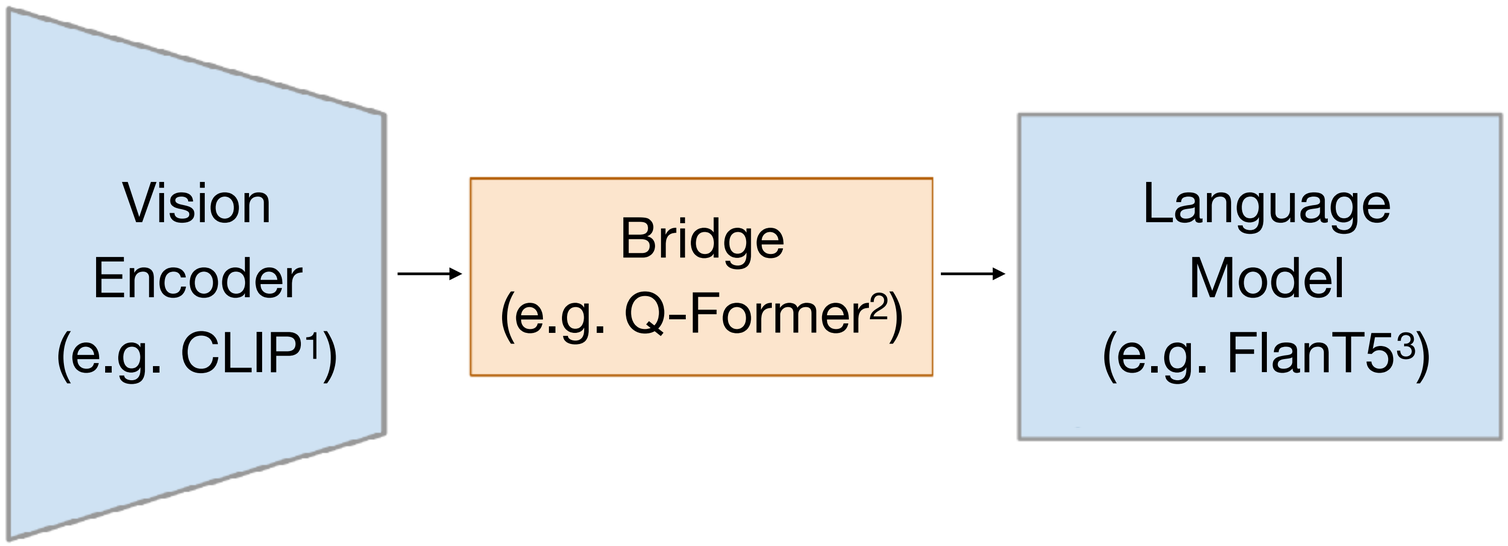
BRAVE🦁: Broadening the visual encoding of vision-language models

ECCV 2024 (Oral)

brave-vlms.epfl.ch

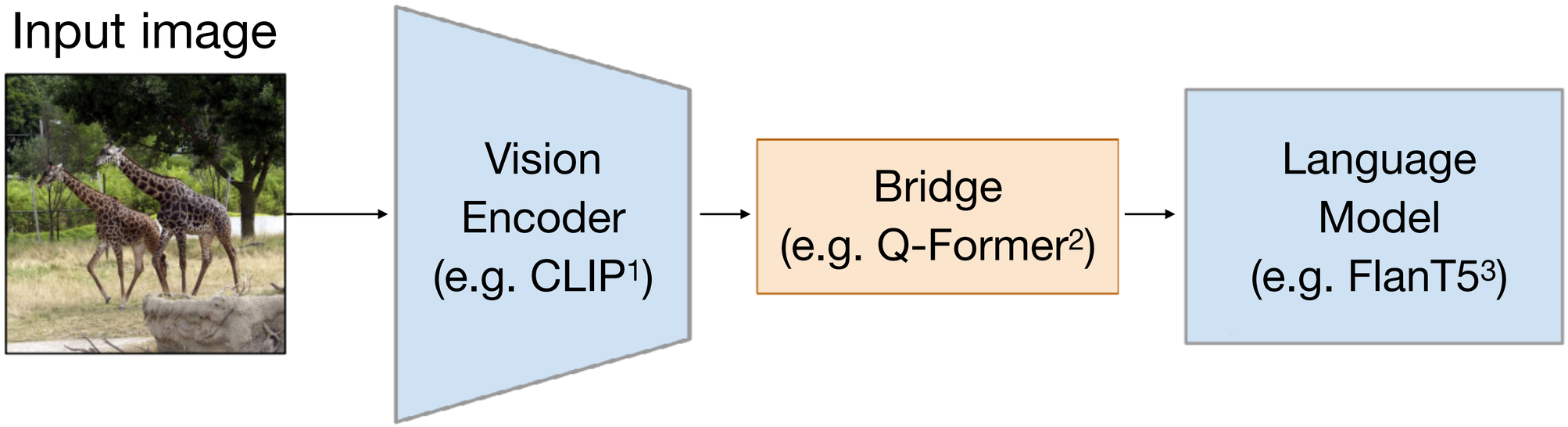
Google EPFL Swiss Federal Institute of Technology TUM

Vision-Language Models (VLMs)



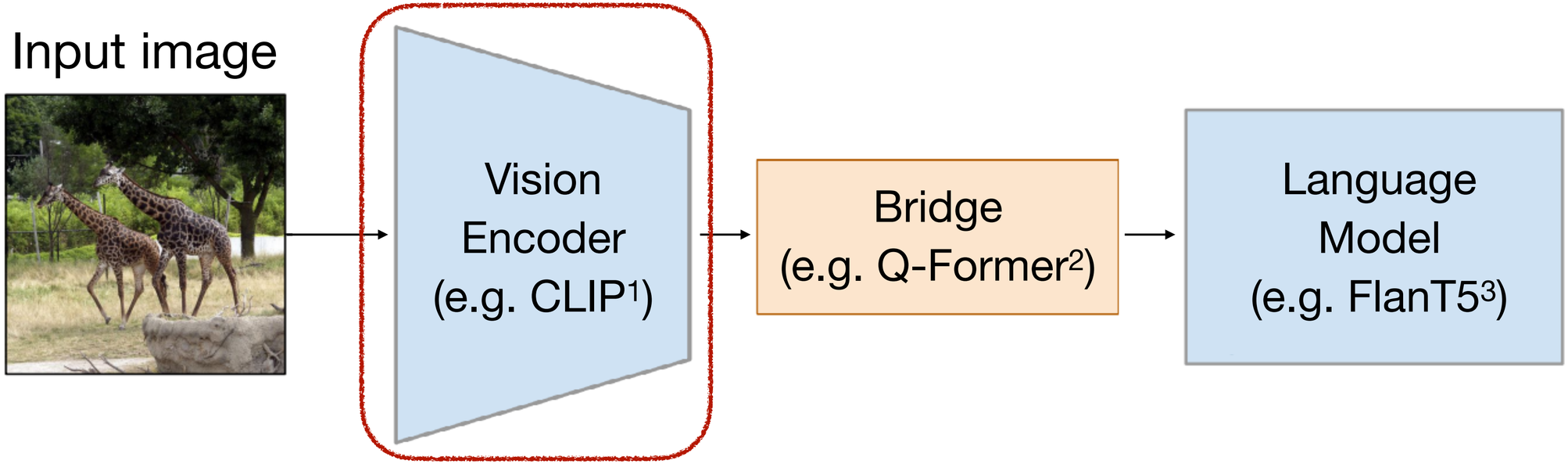
[1]Radford et al. 2021
[2]Li et al. 2023
[3]Chung et al. 2023

Vision-Language Models (VLMs)



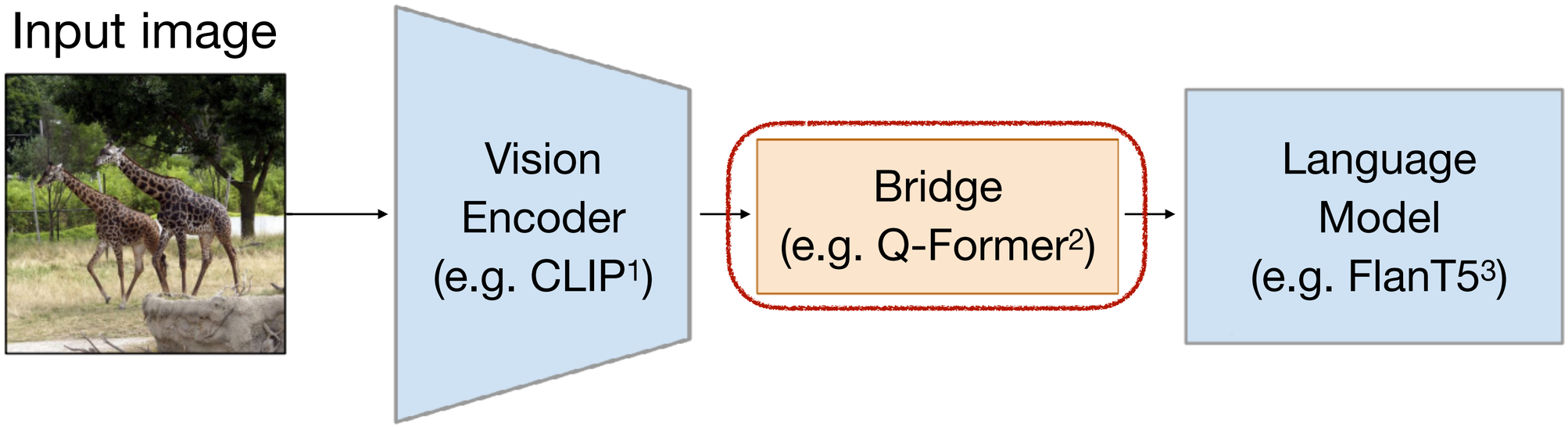
[1]Radford et al. 2021
[2]Li et al. 2023
[3]Chung et al. 2023

Vision-Language Models (VLMs)



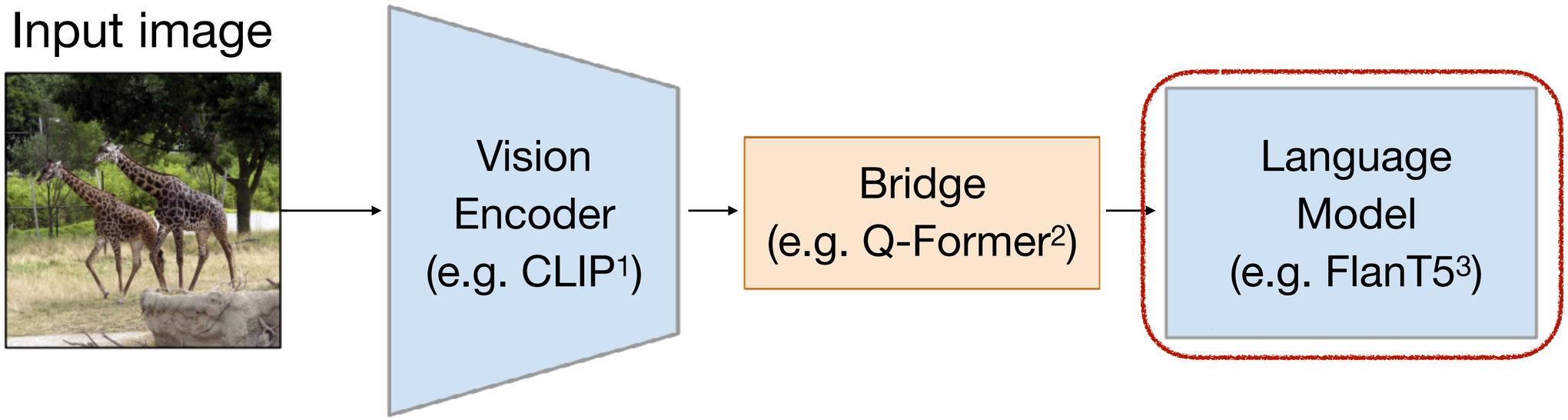
[1]Radford et al. 2021
[2]Li et al. 2023
[3]Chung et al. 2023

Vision-Language Models (VLMs)



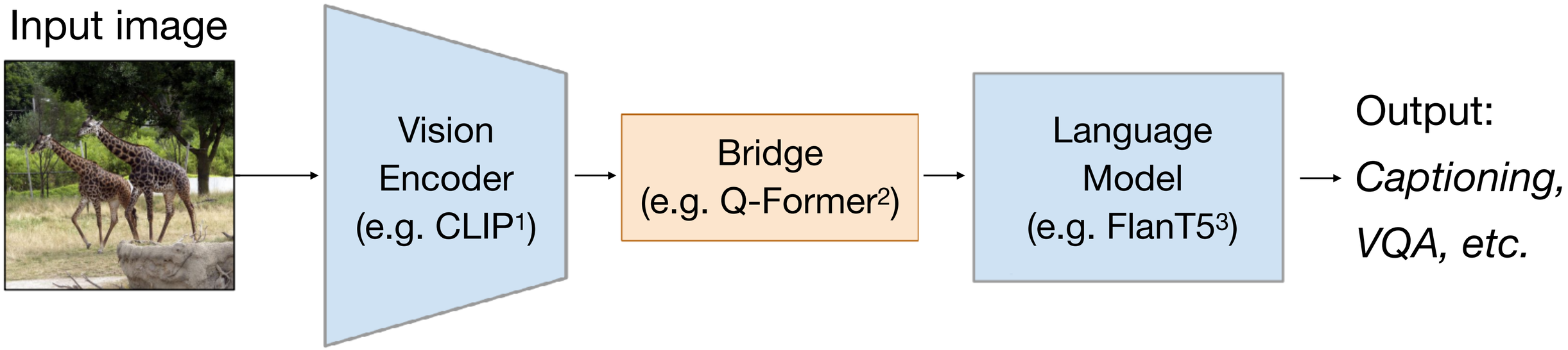
[1]Radford et al. 2021
[2]Li et al. 2023
[3]Chung et al. 2023

Vision-Language Models (VLMs)



[1]Radford et al. 2021
[2]Li et al. 2023
[3]Chung et al. 2023

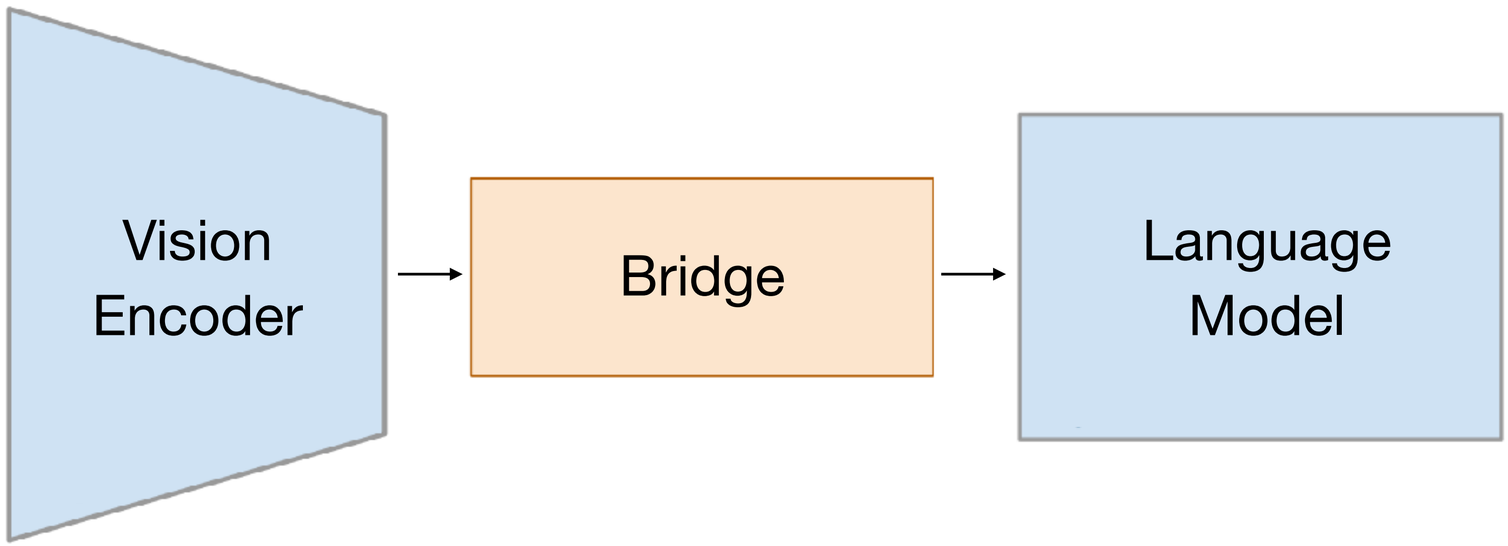
Vision-Language Models (VLMs)



[1]Radford et al. 2021
[2]Li et al. 2023
[3]Chung et al. 2023

VLMs have important shortcomings

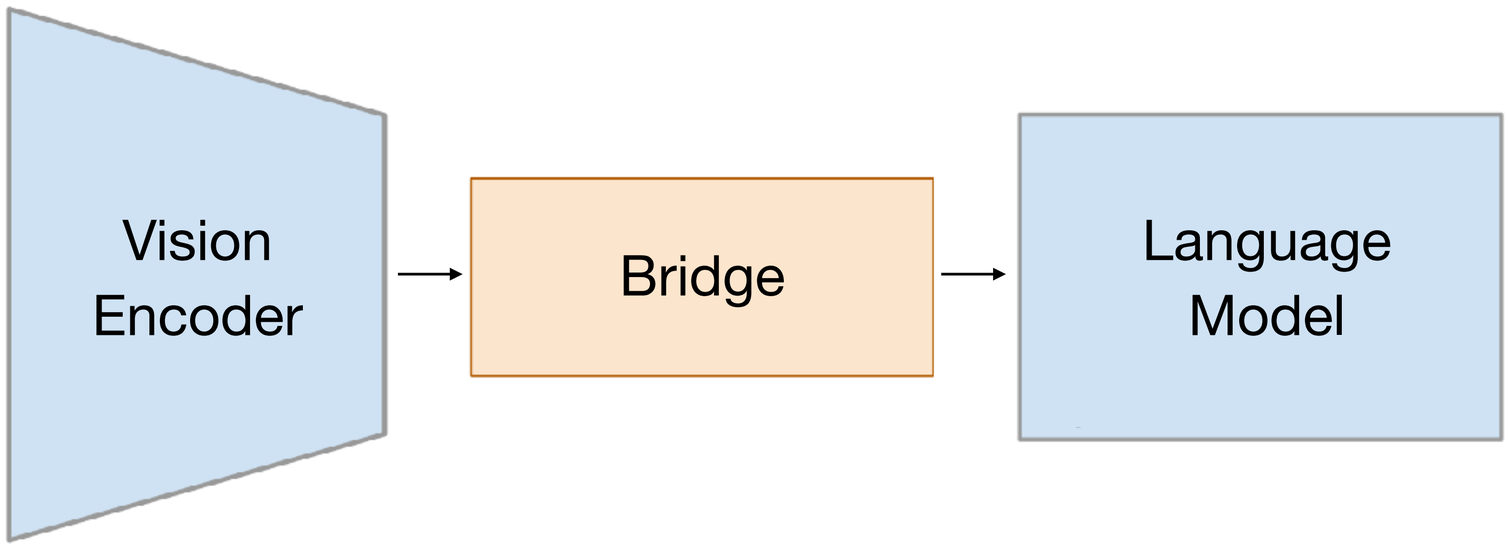
VLMs have important shortcomings



- Limited language capabilities
  - Hallucinations[1,2]
  - Logical faults[3,4]

[1]Bang et al. 2023
[2]Guo et al. 2023
[3]Shen et al. 2023
[4]Thorp et al. 2023

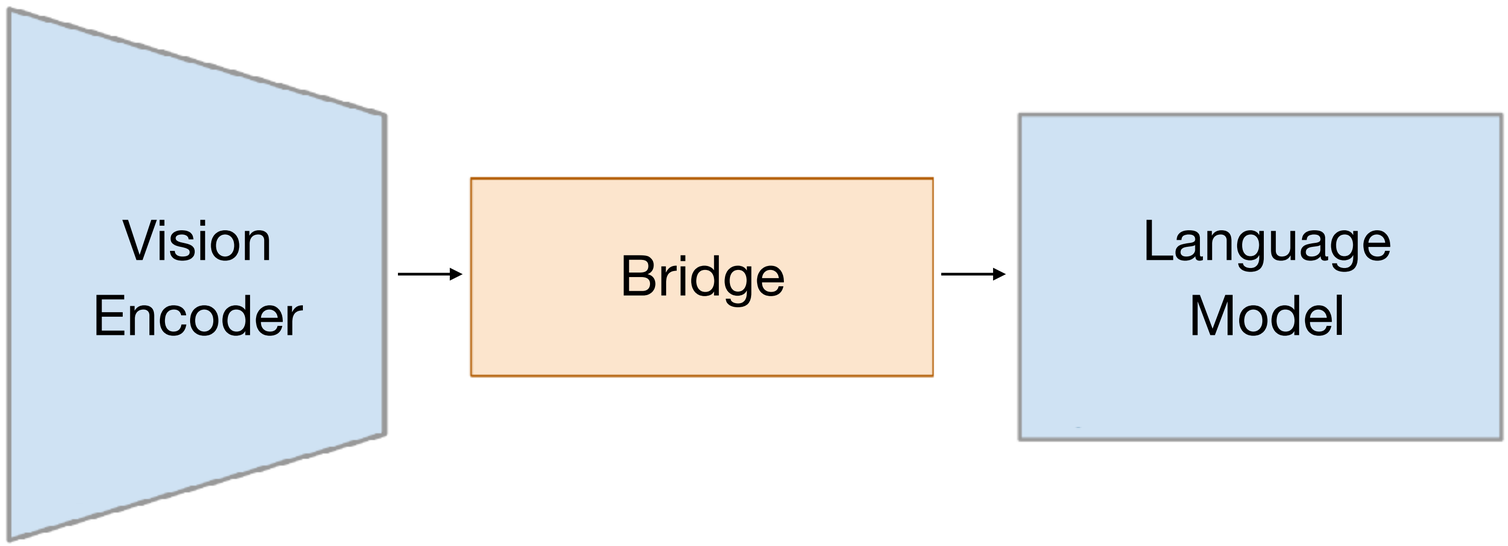
VLMs have important shortcomings



- Limited visual understanding
  - Feature “blindness”[5]
  - Visual hallucinations[6]
- Limited language capabilities
  - Hallucinations[1,2]
  - Logical faults[3,4]

[1]Bang et al. 2023
[2]Guo et al. 2023
[3]Shen et al. 2023
[4]Thorp et al. 2023
[5]Tong et al. 2024
[6]Li et al. 2023

VLMs have important shortcomings



- Limited visual understanding
  - Feature "blindness"[5]
  - Visual hallucinations[6]

- Limited language capabilities
  - Hallucinations[1,2]
  - Logical faults[3,4]

[1]Bang et al. 2023
[2]Guo et al. 2023
[3]Shen et al. 2023
[4]Thorp et al. 2023
[5]Tong et al. 2024
[6]Li et al. 2023

Blindness examples — confusing image pairs[1]

Is there a hand using the mouse in this image?

Are the butterfly's feet visible?

Is the door of the truck open?

[1]Tong et al. 2024

Blindness examples — confusing image pairs[1]

Is there a hand using the mouse in this image?

Are the butterfly's feet visible?

Is the door of the truck open?

InstructBLIP :

LLaVA-1.5 :

[1]Tong et al. 2024

Blindness examples — confusing image pairs[1]

Is there a hand using the mouse in this image?

Yes

No

Are the butterfly's feet visible?

Is the door of the truck open?

InstructBLIP :

LLaVA-1.5 :

[1]Tong et al. 2024

Blindness examples — confusing image pairs[1]

Is there a hand using the mouse in this image?

Yes

No

Are the butterfly's feet visible?

Is the door of the truck open?

InstructBLIP : Yes No ✅

LLaVA-1.5 : No No ❌

[1]Tong et al. 2024

Blindness examples — confusing image pairs[1]

Is there a hand using the mouse in this image?

Yes *No*

Are the butterfly's feet visible?

Yes *No*

Is the door of the truck open?

InstructBLIP :	Yes No ✅	Yes Yes ❌
LLaVA-1.5 :	No No ❌	Yes No ✅

[1]Tong et al. 2024

Blindness examples — confusing image pairs[1]

Is there a hand using the mouse in this image?

Are the butterfly's feet visible?

Is the door of the truck open?

	Yes	*No*	*Yes*	*No*	*Yes*	*No*
InstructBLIP :	Yes	No ✅	Yes	Yes ❌	Yes	Yes ❌
LLaVA-1.5 :	No	No ❌	Yes	No ✅	Yes	Yes ❌

[1]Tong et al. 2024

Blindness examples — confusing image pairs[1]

Is there a hand using the mouse in this image?

Are the butterfly's feet visible?

Is the door of the truck open?

	Yes	*No*	*Yes*	*No*	*Yes*	*No*
InstructBLIP : (EVA Encoder)	Yes	No ✅	Yes	Yes ❌	Yes	Yes ❌
LLaVA-1.5 : (CLIP Encoder)	No	No ❌	Yes	No ✅	Yes	Yes ❌

[1]Tong et al. 2024

Blindness examples — confusing image pairs[1]

Is there a hand using the mouse in this image?

Yes *No*

Are the butterfly's feet visible?

Yes *No*

Is the door of the truck open?

Yes *No*

InstructBLIP : (Ovr. Acc. 16.7%)	Yes No ✅	Yes Yes ❌	Yes Yes ❌
LLaVA-1.5 : (Ovr. Acc. 24.7%)	No No ❌	Yes No ✅	Yes Yes ❌

[1]Tong et al. 2024

Blindness examples — confusing image pairs[1]

Is there a hand using the mouse in this image?

Yes

No

Are the butterfly's feet visible?

Yes

No

Is the door of the truck open?

Yes

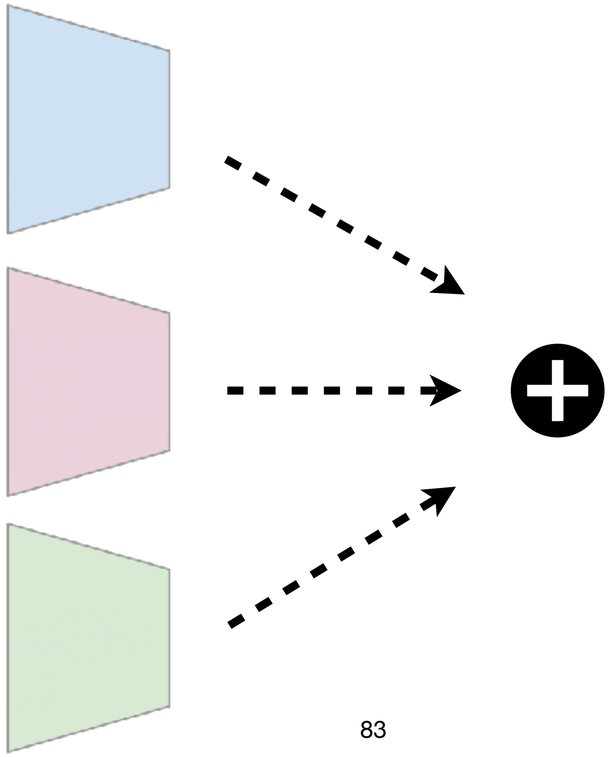
No

InstructBLIP : (Ovr. Acc. 16.7%)	Yes No ✅	Yes Yes ❌	Yes Yes ❌
LLaVA-1.5 : (Ovr. Acc. 24.7%)	No No ❌	Yes No ✅	Yes Yes ❌
BRAVE 🦁 : (Ovr. Acc. 42.0%)	Yes No ✅	Yes No ✅	Yes No ✅

[1]Tong et al. 2024

BRAVE: Broadening the visual encoding of VLMs

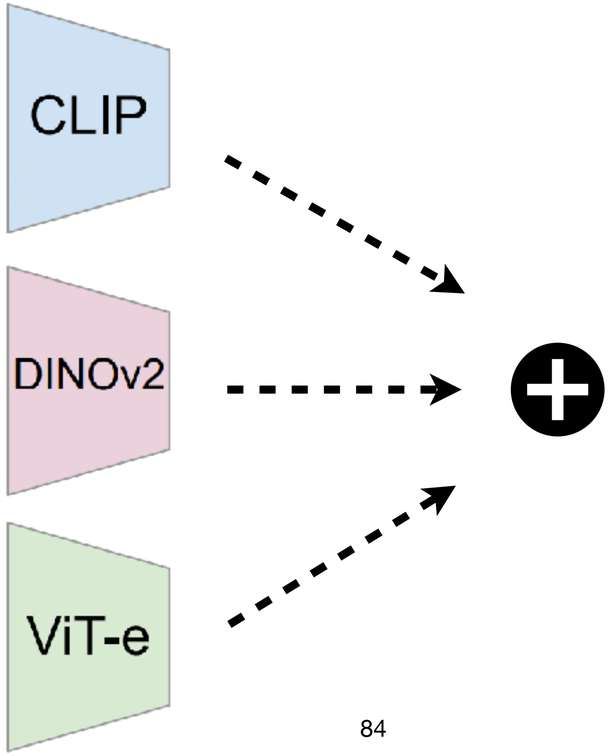
- Core idea from machine learning[1]
  - Different representations -> Different generalization properties
  - Ensemble to create a more complete representation



[1]Geman et al. 1992

BRAVE: Broadening the visual encoding of VLMs

- Core idea from machine learning[1]
  - Different representations -> Different generalization properties
  - Ensemble to create a more complete representation
  - Find the strongest set via **benchmarking**



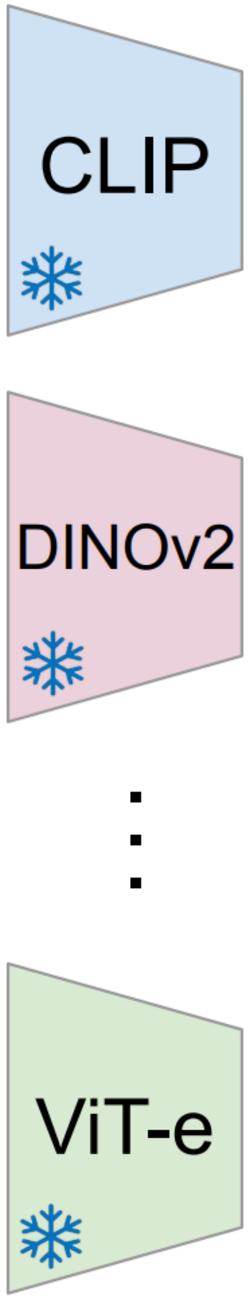
[1]Geman et al. 1992

Benchmarking vision encoders

- **8 different encoders**
  - Different objectives (Masked modeling, contrastive learning, etc.)
  - Different training datasets (LAION-2B, JFT-3B, etc.)
  - Different model sizes (300M to 4B)

Benchmarking vision encoders

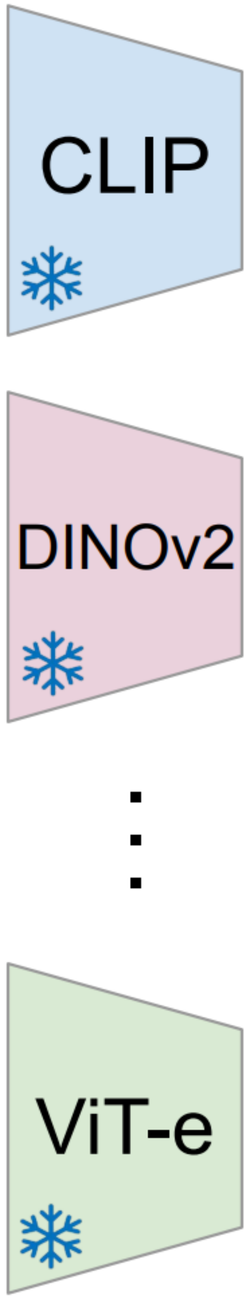
- **8 different encoders**
  - Different objectives (Masked modeling, contrastive learning, etc.)
  - Different training datasets (LAION-2B, JFT-3B, etc.)
  - Different model sizes (300M to 4B)
  - ***CLIP[1], EVA[2], DINOv2[3], SIGLIP[4], OpenCLIP[5], SILC[6], ViT-e[7], ViT-G[8]***
  - **Evaluation tasks: Captioning, VQA**



[1]Radford et al. 2021
[2]Fang et al. 2023
[3]Oquab et al. 2023
[4]Zhai et al. 2023
[5]Cherti et al. 2023
[6]Naeem et al. 2023
[7]Chen et al. 2022
[8]Zhai et al. 2022

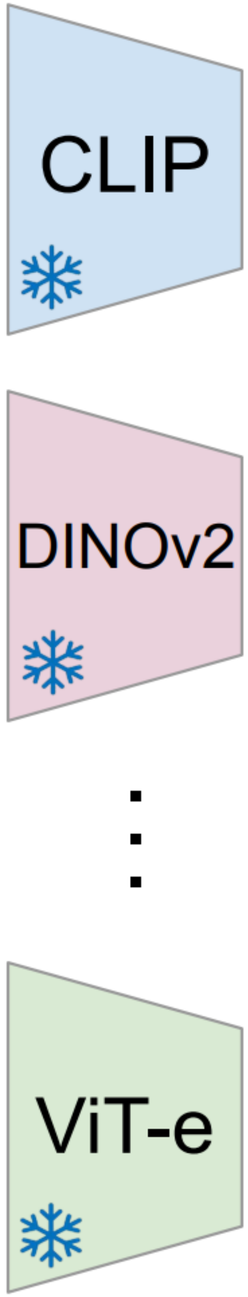
Benchmarking vision encoders Observations

- **No encoder perform consistently well**



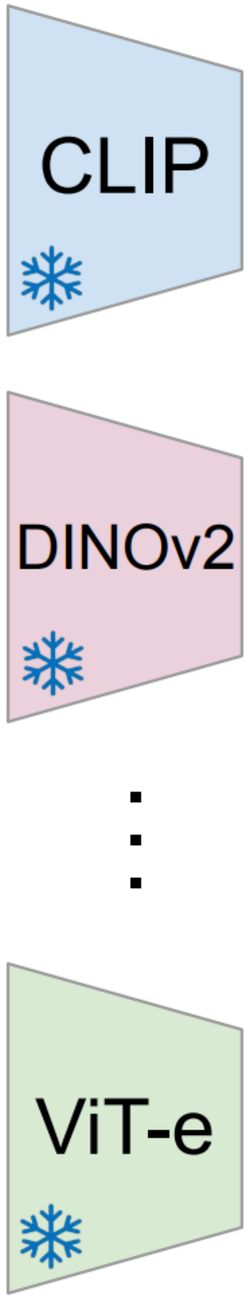
Benchmarking vision encoders Observations

- **No encoder perform consistently well**
  - Using a single encoder is inherently limited



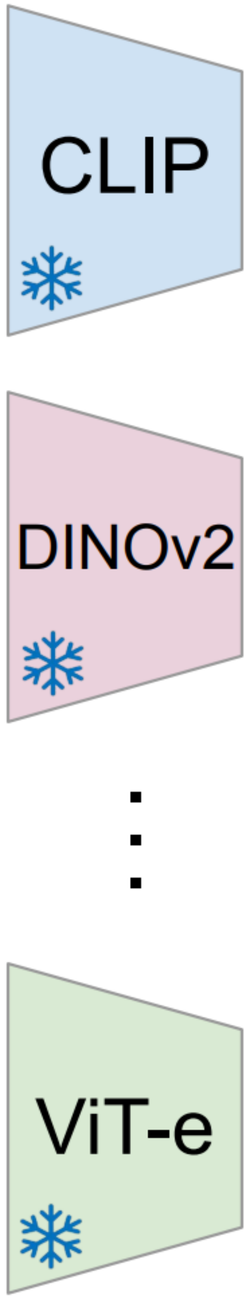
Benchmarking vision encoders
Observations

- **No encoder perform consistently well**
  - Using a single encoder is inherently limited
- **Encoders with different biases can perform similarly**



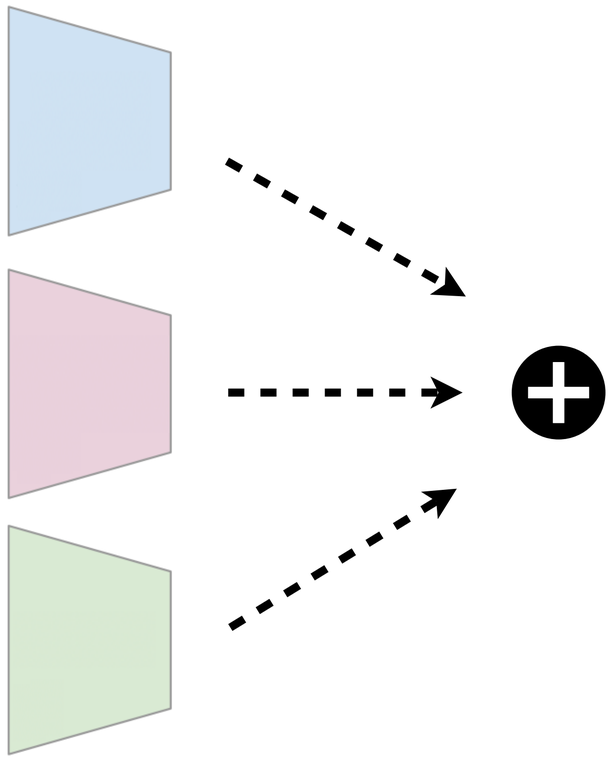
Benchmarking vision encoders
Observations

- **No encoder perform consistently well**
  - Using a single encoder is inherently limited
- **Encoders with different biases can perform similarly**
  - Different cues to exploit

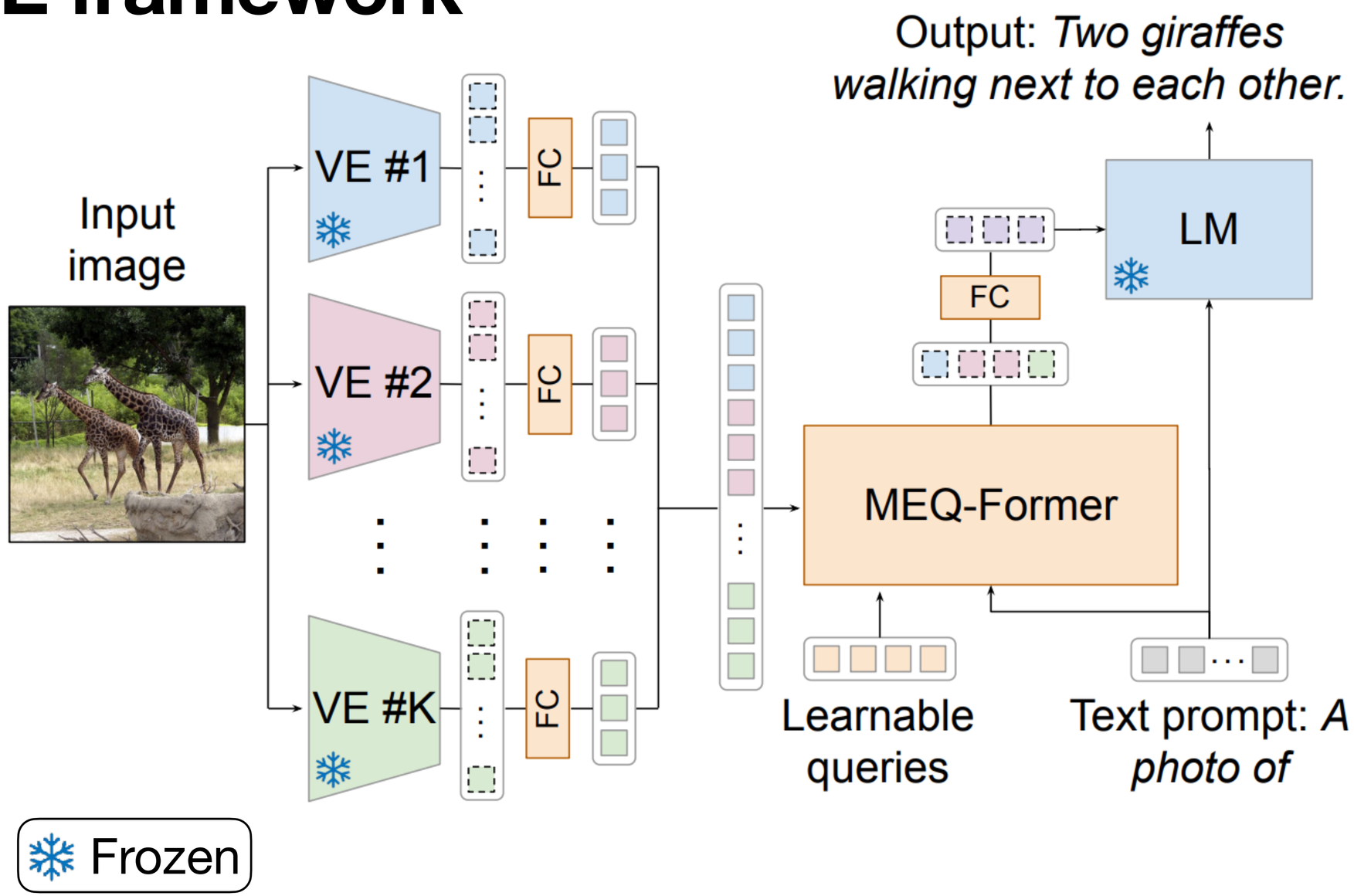


Can we broaden the visual capabilities of VLMs

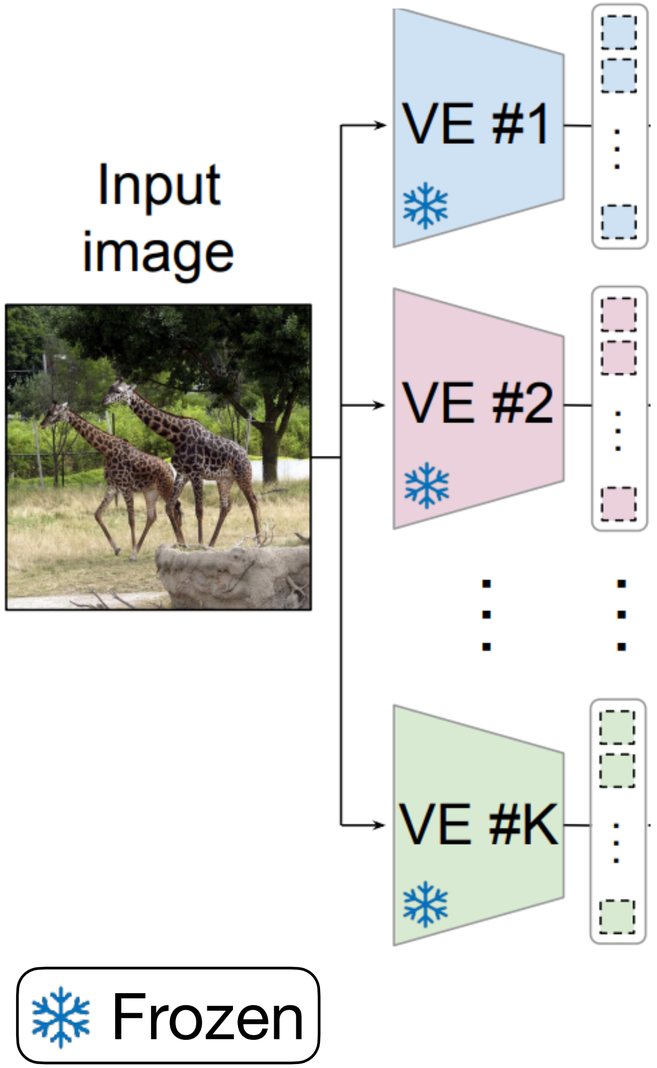
through combining vision encoders with different biases?



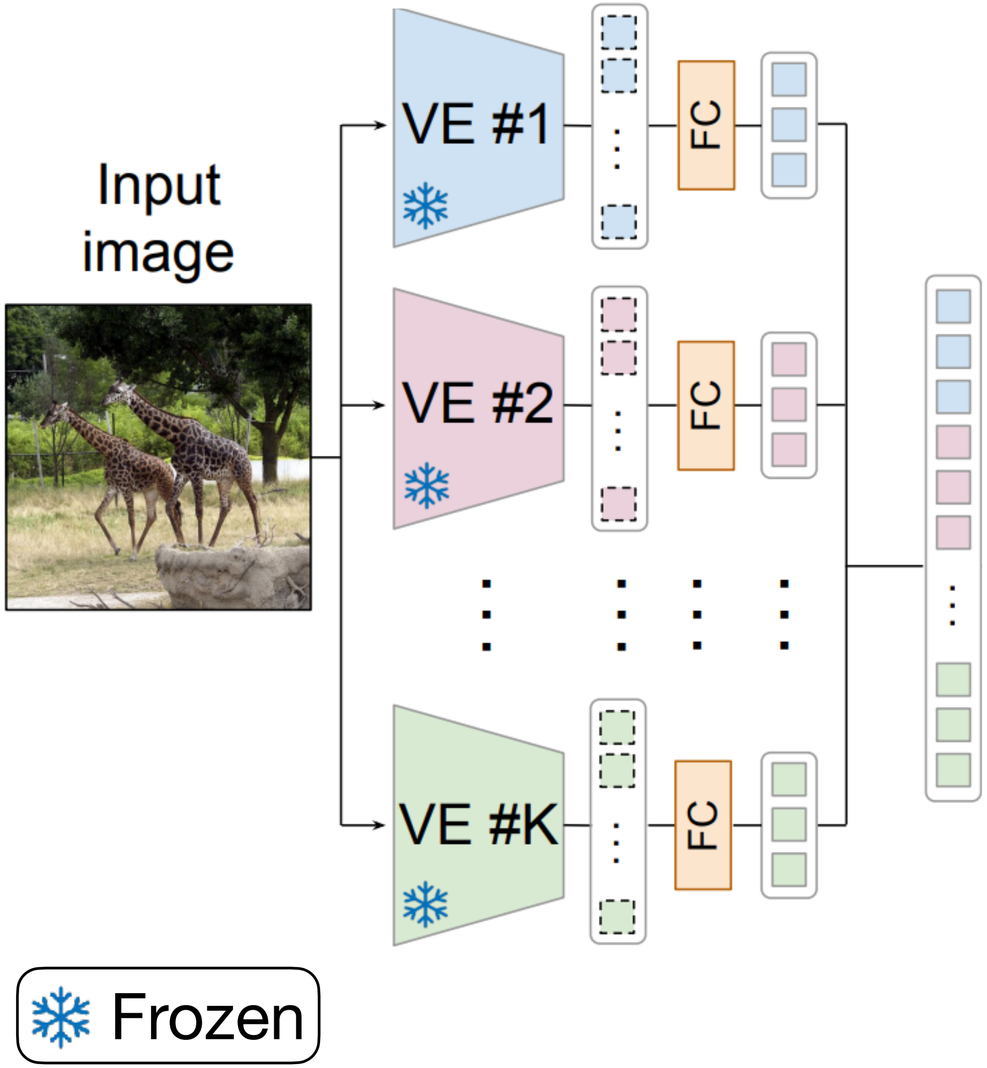
BRAVE framework



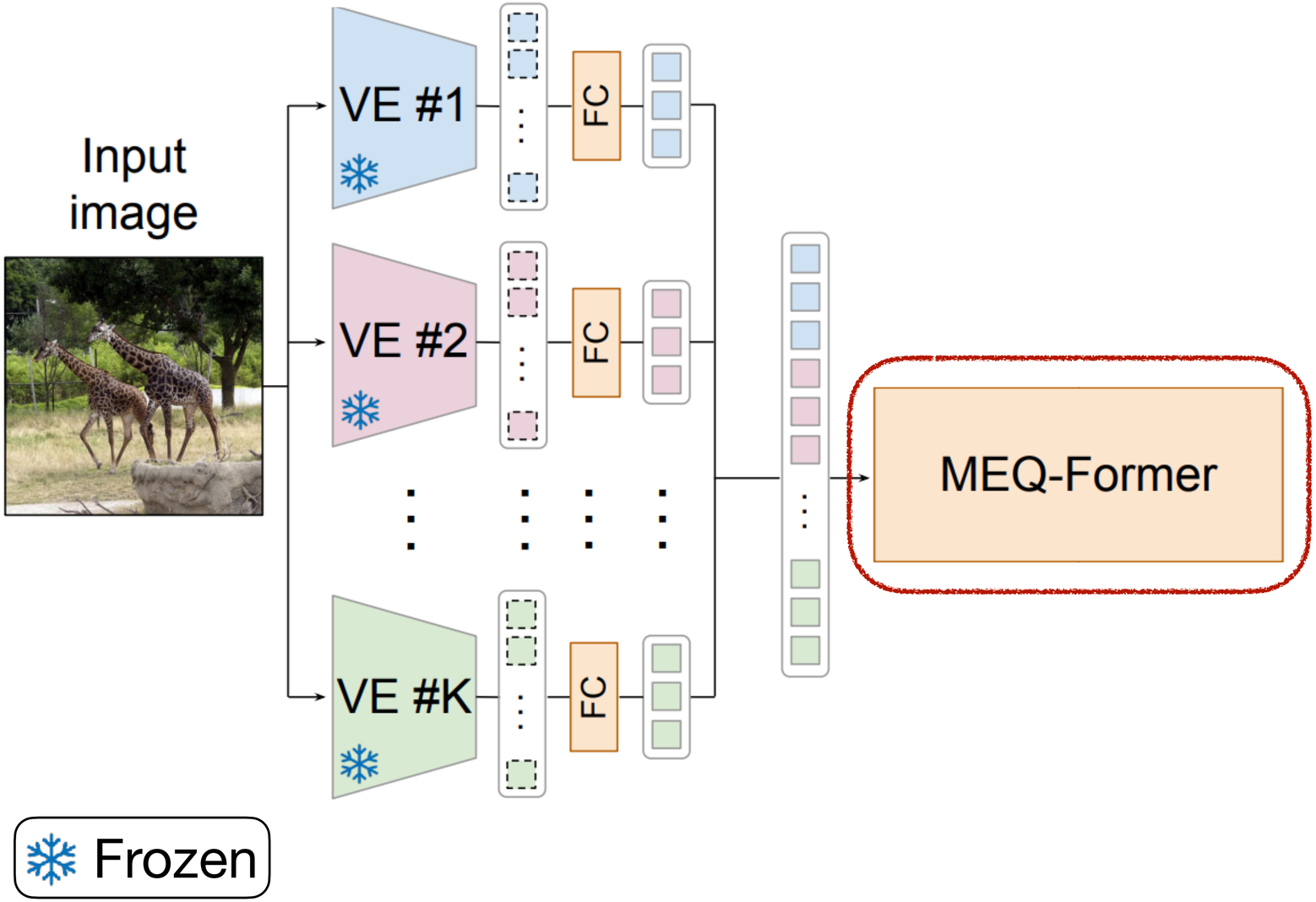
BRAVE framework



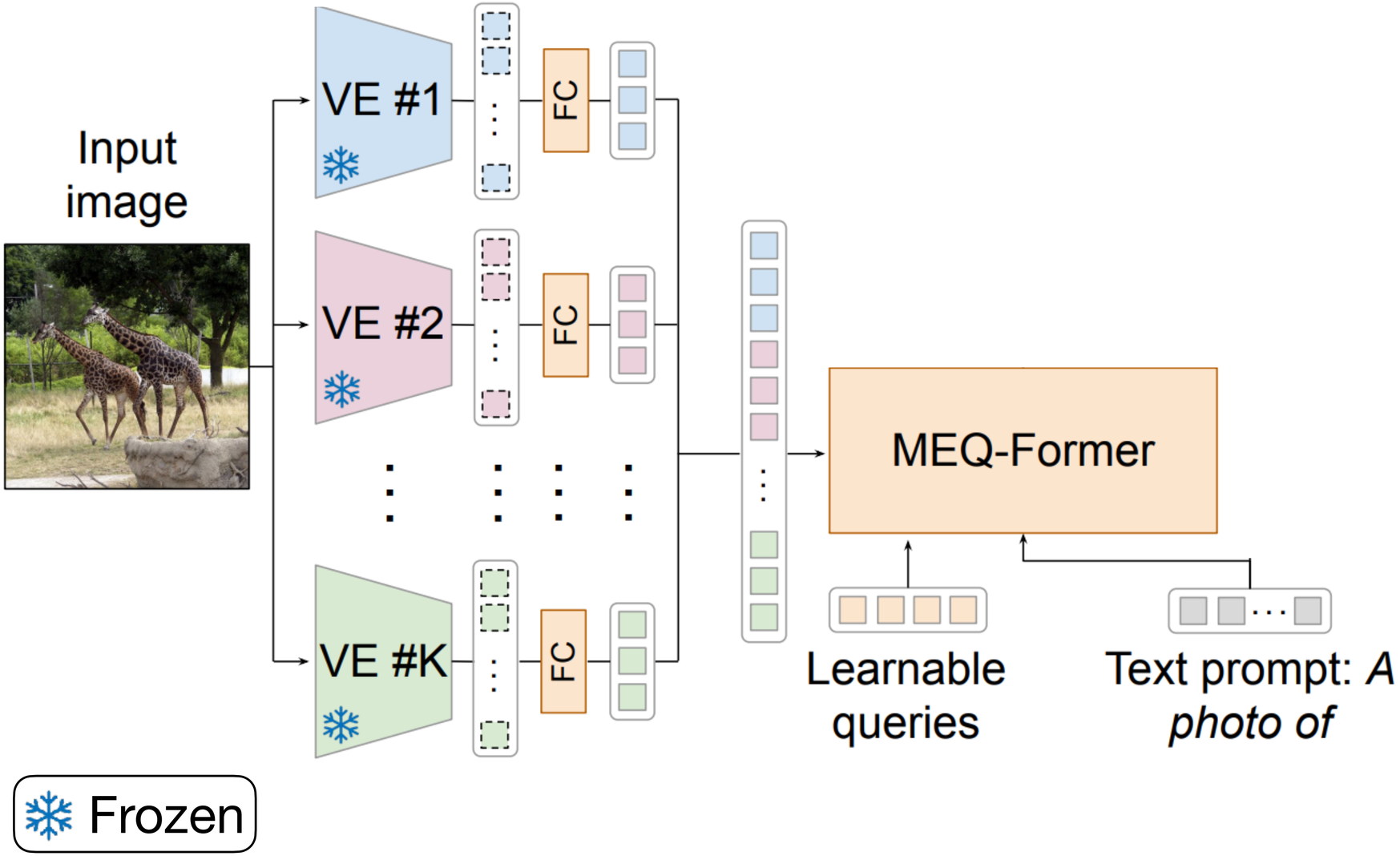
BRAVE framework



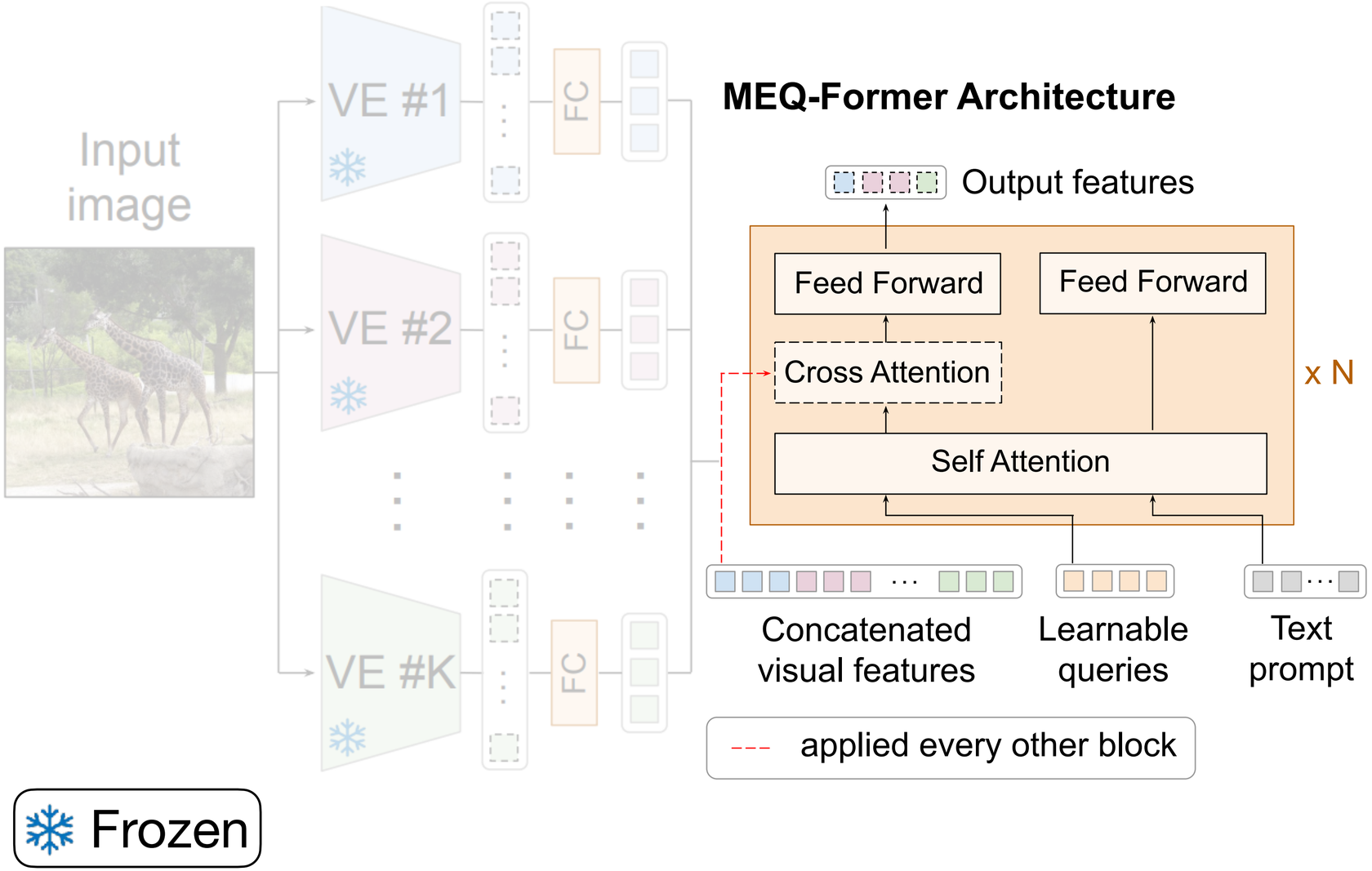
BRAVE framework



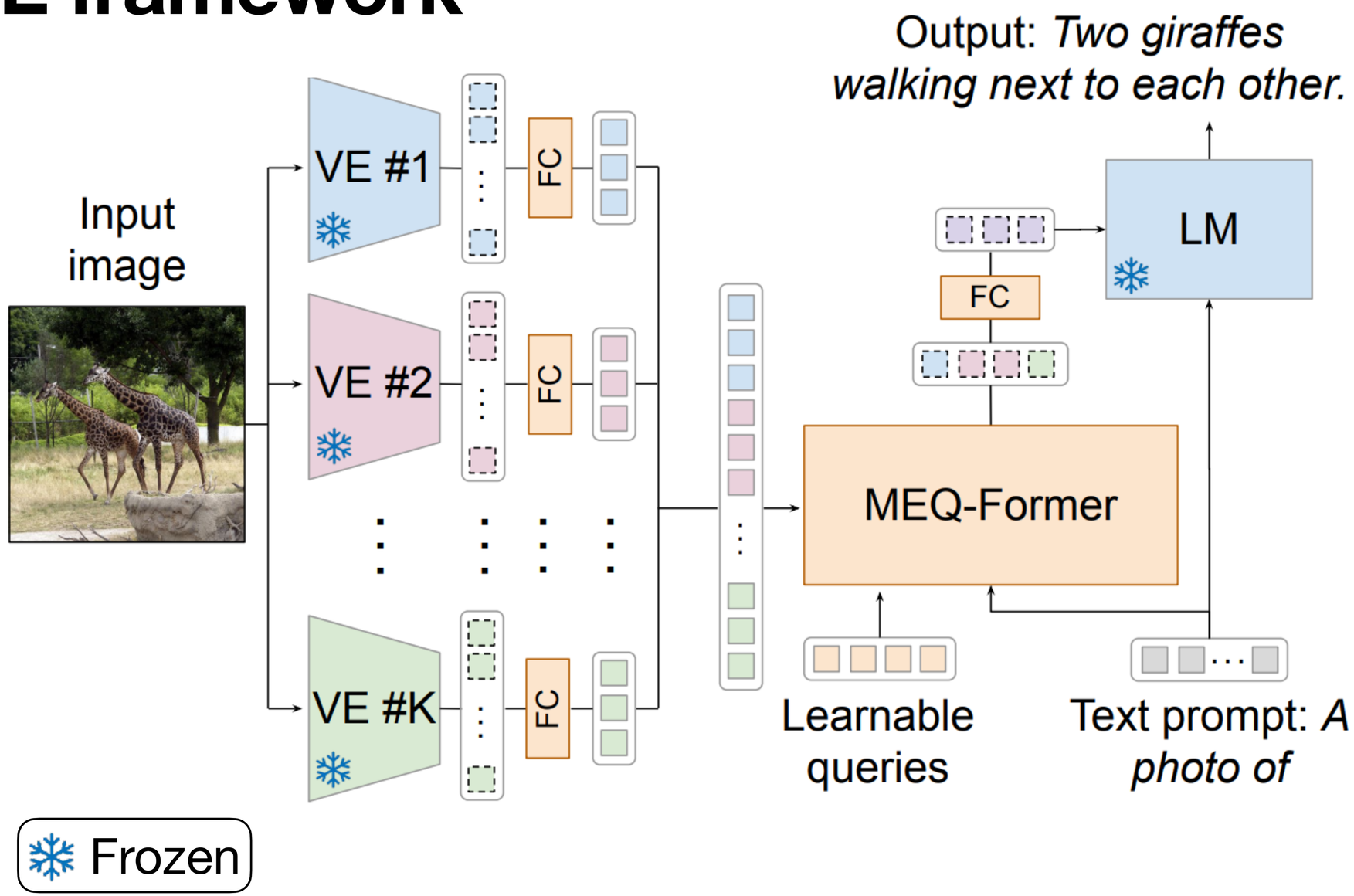
BRAVE framework



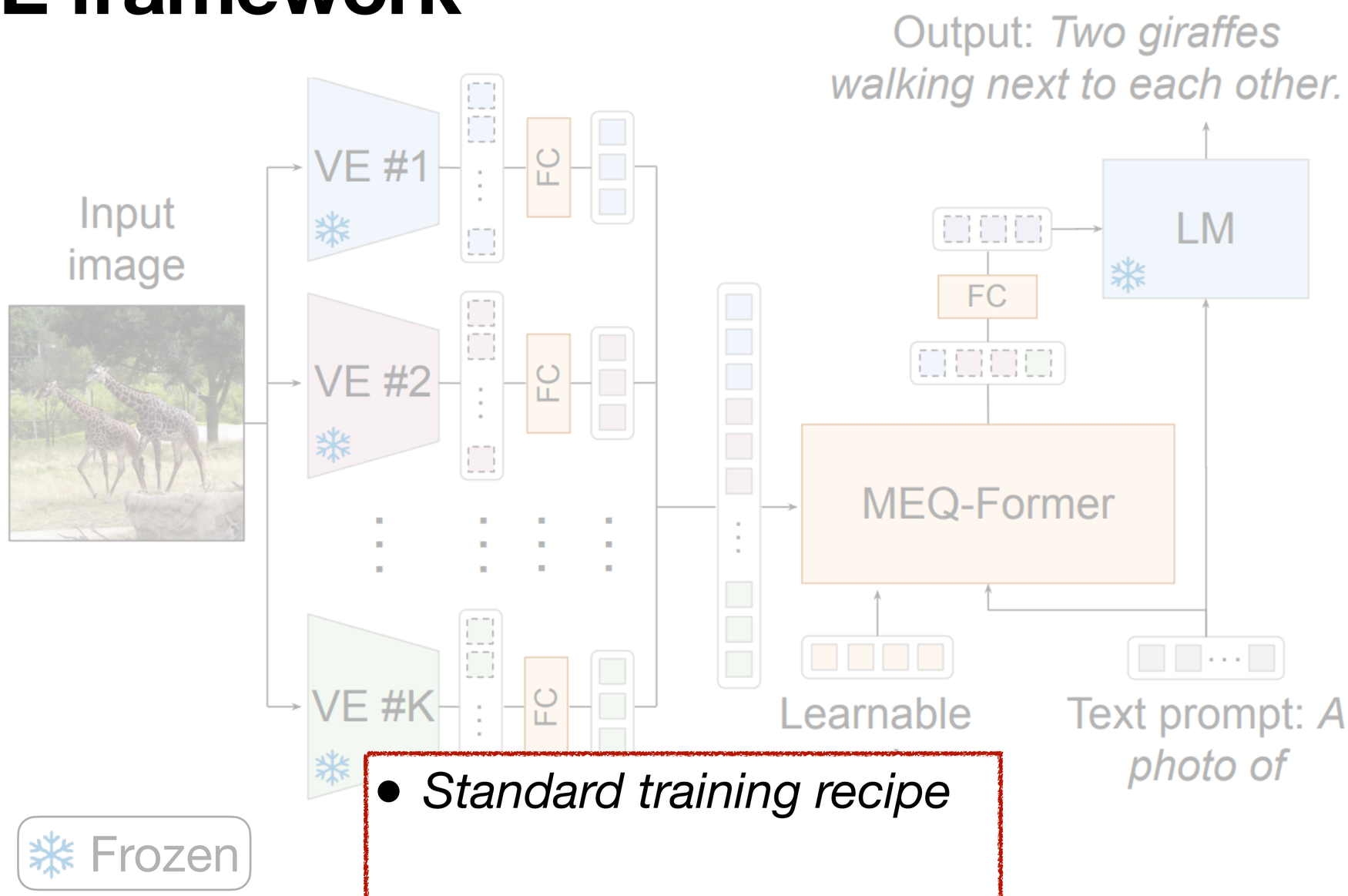
BRAVE framework



BRAVE framework

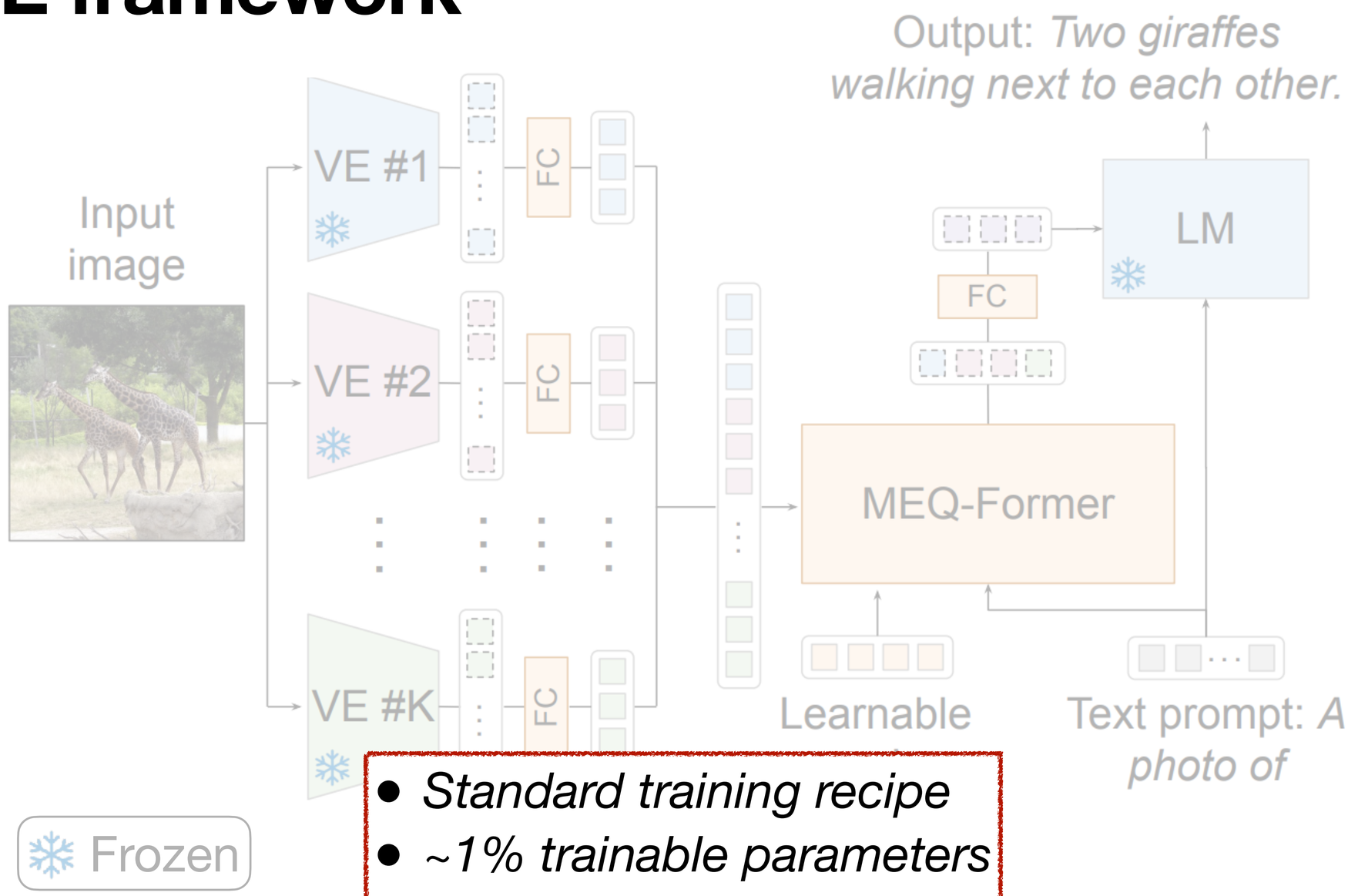


BRAVE framework



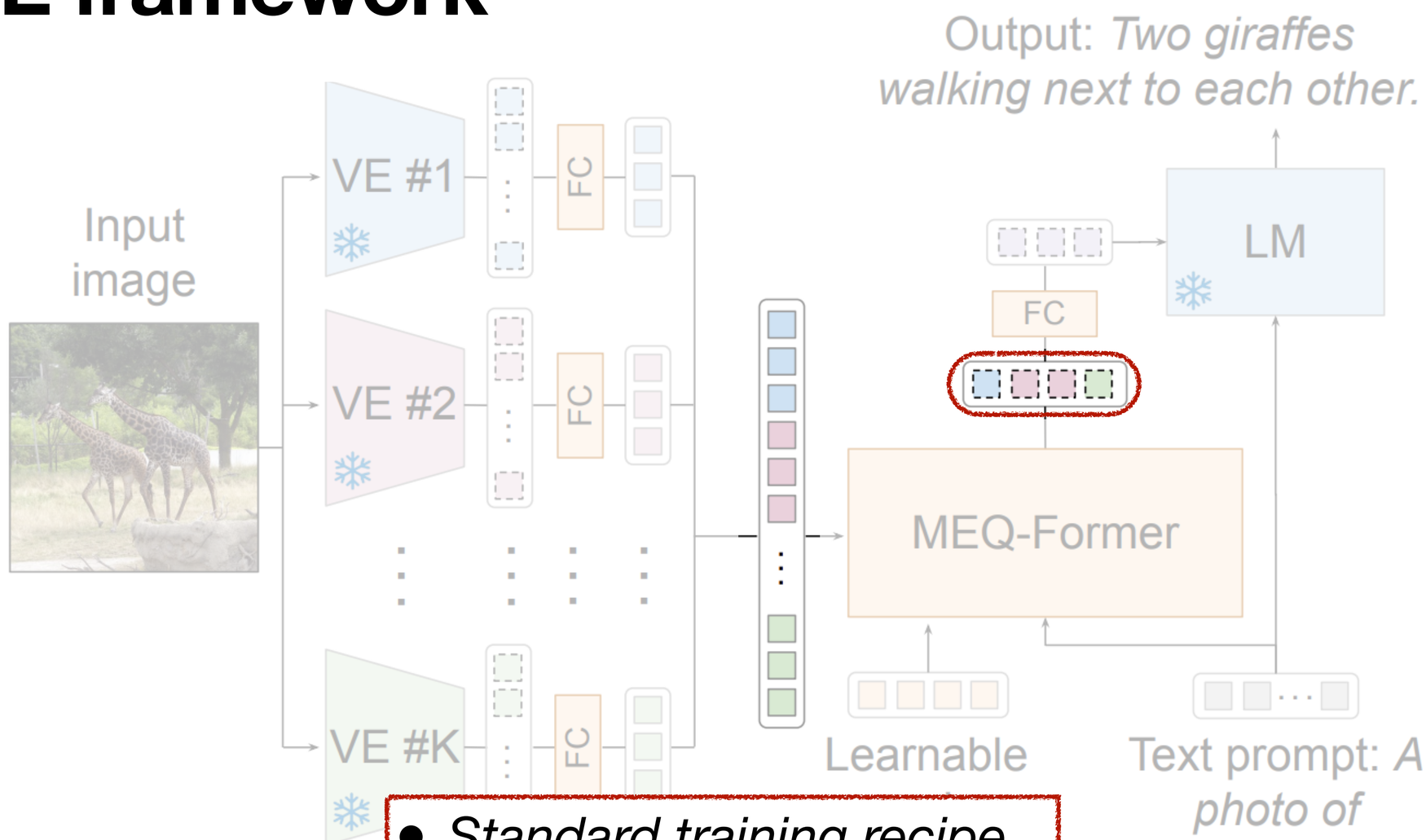
- *Standard training recipe*

BRAVE framework



- *Standard training recipe*
- *~1% trainable parameters*

BRAVE framework



- *Standard training recipe*
- *~1% trainable parameters*
- *Fixed-size visual prompt*

Key results

- State-of-the-art performance for captioning & VQA tasks

Key results

- State-of-the-art performance for captioning & VQA tasks

COCO[1]

General Captioning

Caption: *A large bus sitting next to a very tall building.*

[1]*Chen et al. 2015*

Key results

- State-of-the-art performance for captioning & VQA tasks

COCO[1]

General Captioning

Caption: *A large bus sitting next to a very tall building.*

NoCaps[2]

Novel Object Captioning

Caption: *A crab cake sandwich on a hamburger bun.*

[1]*Chen et al. 2015*
[2]*Agrawal et al. 2019*

Key results

- State-of-the-art performance for captioning & VQA tasks

COCO[1]	NoCaps[2]	VQAv2[3]
General Captioning	Novel Object Captioning	General VQA
Caption: *A large bus sitting next to a very tall building.*	Caption: *A crab cake sandwich on a hamburger bun.*	Q: *What color is the hydrant?* A: *Black and Yellow*

[1]*Chen et al. 2015*
[2]*Agrawal et al. 2019*
[3]*Goyal et al. 2017*

Key results

- State-of-the-art performance for captioning & VQA tasks

COCO[1]

General Captioning

Caption: *A large bus sitting next to a very tall building.*

NoCaps[2]

Novel Object Captioning

Caption: *A crab cake sandwich on a hamburger bun.*

VQAv2[3]

General VQA

Q: *What color is the hydrant?*
A: *Black and Yellow*

OKVQA[4]

Outside Knowledge

Q: *What company makes this sneakers? A: Converse*

[1]Chen et al. 2015
[2]Agrawal et al. 2019
[3]Goyal et al. 2017
[4]Marino et al. 2019

Key results

- State-of-the-art performance for captioning & VQA tasks

COCO[1]

General Captioning

Caption: *A large bus sitting next to a very tall building.*

NoCaps[2]

Novel Object Captioning

Caption: *A crab cake sandwich on a hamburger bun.*

VQAv2[3]

General VQA

Q: *What color is the hydrant?*
A: *Black and Yellow*

OKVQA[4]

Outside Knowledge

Q: *What company makes this sneakers? A: Converse*

GQA[5]

Spatial Reasoning

Q: *On which side of the image is the man? A: Right*

[1]Chen et al. 2015
[2]Agrawal et al. 2019
[3]Goyal et al. 2017
[4]Marino et al. 2019
[5]Hudson et al. 2019

Key results

- State-of-the-art performance for captioning & VQA tasks

COCO[1]

General Captioning

Caption: *A large bus sitting next to a very tall building.*

NoCaps[2]

Novel Object Captioning

Caption: *A crab cake sandwich on a hamburger bun.*

VQAv2[3]

General VQA

Q: *What color is the hydrant?*
A: *Black and Yellow*

OKVQA[4]

Outside Knowledge

Q: *What company makes this sneakers?* A: *Converse*

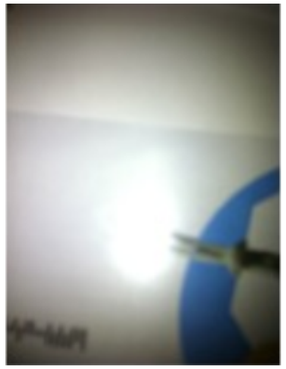
GQA[5]

Spatial Reasoning

Q: *On which side of the image is the man?* A: *Right*

VizWiz-QA[6]

Unanswerable Questions



Q: *Who is this mail for?*
A: *Unanswerable*

[1]Chen et al. 2015
[2]Agrawal et al. 2019
[3]Goyal et al. 2017
[4]Marino et al. 2019
[5]Hudson et al. 2019
[6]Gurari et al. 2018

Key results

- State-of-the-art performance for captioning & VQA tasks

COCO[1]

General Captioning

Caption: *A large b*
next to a very tall building.

NoCaps[2]

Novel Object Captioning

on a hamburger bun.

VQAv2[3]

General VQA

s the hydrant?
A: Black and Yellow

Covers different aspects of VLMs

OKVQA[4]

Outside Knowledge

Q: *What company makes this sneakers? A: Converse*

GQA[5]

Spatial Reasoning

Q: *On which side of the image is the man? A: Right*

VizWiz-QA[6]

Unanswerable Questions

Q: *Who is this mail for?*
A: Unanswerable

[1]*Chen et al. 2015*
[2]*Agrawal et al. 2019*
[3]*Goyal et al. 2017*
[4]*Marino et al. 2019*
[5]*Hudson et al. 2019*
[6]*Gurari et al. 2018*

Key results

- State-of-the-art performance for captioning & VQA tasks
- Improved robustness against hallucinations & feature blindness

POPE[1]

Visual Hallucination

Q: Is there a bottle in the image? A: No.

Q: Is there a surfboard in the image? A: No.

MMVP[2]

Confusing Pairs

Q: Are there cookies stacked on top of other cookies? A (Left): Yes - A (Right): No.

[1]Li et al. 2023
[2]Tong et al. 2024

Key results

- State-of-the-art performance for captioning & VQA tasks
- Improved robustness against hallucinations & feature blindness

POPE[1]

Visual Hallucination

Q: Is there a bottle in the image? A: No.

Q: Is there a surfboard in the image? A: No.

MMVP[2]

Confusing Pairs

Q: Are there cookies stacked on top of other cookies? A (Left): Yes - A (Right): No.

[1]Li et al. 2023
[2]Tong et al. 2024

Key results

- State-of-the-art performance for captioning & VQA tasks
- Improved robustness against hallucinations & feature blindness

POPE[1]

Visual Hallucination

Q: Is there a bottle in the image? A: No.

Q: Is there a surfboard in the image? A: No.

MMVP[2]

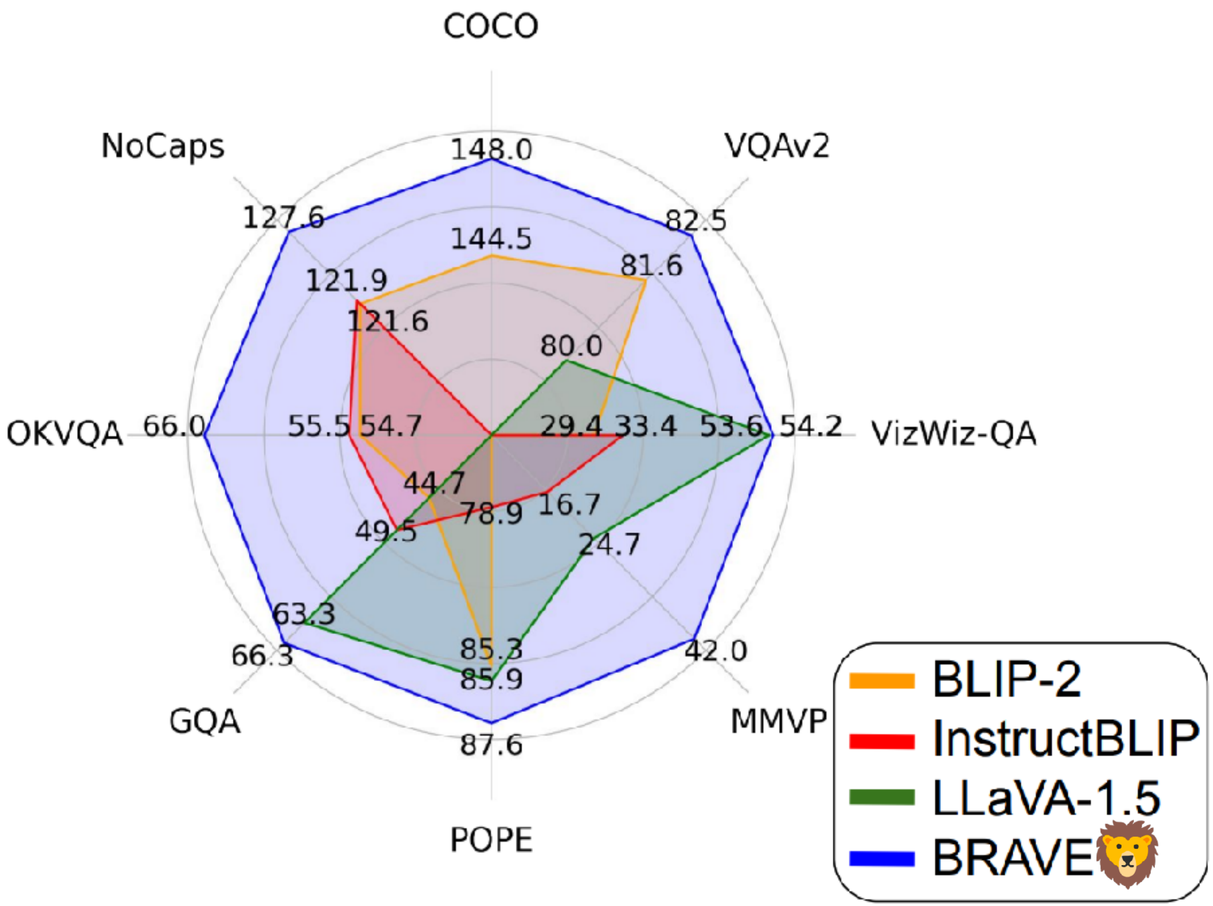
Confusing Pairs

Q: Are there cookies stacked on top of other cookies? A (Left): Yes - A (Right): No.

[1]Li et al. 2023
[2]Tong et al. 2024

Key results

- State-of-the-art performance for captioning & VQA tasks
- Improved robustness against hallucinations & feature blindness



Quantitative results — Captioning

Method	# params		COCO (fine-tuned)	NoCaps (zero-shot, val)		NoCaps (zero-shot, test)	
	Trainable	Total	Karpathy test	out-domain	overall	out-domain	overall
Flamingo [3]	10.6B	80B	138.1	-	-	-	-
SimVLM [85]	632M	632M	143.3	113.7	112.2	-	110.3
Qwen-VL [5]	9.6B	9.6B	-	-	121.4	-	-
BLIP-2 [53]	1.1B	4.1B	144.5	124.8	121.6	-	-
InstructBLIP [23]	188M	14.2B	-	-	121.9	-	-
CoCa [90]	2.1B	2.1B	143.6	-	122.4	-	120.6
GiT2 [81]	5.1B	5.1B	145.0	130.6	126.9	122.3	124.8
PaLI-17B [17]	16.9B	16.9B	**149.1**	-	127.0	126.7	124.4
BRAVE	**116M**	10.3B	148.0	**133.3**	**127.6**	**127.1**	**125.6**

Quantitative results — Captioning

Method	# params		COCO (fine-tuned)	NoCaps (zero-shot, val)		NoCaps (zero-shot, test)	
	Trainable	Total	Karpathy test	out-domain	overall	out-domain	overall
Flamingo [3]	10.6B	80B	138.1	-	-	-	-
SimVLM [85]	632M	632M	143.3	113.7	112.2	-	110.3
Qwen-VL [5]	9.6B	9.6B	-	-	121.4	-	-
BLIP-2 [53]	1.1B	4.1B	144.5	124.8	121.6	-	-
InstructBLIP [23]	188M	14.2B	-	-	121.9	-	-
CoCa [90]	2.1B	2.1B	143.6	-	122.4	-	120.6
GiT2 [81]	5.1B	5.1B	145.0	130.6	126.9	122.3	124.8
PaLI-17B [17]	16.9B	16.9B	**149.1**	-	127.0	126.7	124.4
BRAVE 🦁	**116M**	10.3B	148.0	**133.3**	**127.6**	**127.1**	**125.6**

Quantitative results — VQA

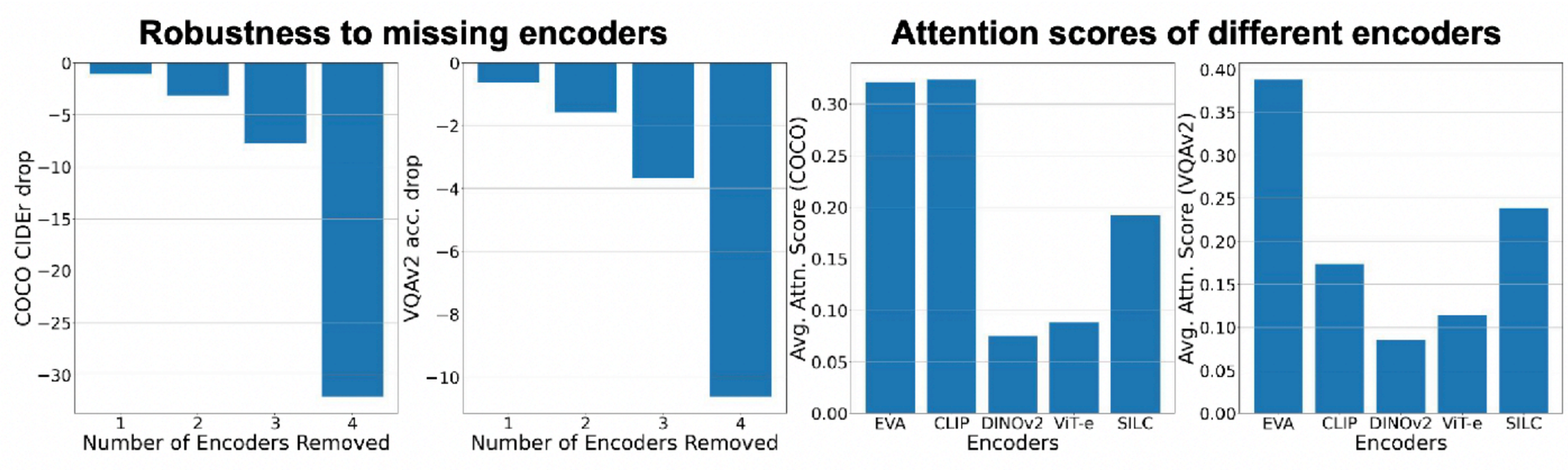
Method	# params		Fine-tuned			Zero-shot			
	Trainable	Total	VQAv2 test-dev	OKVQA val	GQA test-dev	VizWiz-QA test-dev	GQA test-dev	MMVP test	POPE test
SimVLM [91]	632M	632M	80.0	-	-	-	-	-	-
Flamingo [3]	10.2B	80B	82.0	57.8	-	31.6	-	-	-
MiniGPT-v2 [14]	7B	8B	-	57.8	60.1	53.6	-	-	-
GiT2 [87]	5.1B	5.1B	81.7	-	-	-	-	-	-
Qwen-VL [6]	9.6B	9.6B	79.5	58.6	59.3	35.2	-	-	-
SPHINX-2k [61]	13B	16.5B	80.7	62.6	63.1	44.9	-	-	87.2
PaLI-17B [19]	16.9B	16.9B	**84.3**	64.5	-	-	-	-	-
BLIP-2 [56]	1.2B	12.1B	81.6	54.7	-	29.4	44.7	-	85.3
InstructBLIP [25]	188M	14.2B	-	55.5	-	33.4	49.5	16.7	78.9
ShareGPT4V [16]	13.4B	13.4B	81.0	-	64.8	-	-	-	-
LLaVA$^{1.5}$ [64]	13B	13.4B	80.0	-	63.3	53.6	-	24.7	85.9
LLaVA$^{1.6}$ [65]	13B	13.4B	-	46.3	65.4	-	-	-	86.3
LLaVA$^{1.5}$ (I-MoF) [84]	13B	13.6B	79.3	-	-	-	-	31.3	86.7
BRAVE	3B	10.3B	82.5	**66.0**	**66.3**	**54.2**	**52.7**	**42.0**	**87.6**

Quantitative results — VQA

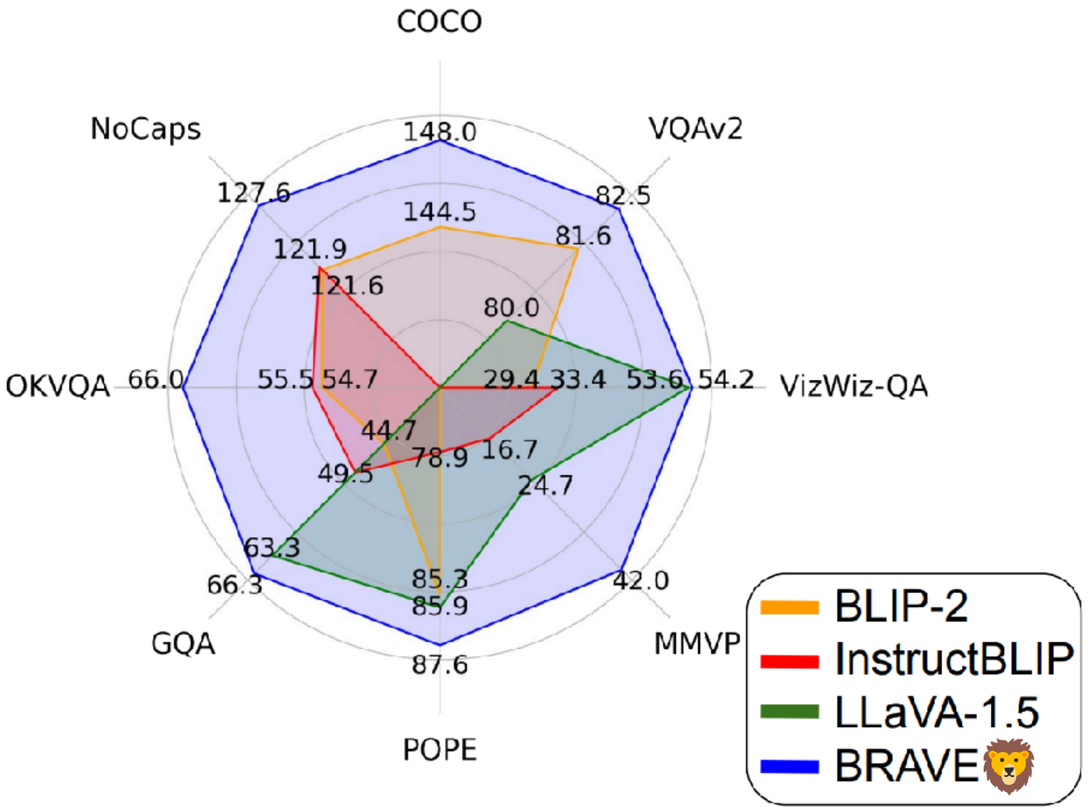
Method	# params		Fine-tuned			Zero-shot			
	Trainable	Total	VQAv2 test-dev	OKVQA val	GQA test-dev	VizWiz-QA test-dev	GQA test-dev	MMVP test	POPE test
SimVLM [91]	632M	632M	80.0	-	-	-	-	-	-
Flamingo [3]	10.2B	80B	82.0	57.8	-	31.6	-	-	-
MiniGPT-v2 [14]	7B	8B	-	57.8	60.1	53.6	-	-	-
GiT2 [87]	5.1B	5.1B	81.7	-	-	-	-	-	-
Qwen-VL [6]	9.6B	9.6B	79.5	58.6	59.3	35.2	-	-	-
SPHINX-2k [61]	13B	16.5B	80.7	62.6	63.1	44.9	-	-	87.2
PaLI-17B [19]	16.9B	16.9B	**84.3**	64.5	-	-	-	-	-
BLIP-2 [56]	1.2B	12.1B	81.6	54.7	-	29.4	44.7	-	85.3
InstructBLIP [25]	188M	14.2B	-	55.5	-	33.4	49.5	16.7	78.9
ShareGPT4V [16]	13.4B	13.4B	81.0	-	64.8	-	-	-	-
LLaVA$^{1.5}$ [64]	13B	13.4B	80.0	-	63.3	53.6	-	24.7	85.9
LLaVA$^{1.6}$ [65]	13B	13.4B	-	46.3	65.4	-	-	-	86.3
LLaVA$^{1.5}$ (I-MoF) [84]	13B	13.6B	79.3	-	-	-	-	31.3	86.7
BRAVE	3B	10.3B	82.5	**66.0**	**66.3**	**54.2**	**52.7**	**42.0**	**87.6**

More results & analysis

- Qualitative results on captioning and VQA
- Ablations of design choices (training data, fine-tuning, LLM, etc.)
- Contribution of different vision encoders



BRAVE🦁: Broadening the visual encoding of vision-language models



Leverage off-the-shelf vision encoders for strong performance

ECCV 2024 (Oral)

brave-vlms.epfl.ch

Publication map & Thesis outline

Part I: Handling Modalities

Robustness Mechanisms

- Cross-task Consistency (CVPR 2020, Best Paper Nominee)
- Cross-domain Ensembles (ICCV 2021, Oral)
- Rapid Network Adaptation (ICCV 2023)
- OOD Generalization via Diversity (ICLR 2024)

Robust Pseudolabelers & Vision Encoders

- 3D Common Corruptions (CVPR 2022, Oral) [Demo]
- Broadening Vision in VLMs (ECCV 2024, Oral)

Part II: Multimodal Training

Pretraining Framework

- 4M: Multimodal Modeling (NeurIPS 2023, Spotlight)

Architecture

Training the Unified Model

- 4M-21: Any-to-Any Models (NeurIPS 2024) [Demo]

Data

Model

Status check

Part III: Analysis

Internet-based Generalists

- Evaluating MFMs on Vision (arXiv 2025)

Real-world Specialization

- Test-space Multimodal Training (arXiv 2025)

Variable Tokenization

- Resampling Images via FlexTok (arXiv 2025)

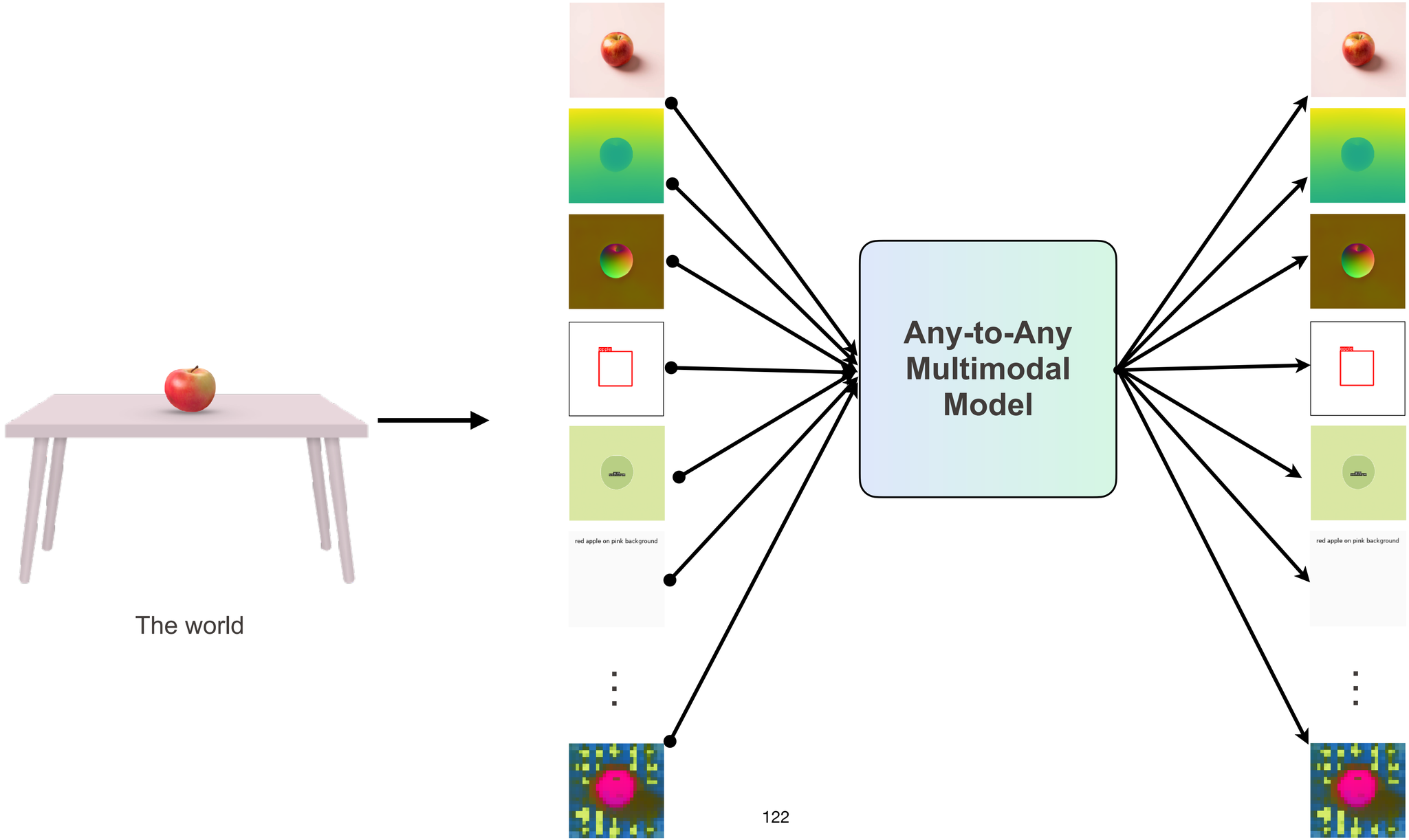
4M-21: An Any-to-Any Vision Model for Tens of Tasks and Modalities

NeurIPS 2024

4m.epfl.ch

Swiss Federal Institute of Technology

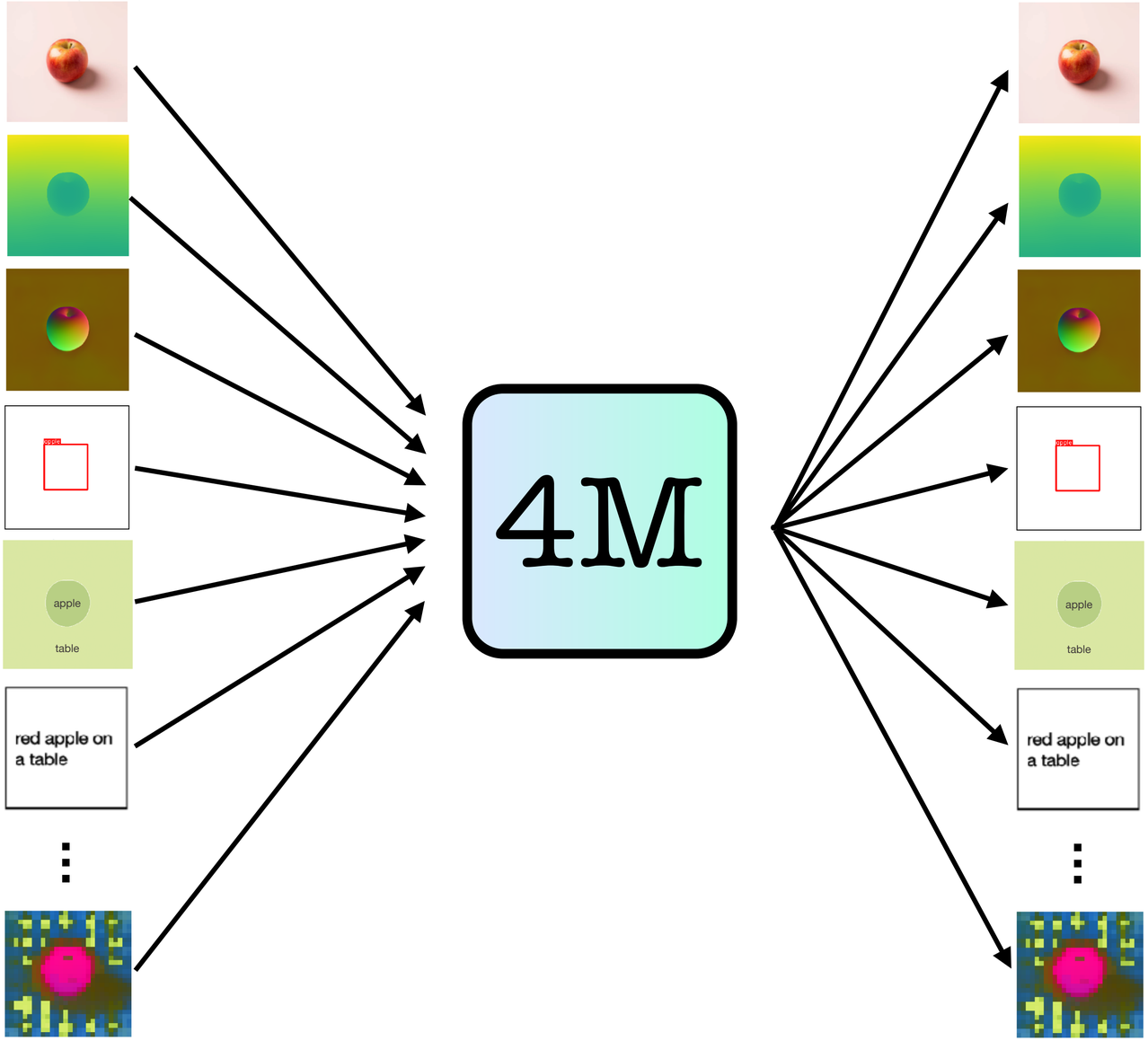
Multimodal Foundation Model (MFM)



4M overview

4M: A training framework for vision foundation models

- Anything in, anything out (**any-to-any**)
- Scale to large model sizes
- Benefit from large datasets

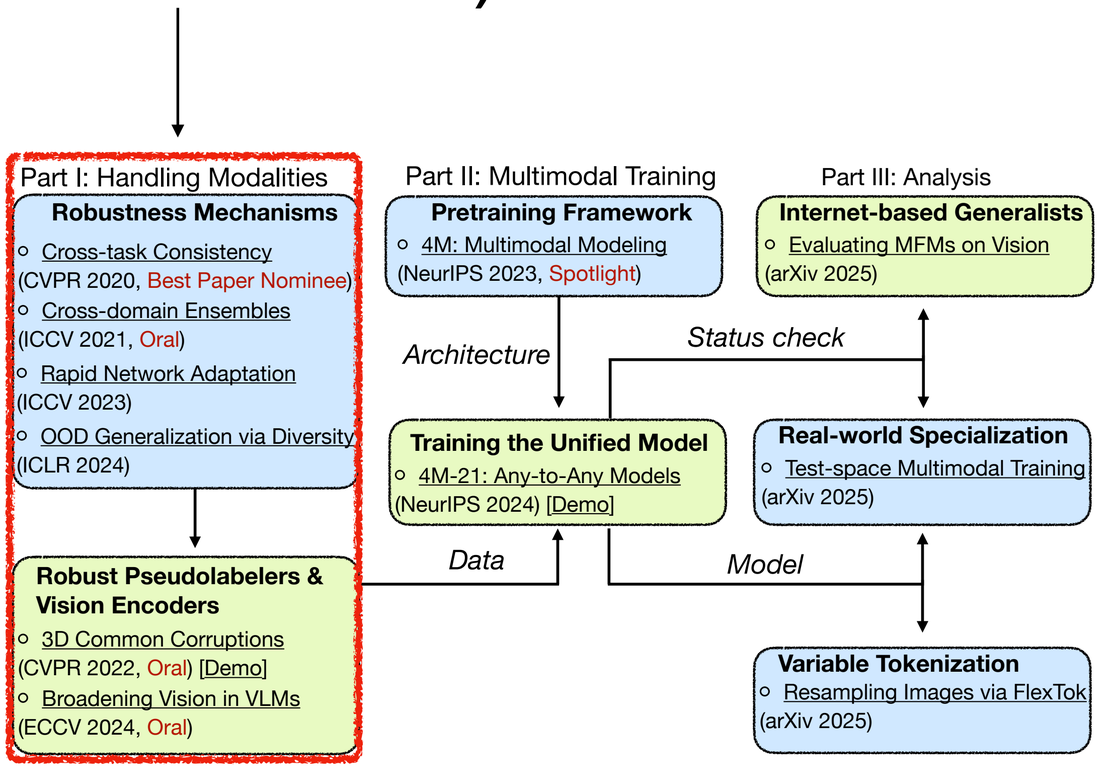


How to scale to several modalities?

- Building an aligned dataset (focus of Part 1)
- Tokenization
- Masked modeling

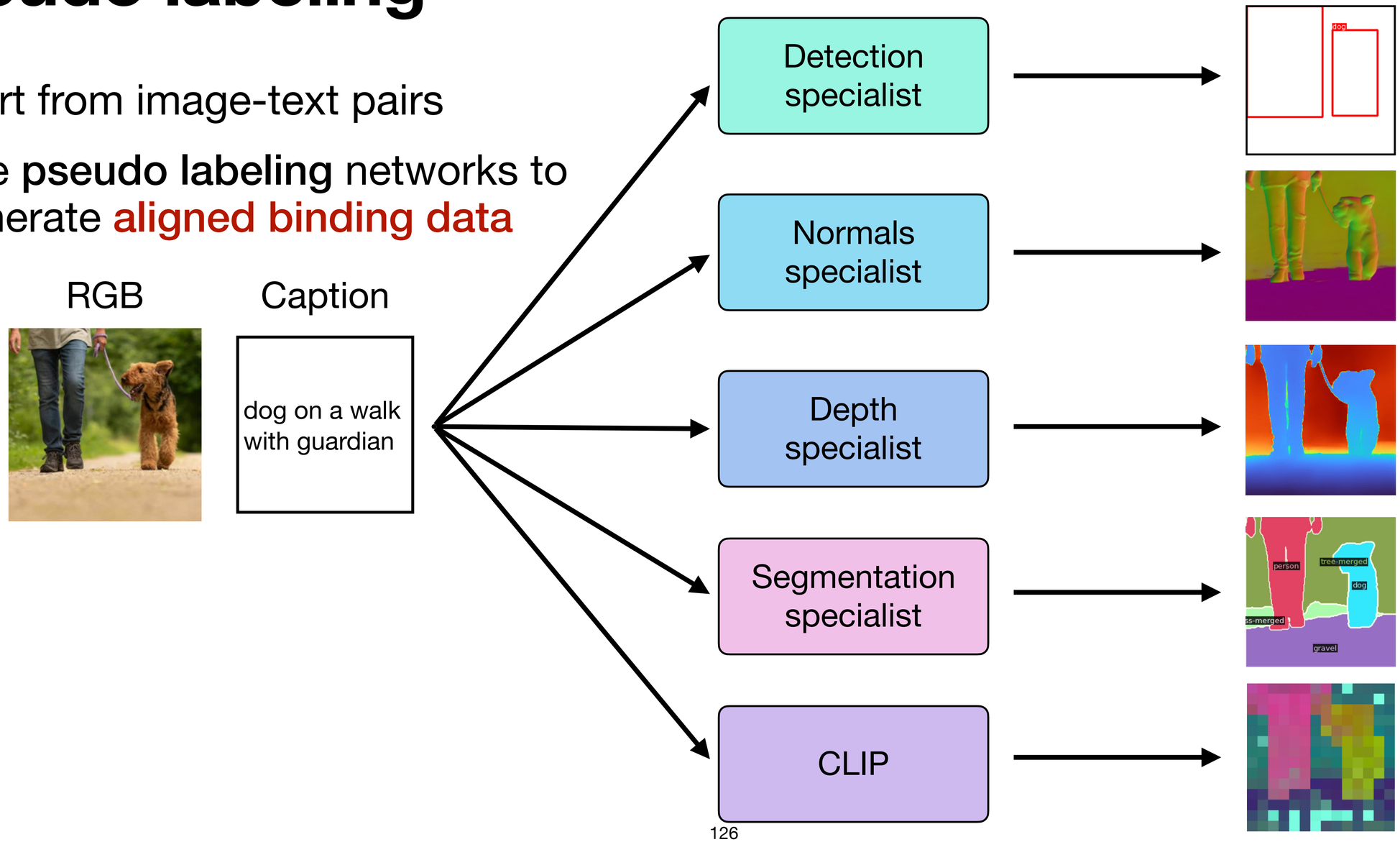
How to scale to several modalities?

- Building an aligned dataset (focus of Part 1)
- Tokenization
- Masked modeling



Building an aligned multimodal dataset through pseudo labeling

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

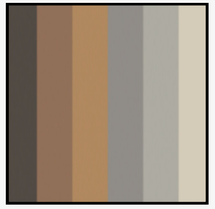


4M-21 modalities

RGB modalities

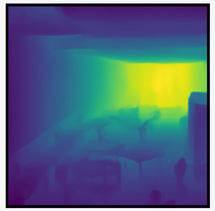
RGB

Color palette

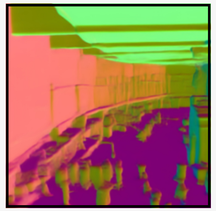


Geometric modalities

Depth



Surface normals

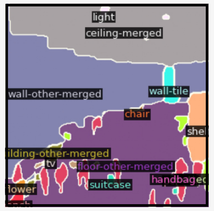


3D human poses

Semantic modalities

Bounding boxes

Semantic segmentation



SAM instances

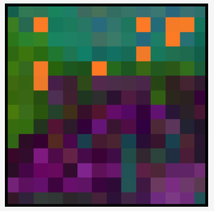
Edge modalities

SAM edges

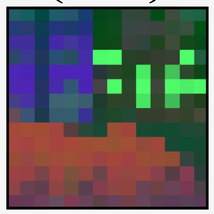
Canny edges

Feature map modalities

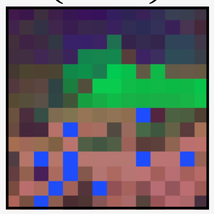
CLIP features (dense)



DINOv2 features (dense)

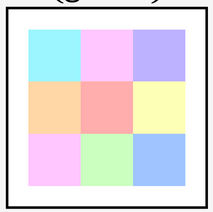


ImageBind features (dense)

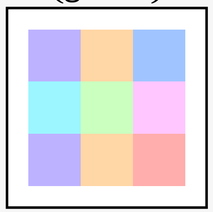


Global feature modalities

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%
...

How to train?

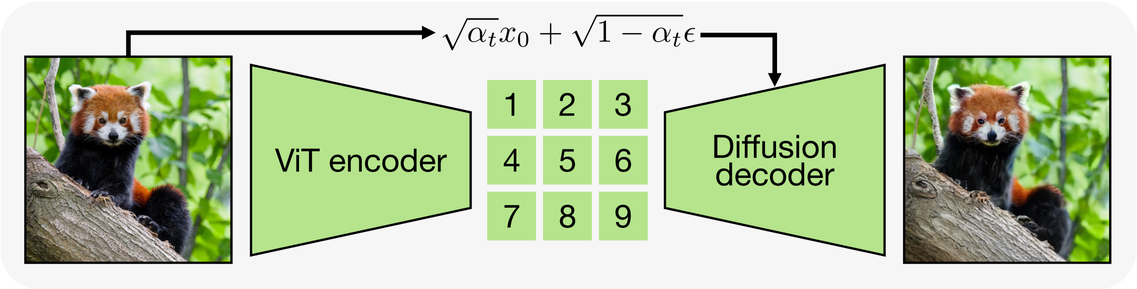
- Building an aligned dataset
- Tokenization
- Masked modeling

Multimodal tokenization

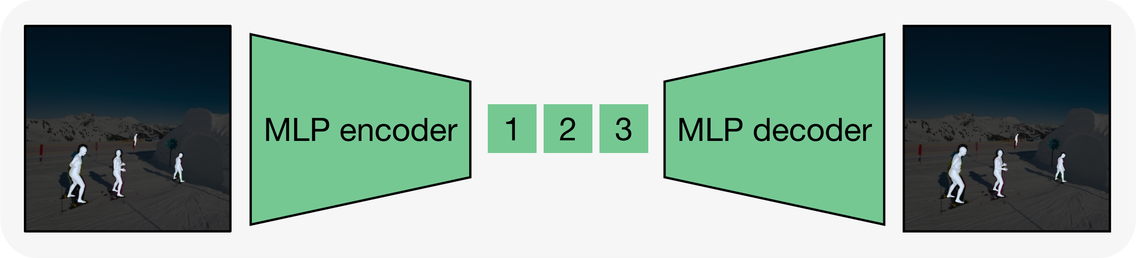
- Discrete sequences for all modalities
- Abstract away modality-specific details
- Scalable training

Multimodal tokenization

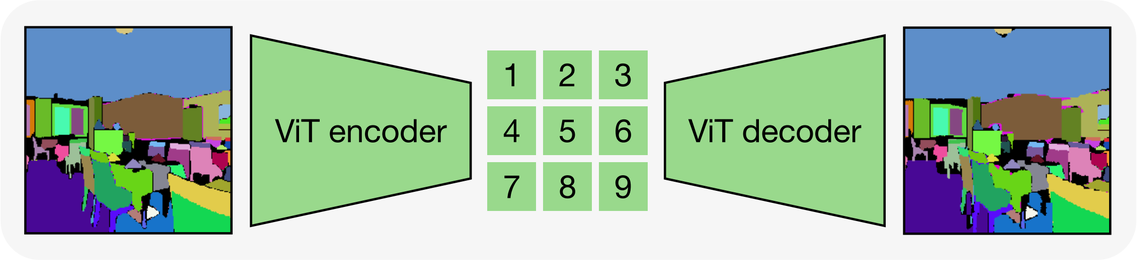
Spatial VQ-VAE with diffusion decoder: RGB, normal, depth, edges



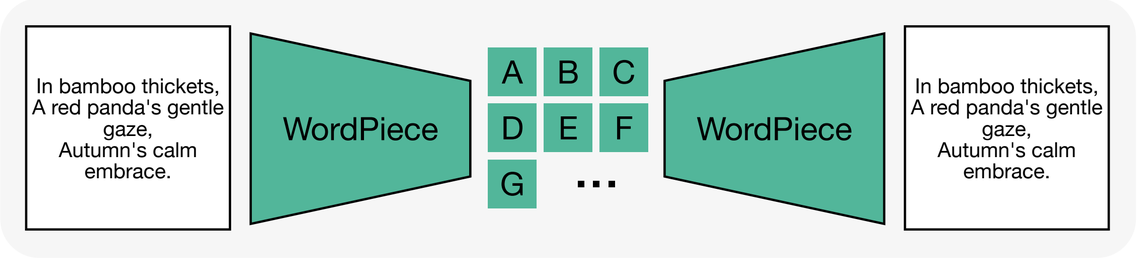
MLP VQ-VAE: Human poses, DINOv2 & ImageBind global tokens



Spatial VQ-VAE: Segmentation, CLIP, DINOv2, ImageBind, SAM inst.



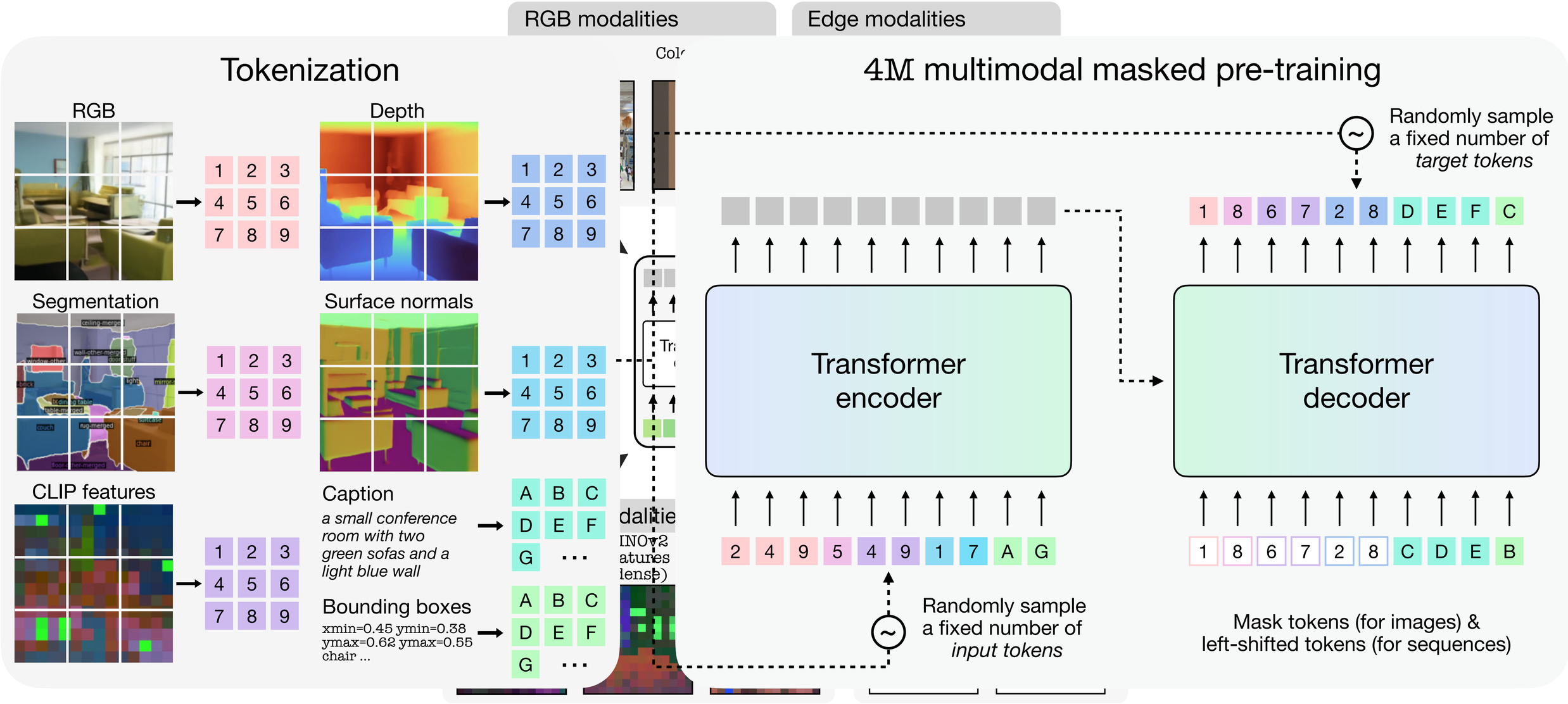
Text tokenizer: Text, bounding boxes, metadata, color palette



How to train?

- Building an aligned dataset
- Tokenization
- Masked modeling

Training 4M-21



Resulting capabilities

Resulting capabilities

- Any-conditional, steerable generation

Resulting capabilities

- Any-conditional, steerable generation
- Cross-modal retrieval

Resulting capabilities

- Any-conditional, steerable generation
- Cross-modal retrieval
- Solving vision tasks out-of-the-box

Resulting capabilities

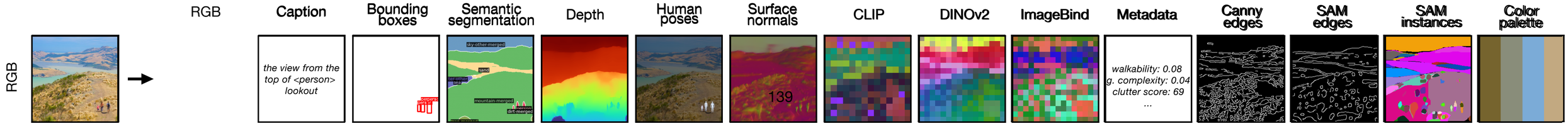
- Any-conditional, steerable generation
- Cross-modal retrieval
- Solving vision tasks out-of-the-box
- Multi-sensory fusion

Generate <u>any modality</u> conditioned on <u>any other</u>

RGB

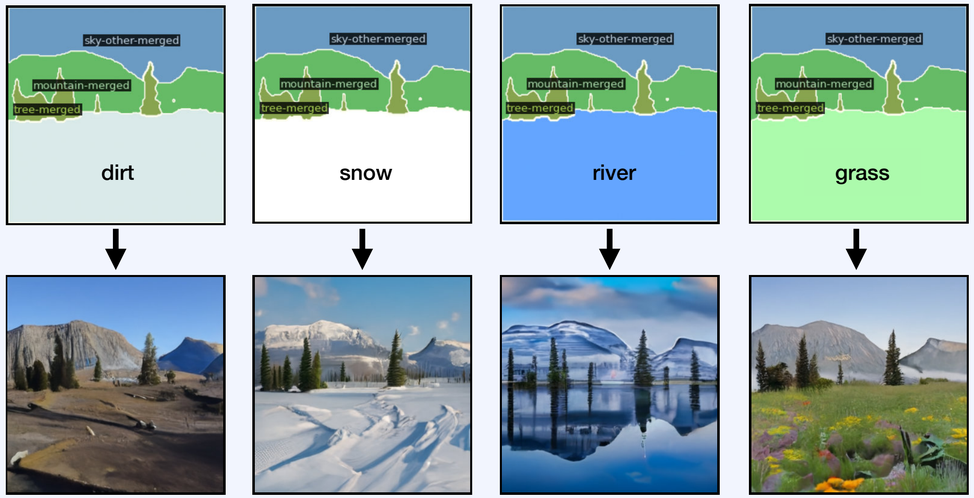
↓

Generate <u>any modality</u> conditioned on <u>any other</u>

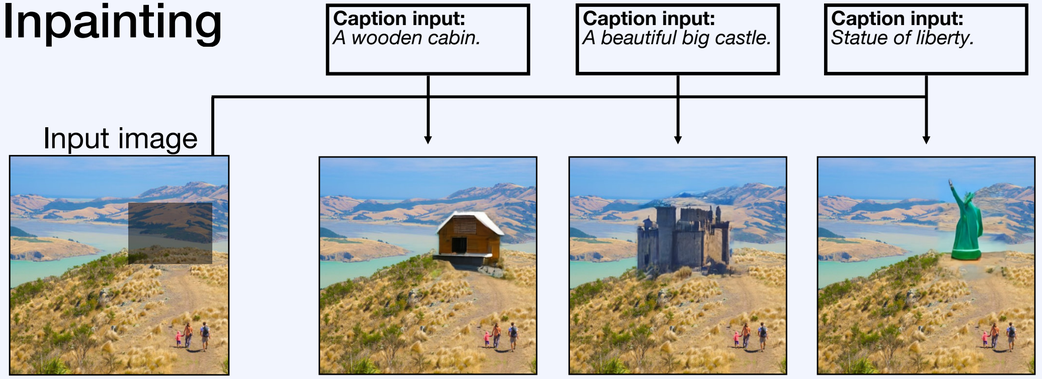


Steerable generation & multimodal image editing

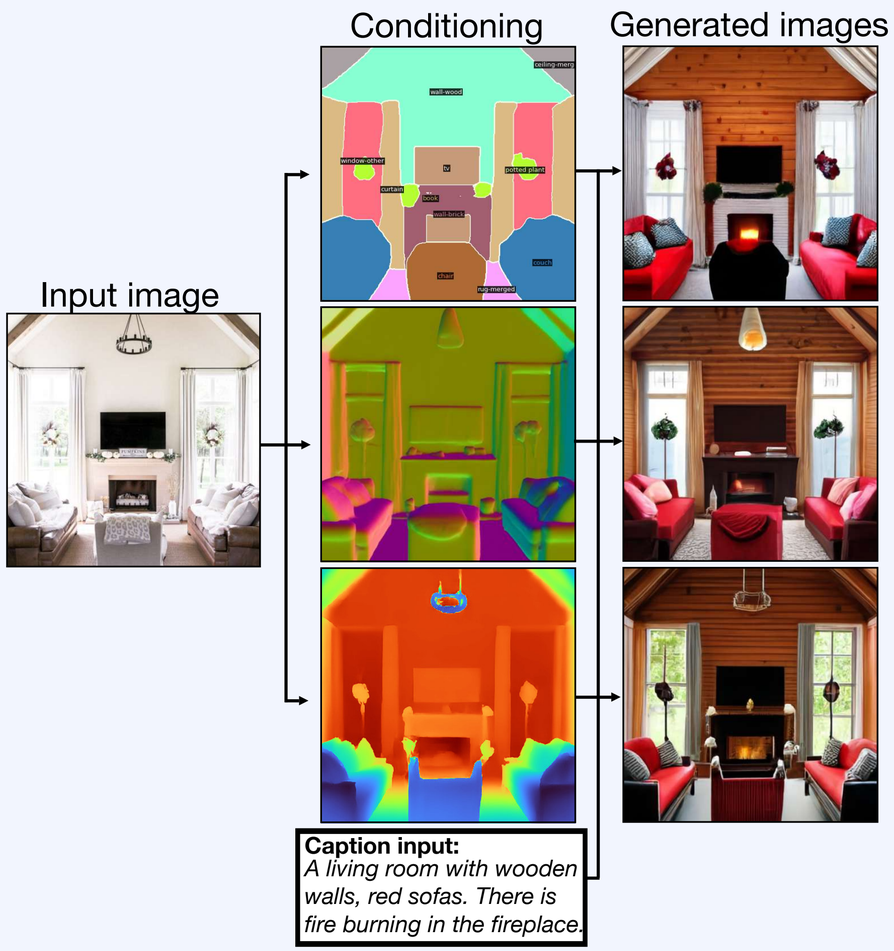
Semantic editing



Inpainting



Grounded image generation



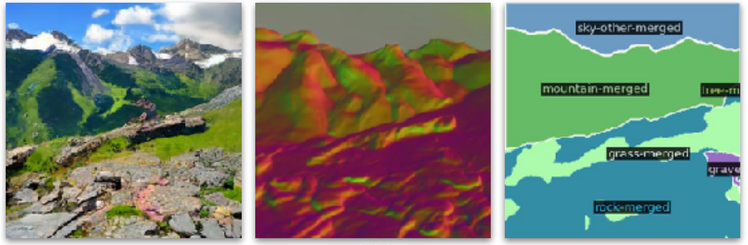
Steerable multimodal data generation with metadata

SAM clutter score: 0.00

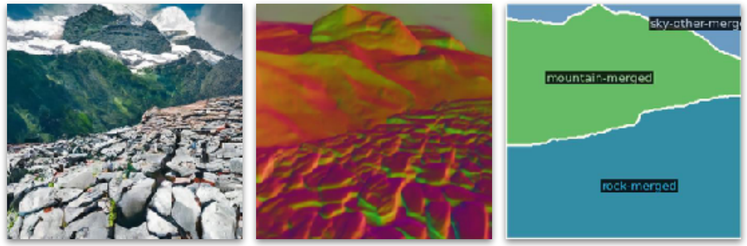
+ Caption input: *a photo of a mountain scene*

Low SAM clutter score

Medium SAM clutter score



High SAM clutter score



Fine-grained control through <u>multimodal guidance & weighting</u>

RGB input

Extracted depth

Caption input

an oil painting of a blue flower

Fixed weight

🔒 2.0

Varying weight

Negative geometric conditioning | Weak geometric conditioning | Stronger geometric conditioning

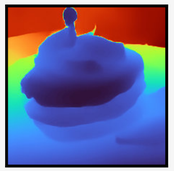
Guidance weight: -1.00

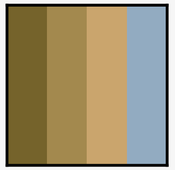
Multimodal retrieval with 4M-21

Any-to-RGB retrieval

Query

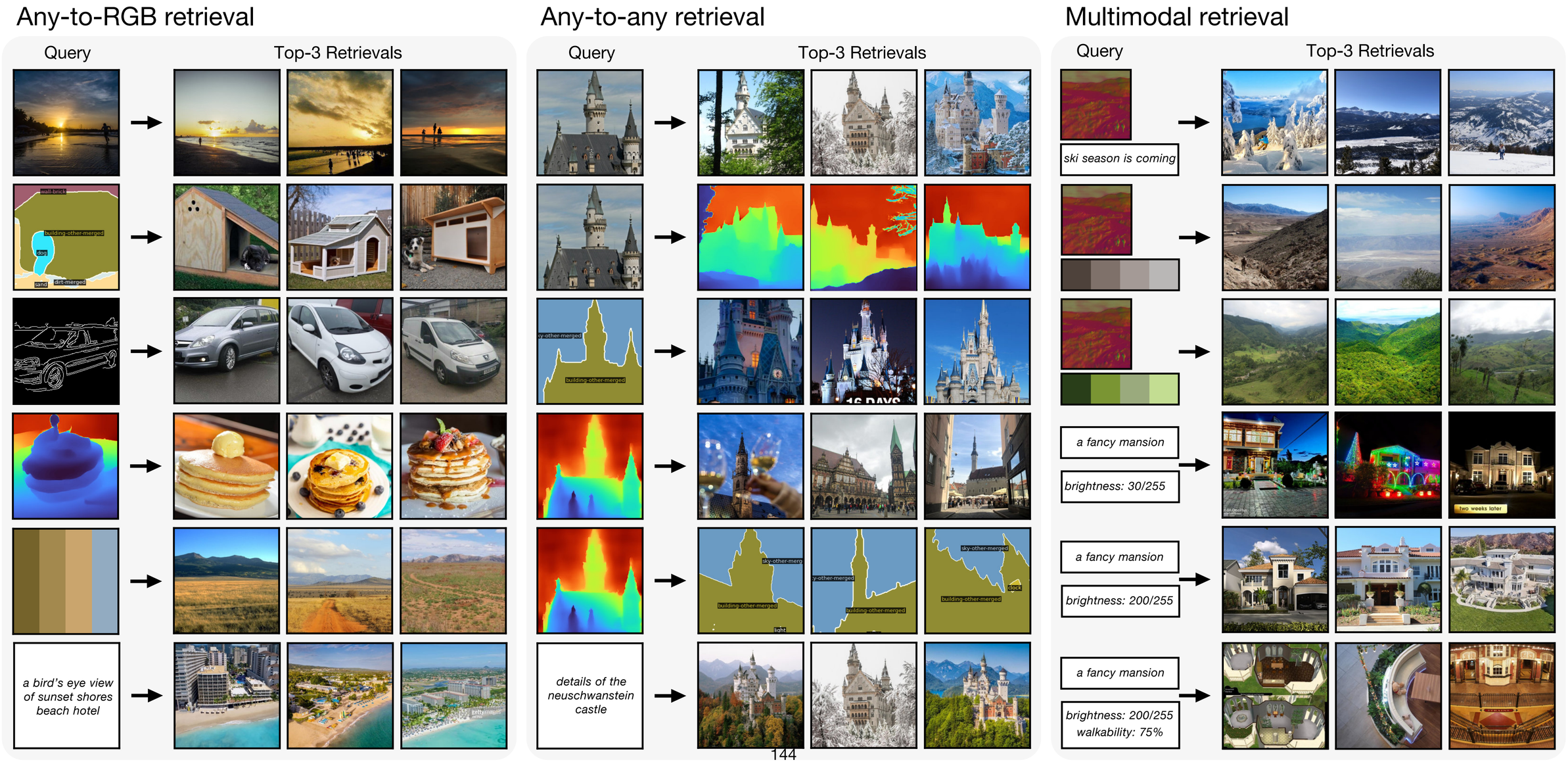
Top-3 Retrievals



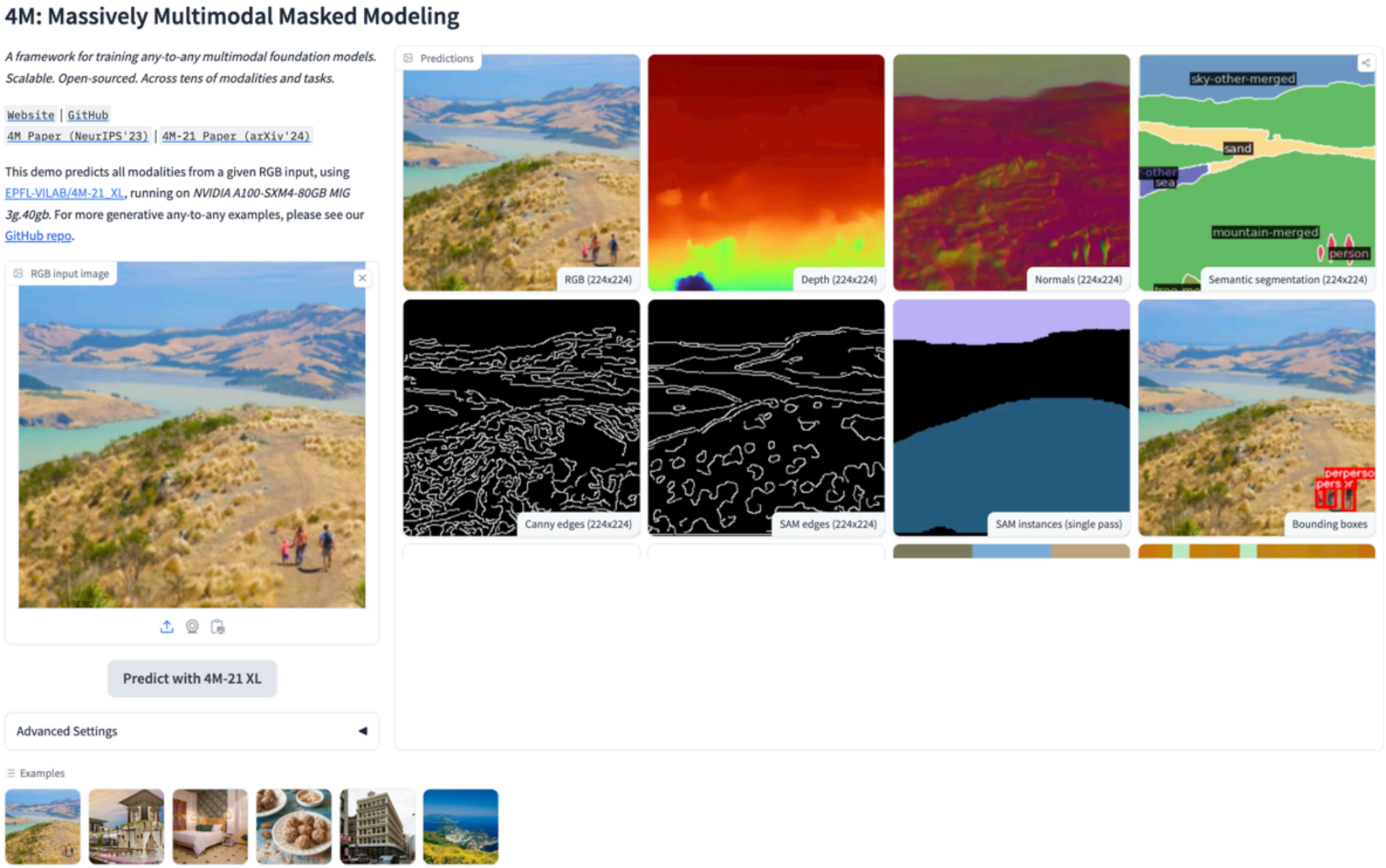


a bird's eye view of sunset shores beach hotel

Multimodal retrieval with 4M-21



Demos: 4m.epfl.ch



Quantitative comparisons

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

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

Multimodal transfer

	NYUv2-S mIoU ↑		Hypersim mIoU ↑		ARKitScenes AP^{3D}↑	
Method	RGB	RGB-D	RGB	RGB-D	RGB	RGB-D
4M 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 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 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**

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

Multimodal transfer

	NYUv2-S mIoU ↑		Hypersim mIoU ↑		ARKitScenes AP^{3D}↑	
Method	RGB	RGB-D	RGB	RGB-D	RGB	RGB-D
4M 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 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 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 [32]	IN1K	86M	84.2	46.1	89.1	30.9
DeiT III B [70]	IN21K		85.4	49.0	87.4	36.1
MultiMAE B [7]	IN1K		84.0	46.2	89.0	34.2
DINOv2 B [53]	LVD142M		85.3	51.6	92.2	38.1
4M B [50]	CC12M		84.5	50.1	92.0	40.3
4M 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 [32]	IN1K	303M	86.8	51.8	93.6	36.2
DeiT III L [70]	IN21K		87.0	52.0	89.6	40.3
DINOv2 L [53]	LVD142M		86.7	53.4	94.1	42.8
4M L [50]	CC12M		86.6	53.4	94.4	46.8
4M 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 [53]	LVD142M	1.1B	**88.0**	**58.7**	92.5	45.3
4M XL [50]	CC12M	1.2B	87.0	55.0	96.1	48.1
4M XL (Ours)	COYO		87.1	56.1	**96.5**	47.3
Ours XL	CC12M+COYO+C4		87.1	56.0	**96.5**	**48.4**

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

- **Can solve several tasks out-of-the-box**

Multimodal transfer

	NYUv2-S mIoU ↑		Hypersim mIoU ↑		ARKitScenes AP^{3D}↑	
Method	RGB	RGB-D	RGB	RGB-D	RGB	RGB-D
4M 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 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 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 [32]	IN1K	86M	84.2	46.1	89.1	30.9
DeiT III B [70]	IN21K		85.4	49.0	87.4	36.1
MultiMAE B [7]	IN1K		84.0	46.2	89.0	34.2
DINOv2 B [53]	LVD142M		85.3	51.6	92.2	38.1
4M B [50]	CC12M		84.5	50.1	92.0	40.3
4M 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 [32]	IN1K	303M	86.8	51.8	93.6	36.2
DeiT III L [70]	IN21K		87.0	52.0	89.6	40.3
DINOv2 L [53]	LVD142M		86.7	53.4	94.1	42.8
4M L [50]	CC12M		86.6	53.4	94.4	46.8
4M 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 [53]	LVD142M	1.1B	**88.0**	**58.7**	92.5	45.3
4M XL [50]	CC12M	1.2B	87.0	55.0	96.1	48.1
4M XL (Ours)	COYO		87.1	56.1	**96.5**	47.3
Ours XL	CC12M+COYO+C4		87.1	56.0	**96.5**	**48.4**

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

- **Can solve several tasks out-of-the-box**
- **Strong transfer (unimodal, multimodal)**

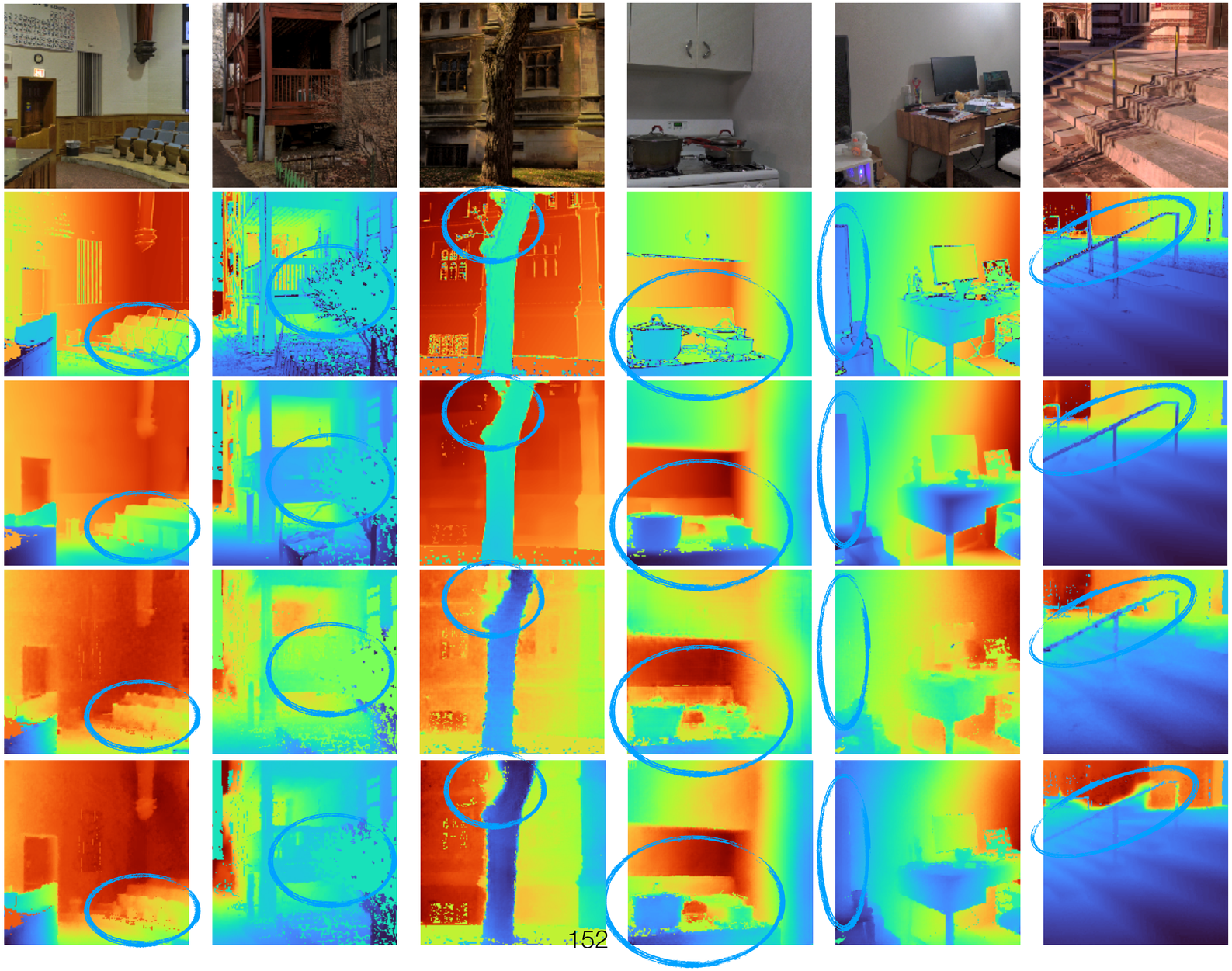
Multimodal transfer

	NYUv2-S mIoU ↑		Hypersim mIoU ↑		ARKitScenes AP^{3D}↑	
Method	RGB	RGB-D	RGB	RGB-D	RGB	RGB-D
4M 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 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 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 [32]	IN1K	86M	84.2	46.1	89.1	30.9
DeiT III B [70]	IN21K		85.4	49.0	87.4	36.1
MultiMAE B [7]	IN1K		84.0	46.2	89.0	34.2
DINOv2 B [53]	LVD142M		85.3	51.6	92.2	38.1
4M B [50]	CC12M		84.5	50.1	92.0	40.3
4M 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 [32]	IN1K	303M	86.8	51.8	93.6	36.2
DeiT III L [70]	IN21K		87.0	52.0	89.6	40.3
DINOv2 L [53]	LVD142M		86.7	53.4	94.1	42.8
4M L [50]	CC12M		86.6	53.4	94.4	46.8
4M 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 [53]	LVD142M	1.1B	**88.0**	**58.7**	92.5	45.3
4M XL [50]	CC12M	1.2B	87.0	55.0	96.1	48.1
4M XL (Ours)	COYO		87.1	56.1	**96.5**	47.3
Ours XL	CC12M+COYO+C4		87.1	56.0	**96.5**	**48.4**

Out-of-the-box comparison to Unified-IO 1 & 2



RGB Input

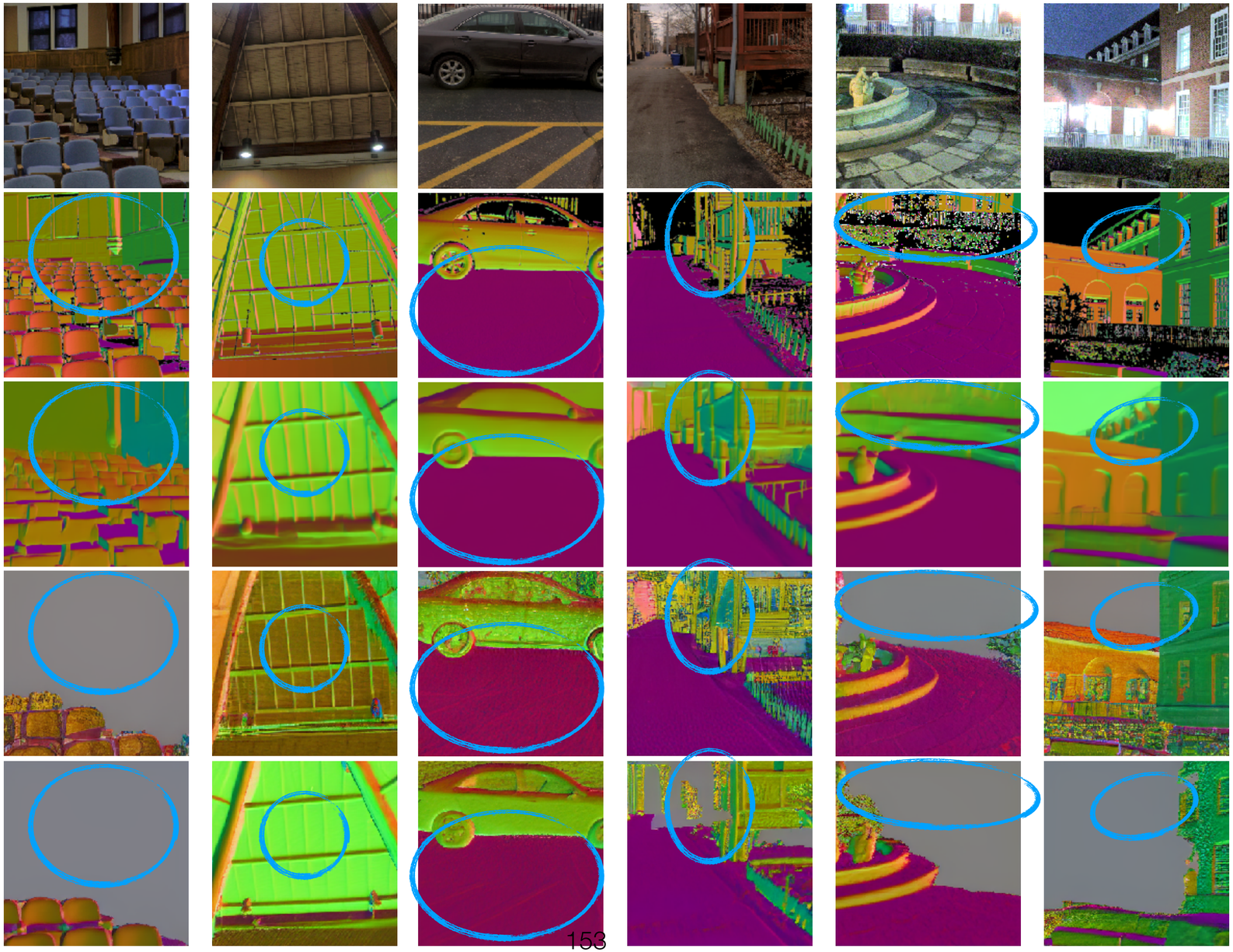
Ground Truth

Ours XL (3B param.) — **0.68** MAE (↓)

Unified-IO XL (3B param.) — 0.82 MAE (↓)

Unified-IO 2 XXL (7B param.) — 0.84 MAE (↓)

Out-of-the-box comparison to Unified-IO 1 & 2



RGB Input

Ground Truth

Ours XL (3B param.) — **20.8** MAE (↓)

Unified-IO XL (3B param.) — 31.0 MAE (↓)

Unified-IO 2 XXL (7B param.) — 37.4 MAE (↓)

Out-of-the-box comparison to Unified-IO 1 & 2

RGB Input

Ground Truth

Ours XL
(3B param.)

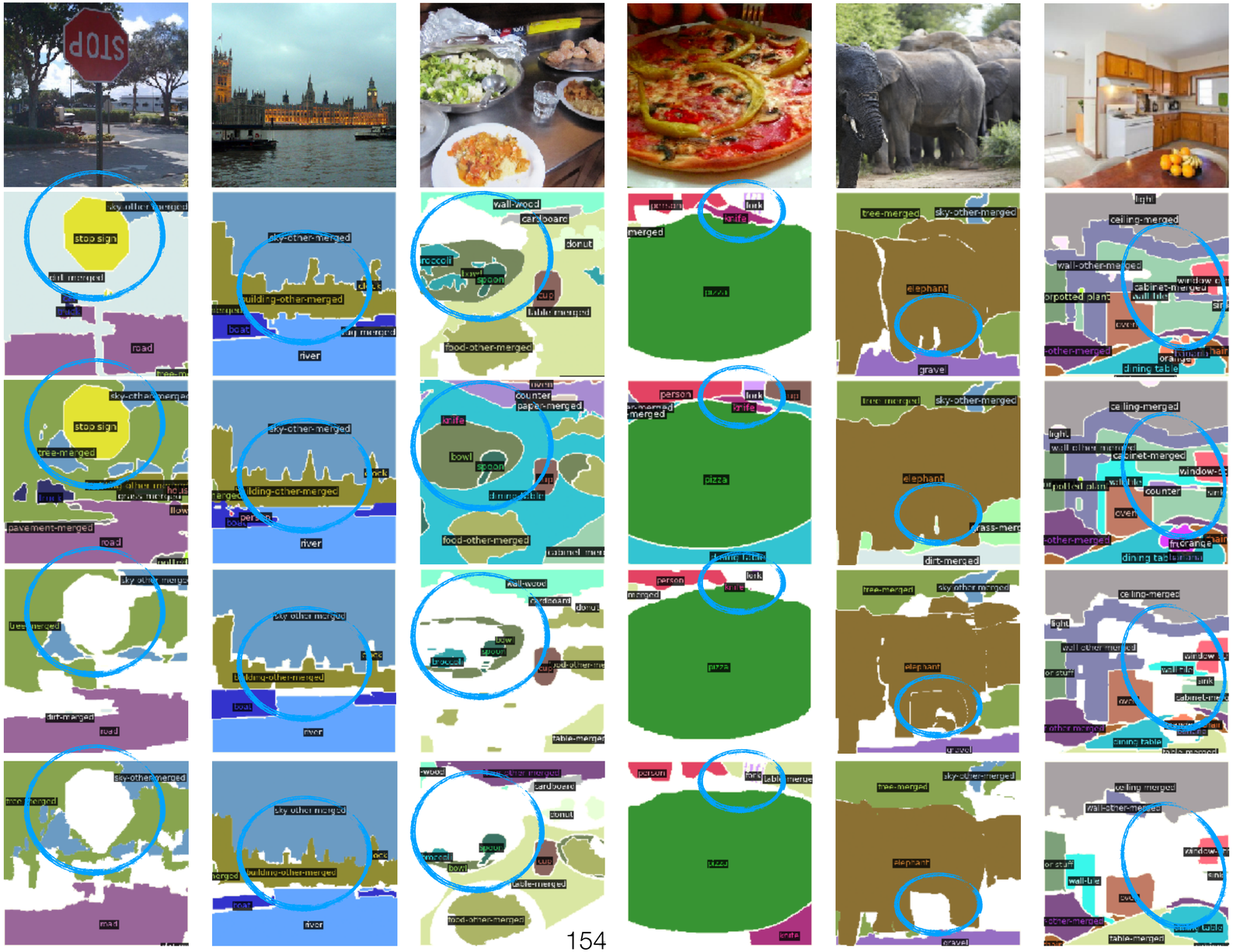
48.1 mIoU (↑)

Unified-IO XL
(3B param.)

44.3 mIoU (↑)

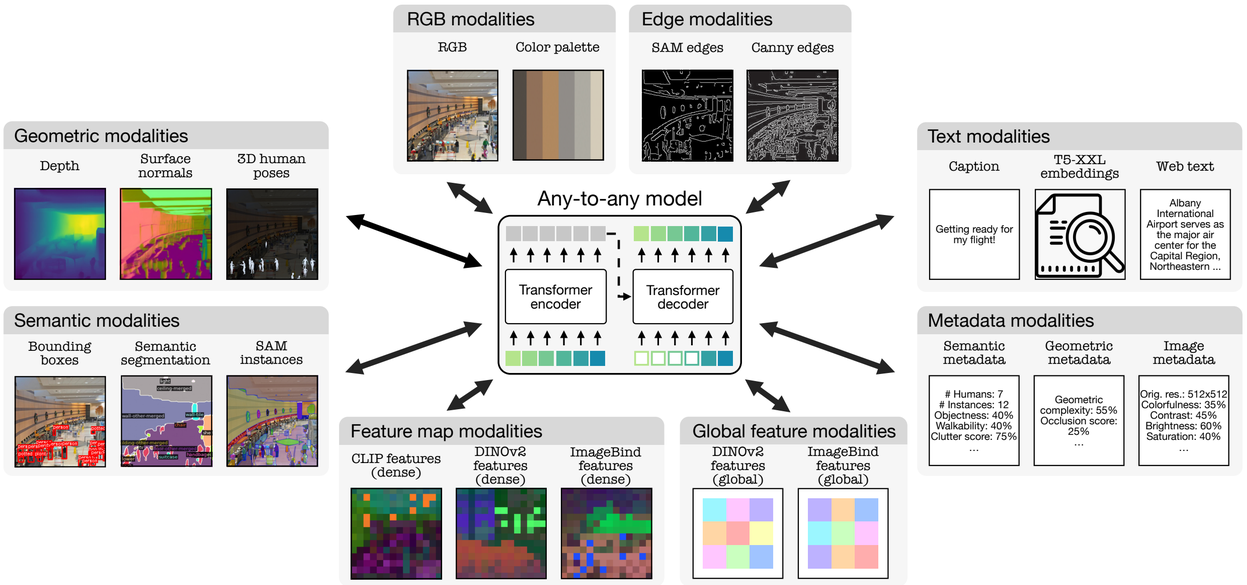
Unified-IO 2 XXL
(7B param.)

41.7 mIoU (↑)



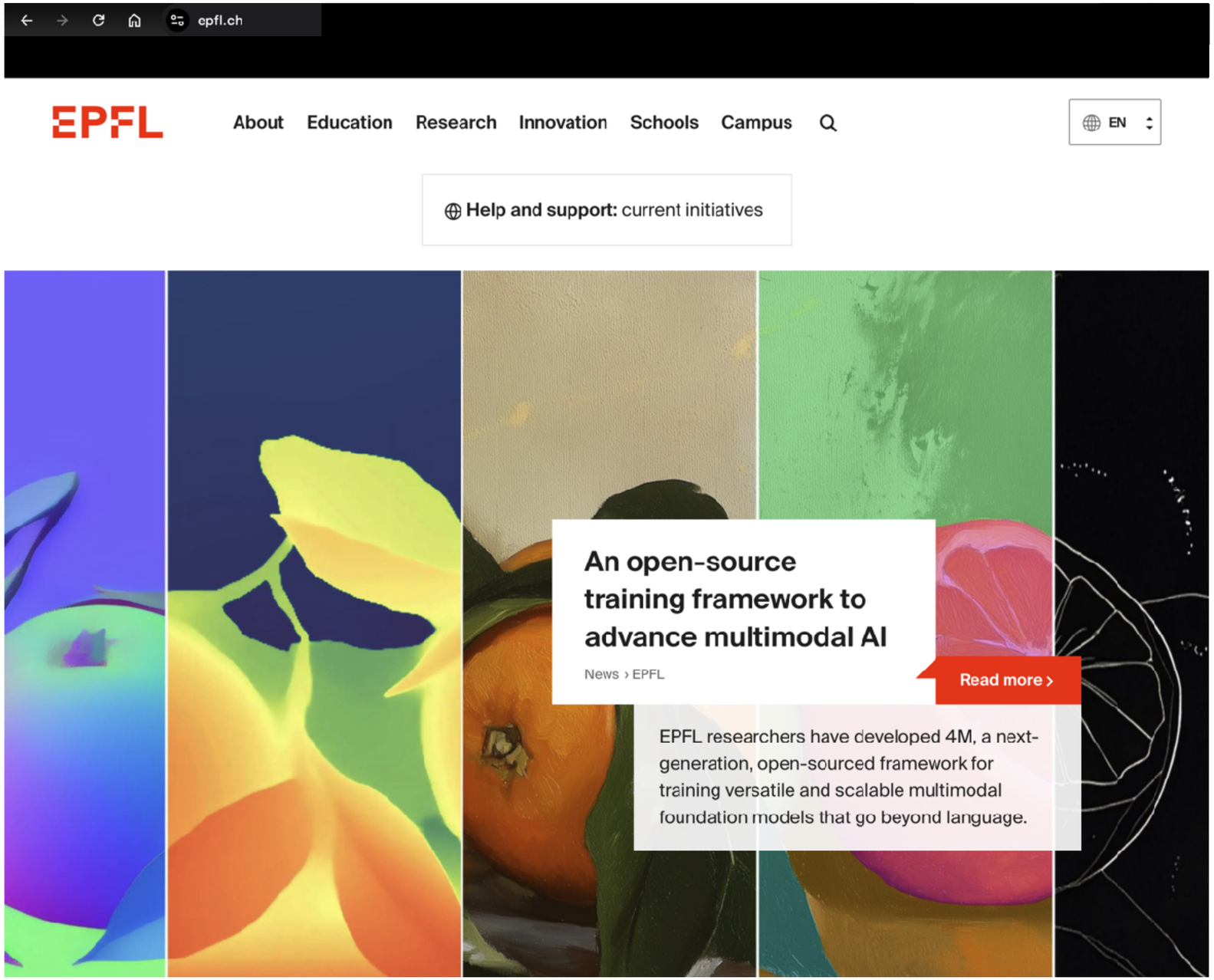
An Any-to-Any Vision Model for Tens of Tasks and Modalities

- 3B param. multimodal & multitask model trained on **21 diverse types of modalities**, across 1T tokens
- Predict **any modality from any other**
- Unlocks exciting **out-of-the-box capabilities**, all **in a single model**:
  - Standard vision tasks, competitive with single-task expert models
  - Any-conditional, steerable generation
  - Cross-modal retrieval
  - Multi-sensory fusion
- **Open sourced** training code and released tokenizers and models (1.7k stars on Github)



4m.epfl.ch

Featured at EPFL front page



Other domains

IBM | Research | Focus areas | Blog | Publications | Careers | About

Home
↳ Publications

Date

09 Dec 2024

Publication

AGU 2024

Authors

Johannes Jakubik
Benedikt Blumenstiel
Stefano Maurogiovanni
Paolo Fraccaro
Rocco Sedona
Gabriele Cavallaro
Thomas Brunschwiler

• *IBM-affiliated at time of publication*

Topics

AI
Accelerated Discovery

Share

Talk

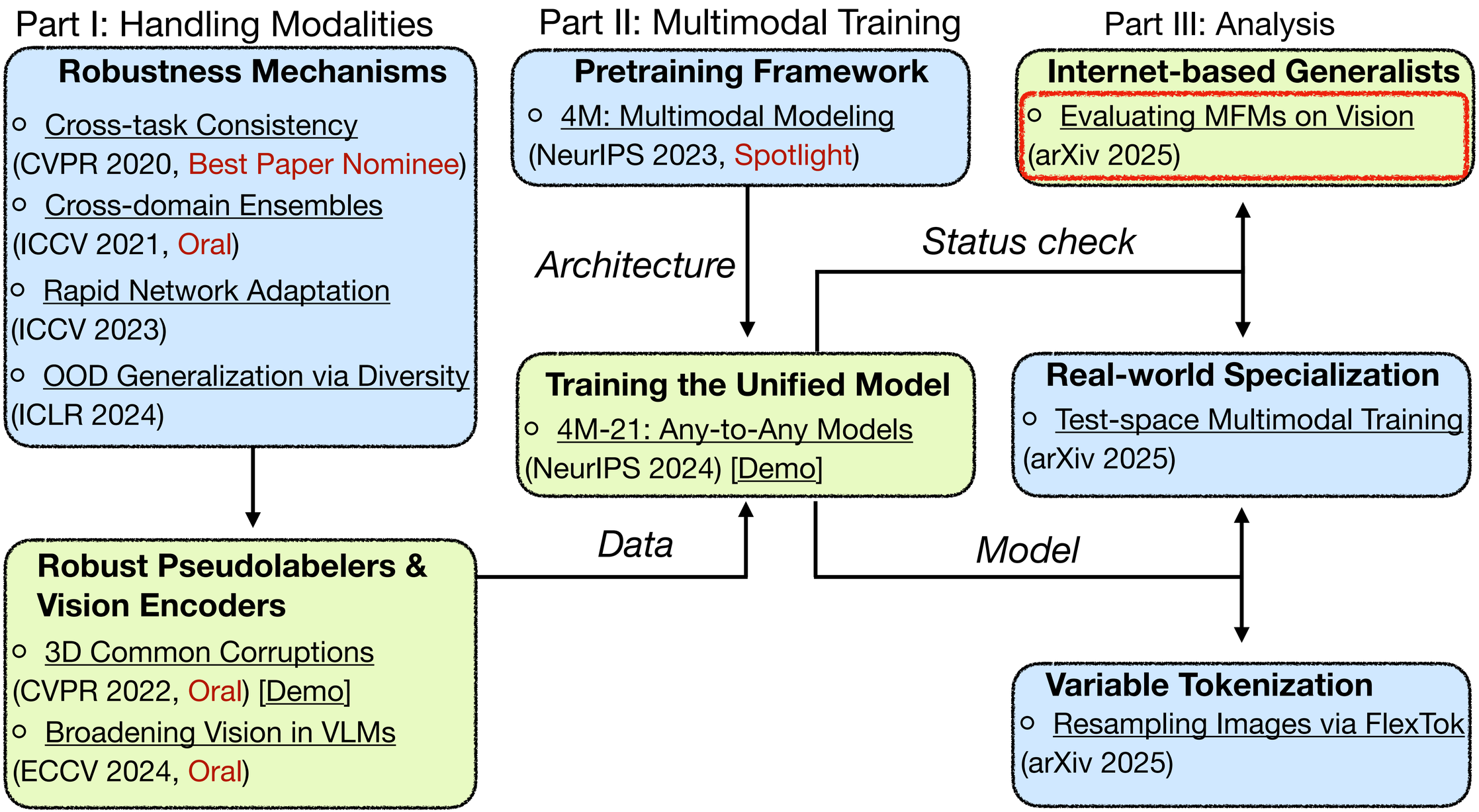
4M4EO – Massively Multi-Modal Masked Autoencoders for Earth Observation

AGU 2024

Abstract

Earth observation generates datasets from diverse modalities such as different satellite missions, a range of higher-level data products, and corresponding meta-data. Deep learning has been struggling with leveraging diverse sets of multi-modal data within a single architecture due to the missing spatio-temporal alignment of the modalities, the partial unavailability of certain modalities, and the vastly different modeling requirements for each of the modalities. To mitigate those challenges, we collect a global-scale dataset of more than 9 million spatio-temporal aligned samples from the Copericus and Landsat missions, higher-level products like Digital Elevation Maps, Land-Use-Land-Change and Canopy-Height datasets, as well as meta-data like textual descriptions and Open Street Map. On top of this data and significant HPC compute, we train a single multi-modal foundation model inspired by models developed for the natural image domain, such as 4M [1]. Our 4M4EO model includes modality-specific tokenizers and a joint masked latent space reconstruction which deals with partial spatio-temporal unavailability. Our approach is two-staged: First, it leverages a single, task-specific encoder-decoder architecture per modality. Second, the resulting embeddings are combined via masked latent space reconstruction. Thus, the approach first learns to embed the data in individual latent spaces and then learns to align the latent spaces through masked latent space reconstruction. We delve in diverse experiments, such as the design of the task-specific encoder-decoder in stage 1 comparing vector-quantization with finite-scalar-quantization as well as diffusion-based with GAN-based approaches. For the second stage, we provide a detailed comparison of different late fusion approaches comparing MLP-based late fusion with masked latent space reconstruction. We believe that our work is pivotal for achieving massive multi-modality for Earth Observation.

Publication map & Thesis outline



Evaluating Multimodal Foundation Models on Standard Computer Vision Tasks

arXiv 2025

fm-vision-evals.epfl.ch

Swiss Federal Institute of Technology

(Leading) MFMs are exciting!

- Impressive range of capabilities

Write a haiku for this image

Dancing robots whirl,
Colors burst, music unfolds,
Joy knows no bounds here.

But How Do They Perform on Vision Tasks?

What city is this? If I want to tour around the landmark in this picture, where should I go?

New York City

This is New York City, and the prominent landmark in the image is the Washington Square Arch, which is located in the Greenwich Village neighborhood. To tour around this landmark, you should visit the nearby Washington Square Park, which is a popular spot for locals and tourists alike. The park offers a variety of activities, including walking paths, a fountain, and a statue of George Washington. Additionally, the surrounding area is rich with historical buildings, art galleries, and restaurants, making it a vibrant and culturally diverse place to explore.

LMSYS Chatbot Arena Leaderboard

Vote!

Blog | GitHub | Paper | Dataset | Twitter | Discord | Kaggle Competition

This is a mirror of the live leaderboard created and maintained by the LMSYS Organization. Please link to leaderboard.lmsys.org for citation purposes.

LMSYS Chatbot Arena is a crowdsourced open platform for LLM evals. We've collected over 1,000,000 human pairwise comparisons to rank LLMs with the Bradley-Terry model and display the model ratings in Elo-scale. You can find more details in our paper. Chatbot arena is dependent on community participation, please contribute by casting your vote!

Arena | NEW: Arena (Vision) | Full Leaderboard

Total #models: 11. Total #votes: 29,500. Last updated: 2024-07-06.

Code to recreate leaderboard tables and plots in this notebook. You can contribute your vote at chat.lmsys.org!

Category: Overall

Overall Questions

#models: 11 (100%) #votes: 29,500 (100%)

Rank* (UB)	Model	Arena Score	95% CI	Votes	Organization	License	Knowledge Cutoff
1	GPT-4o-2024-05-13	1225	+7/-6	6002	OpenAI	Proprietary	2023/10
2	Claude 3.5 Sonnet	1210	+7/-4	8536	Anthropic	Proprietary	2024/4
3	GPT-4-Turbo-2024-04-09	1168	+5/-7	5538	OpenAI	Proprietary	2023/12
3	Gemini-1.5-Pro-API-0514	1162	+6/-6	5956	Google	Proprietary	2023/11
5	Gemini-1.5-Flash-API-0514	1083	+6/-6	5952	Google	Proprietary	2023/11
5	Claude 3 Opus	1082	+7/-7	6063	Anthropic	Proprietary	2023/8
7	Claude 3 Sonnet	1050	+6/-7	6112	Anthropic	Proprietary	2023/8

- How good are these models at standard vision benchmarks (like COCO)?
- How do these models stack up against specialized vision models?

Challenges

Provide a segmentation mask for this image

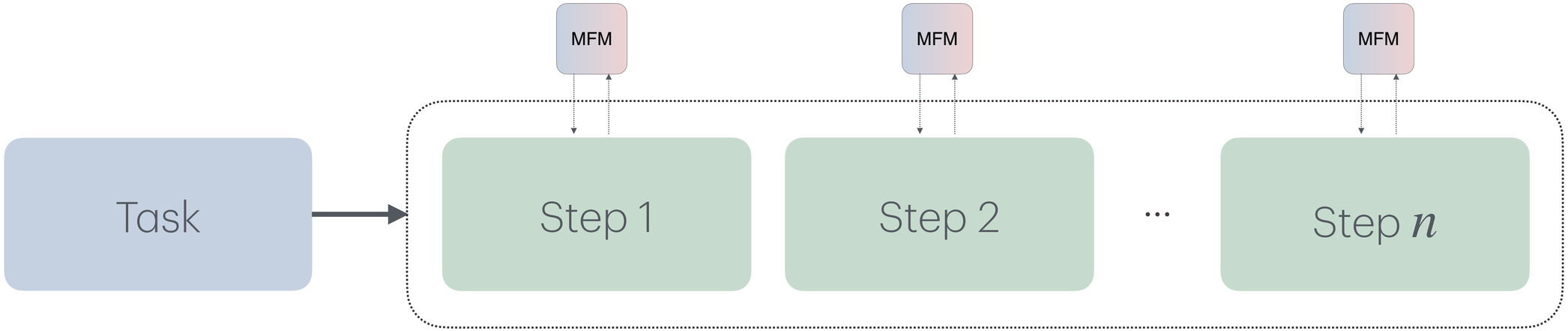
...

Provide bounding box coordinates for the person

[0.3, 0.2, 0.5, 0.5]

1. Cannot provide dense outputs
2. Subpar when directly prompted

Solution



Difficult to solve directly

More manageable

Solution

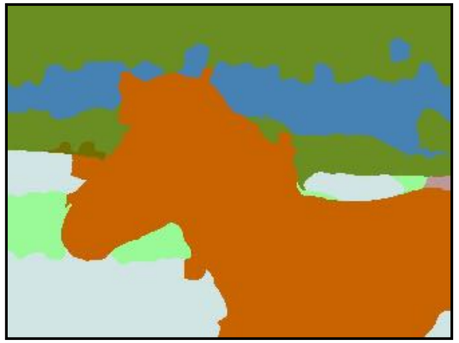
Input Image

Classification

Zebra

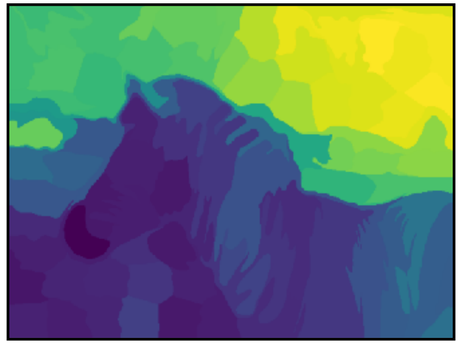
Object Detection

Segmentation

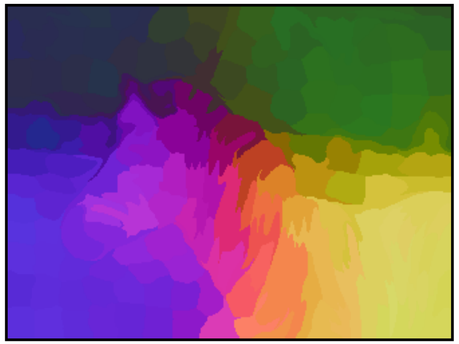


Grouping

Depth

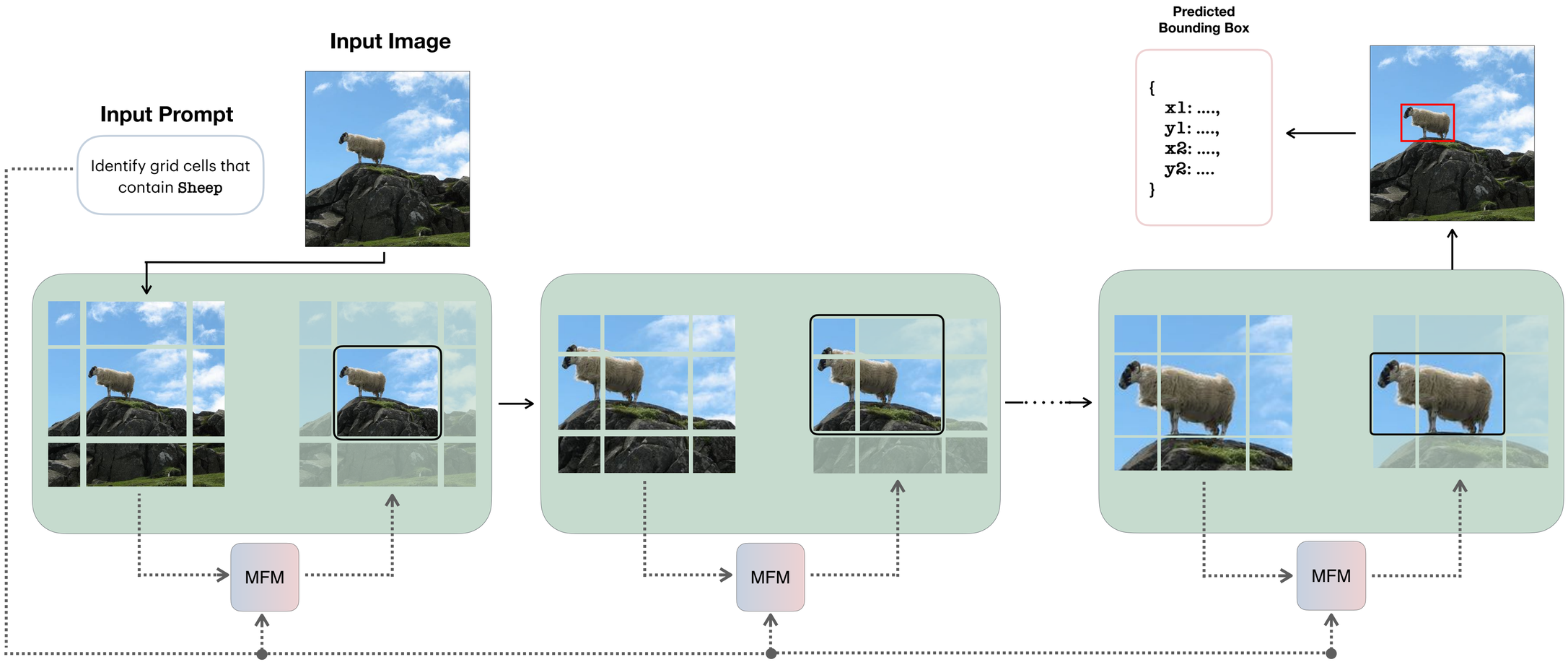


Surface Normals

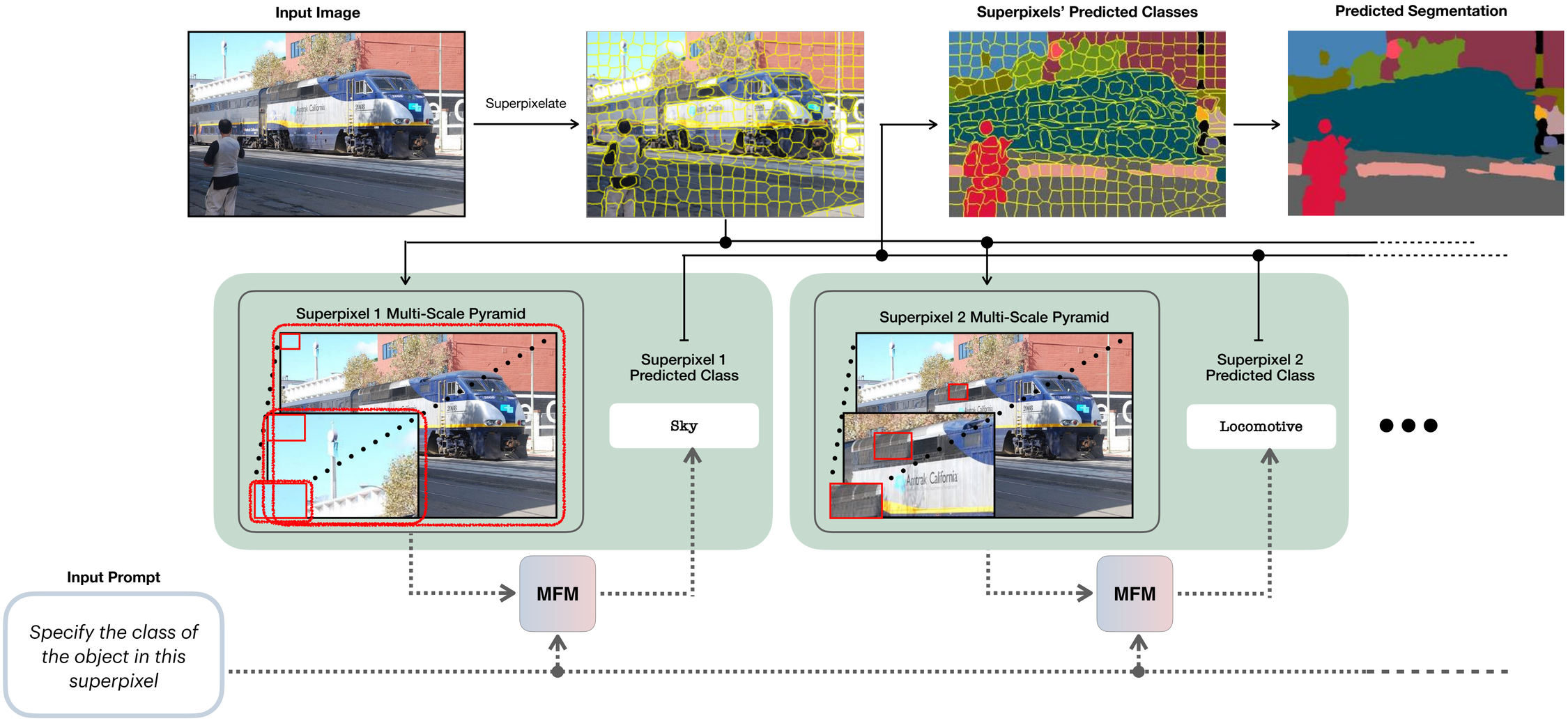


These are GPT-4o predictions!

Object Detection

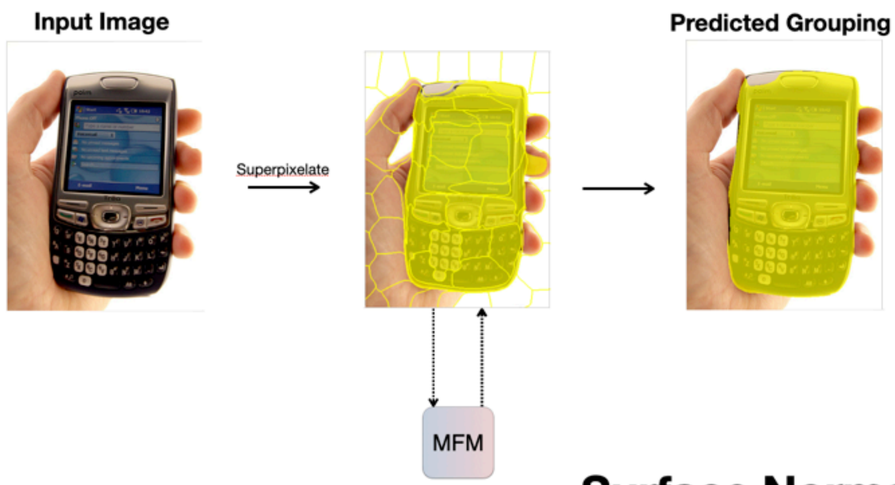


Semantic Segmentation

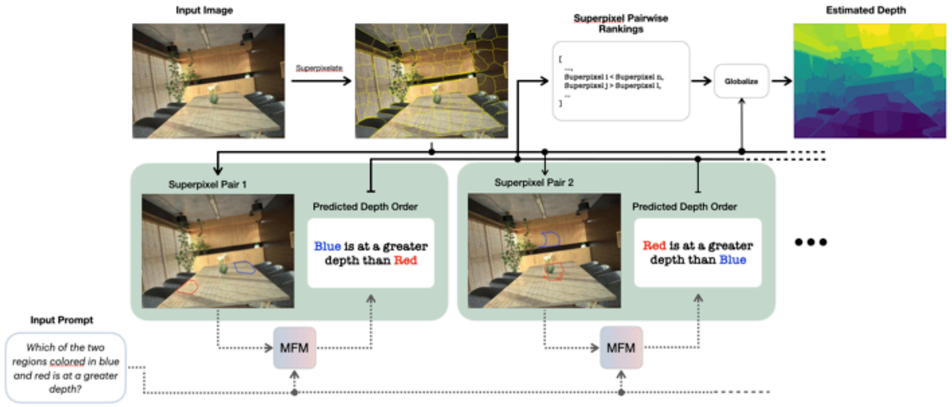


Similar idea, more tasks

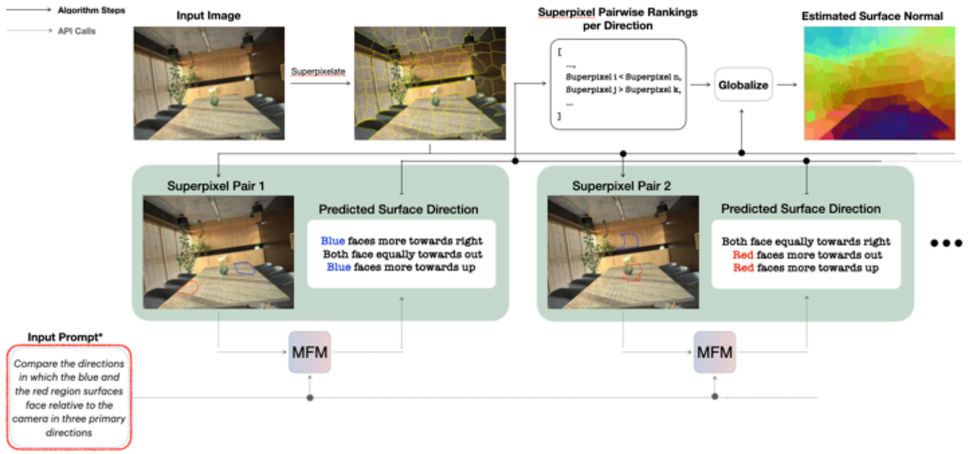
Grouping



Depth Prediction



Surface Normal Prediction



Experimental Results: Overview

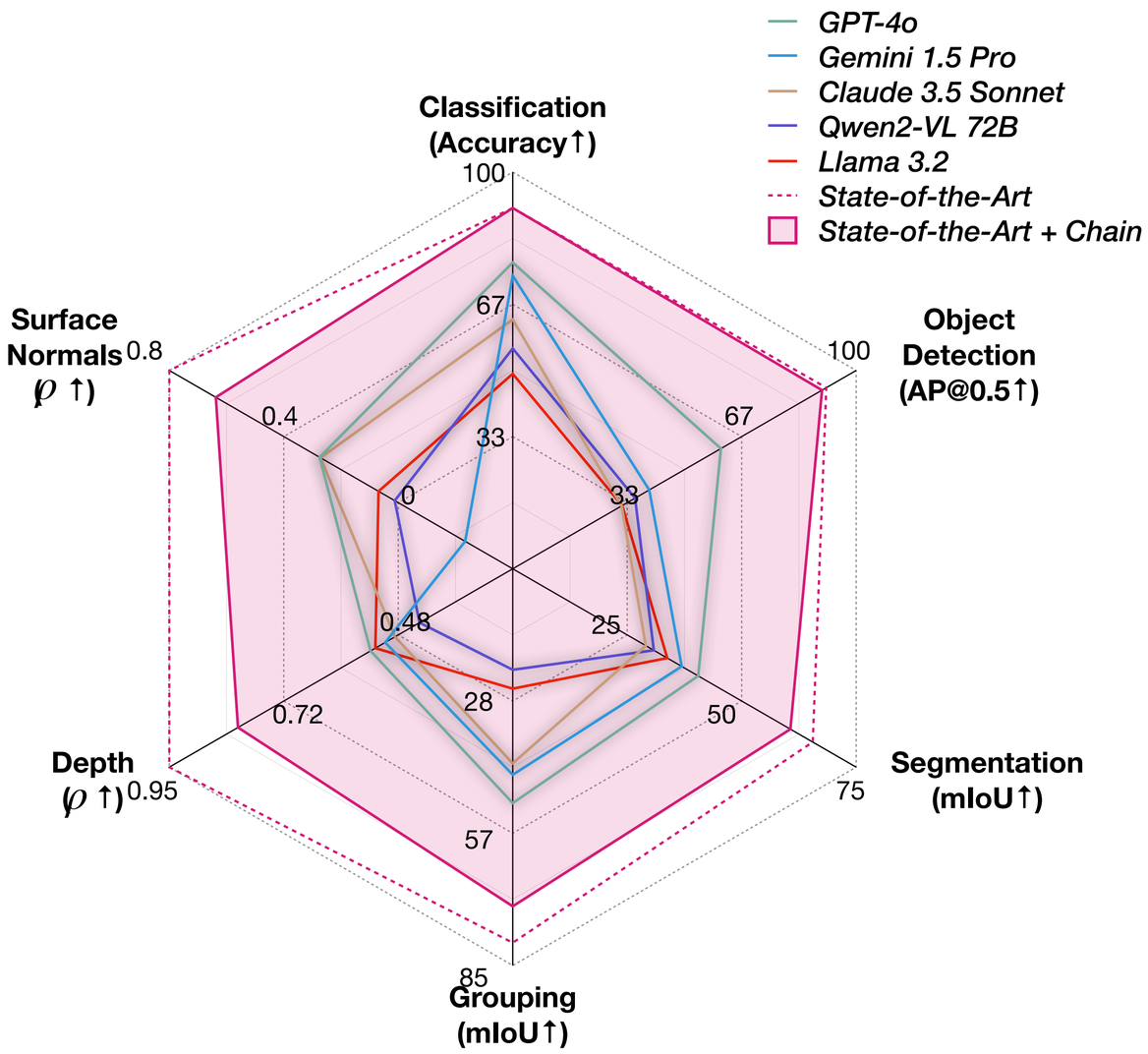
Tasks

- Classification, Object Detection, Semantic Segmentation, Grouping, Depth Prediction, Surface Normal Prediction

Datasets

- Classification: **ImageNet** (and variants)
- Object Detection, Segmentation, Grouping: **COCO**
- Depth and surface normal prediction: **Hypersim**

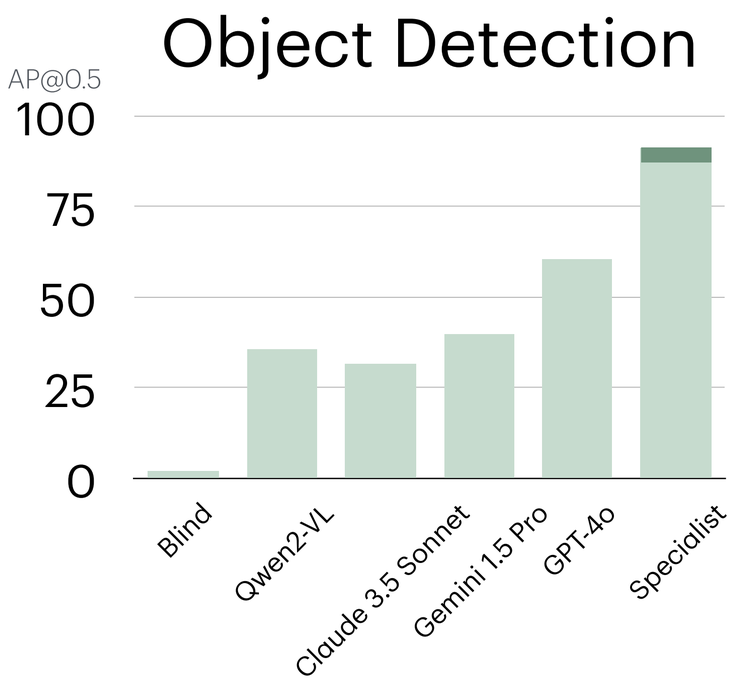
Experimental Results: Takeaways

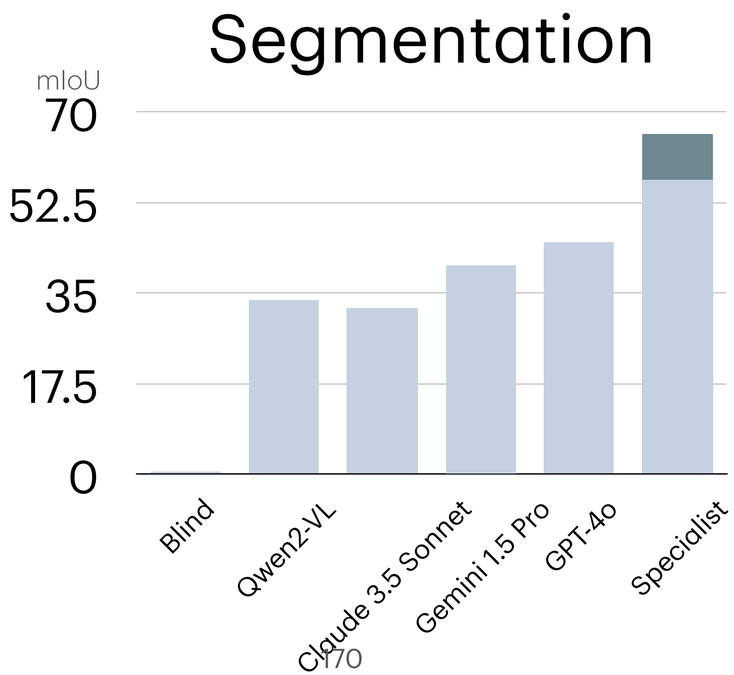


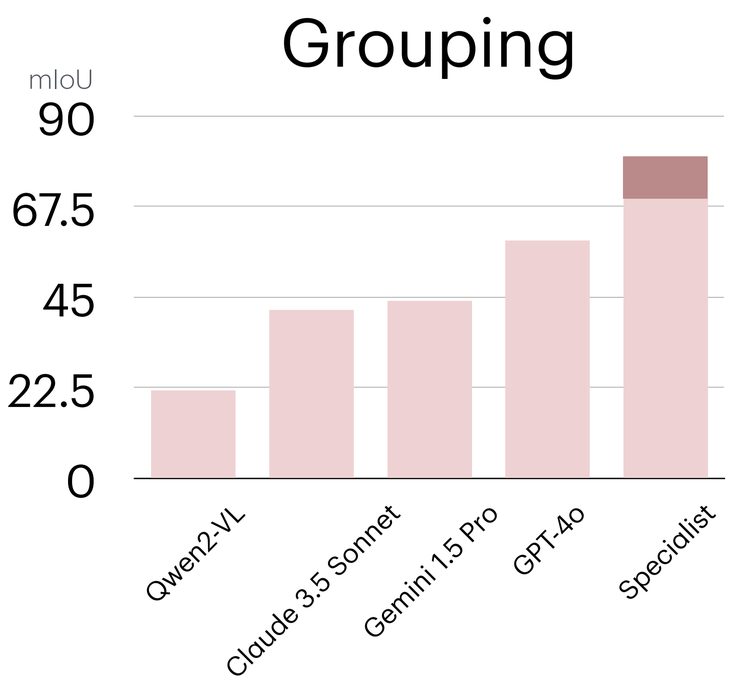
Key Takeaways

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

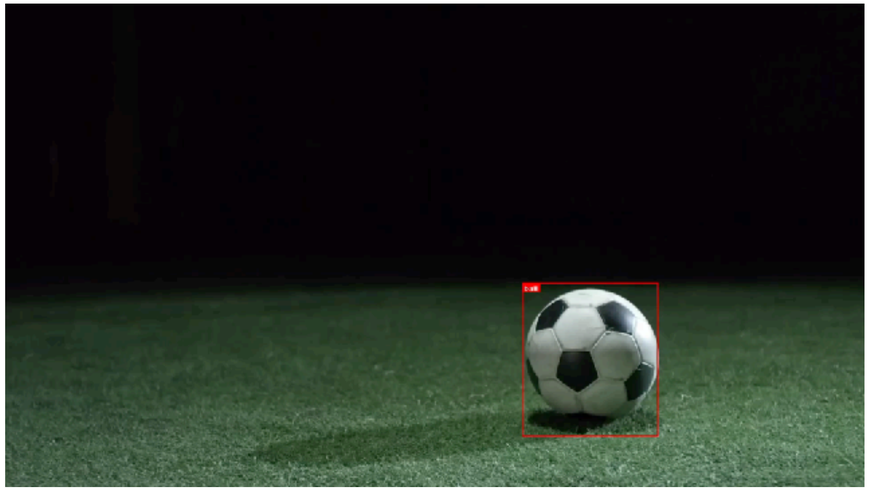
Experimental Results: Semantic Tasks

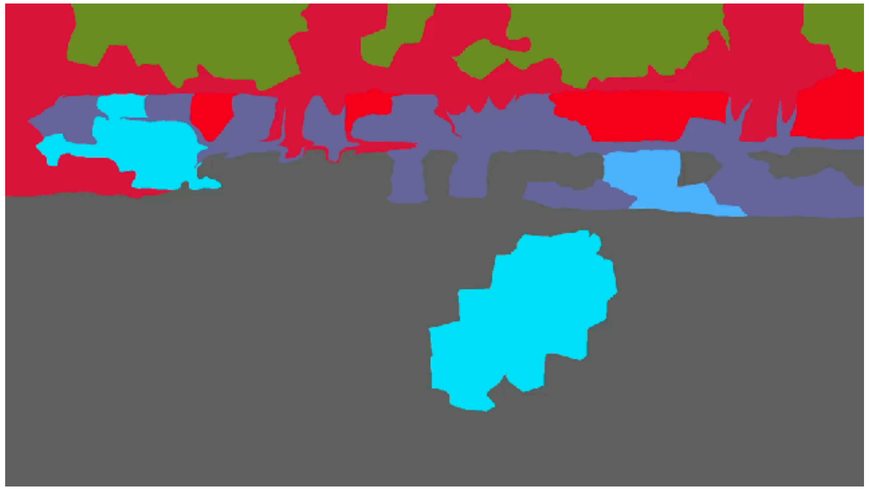




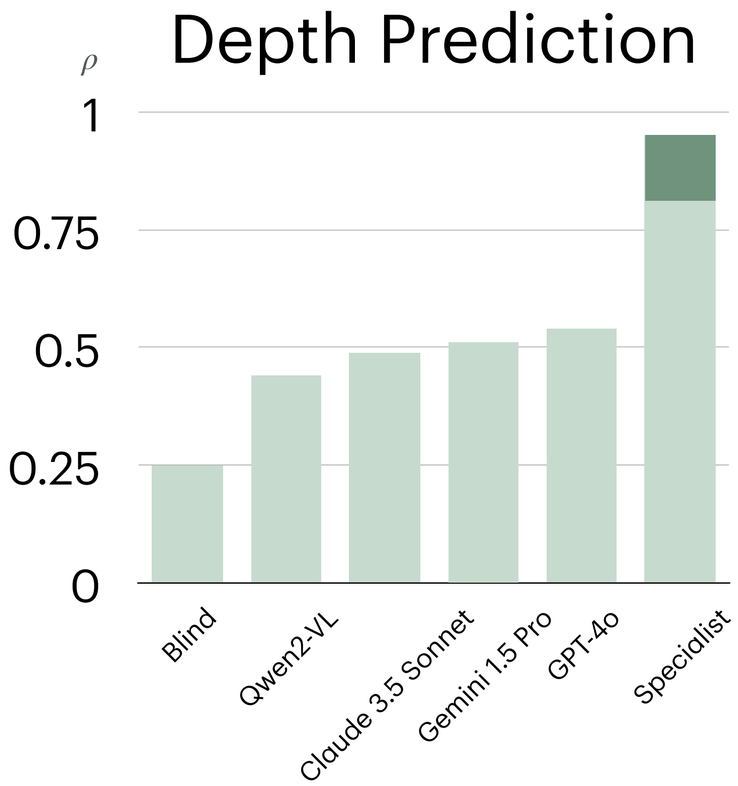


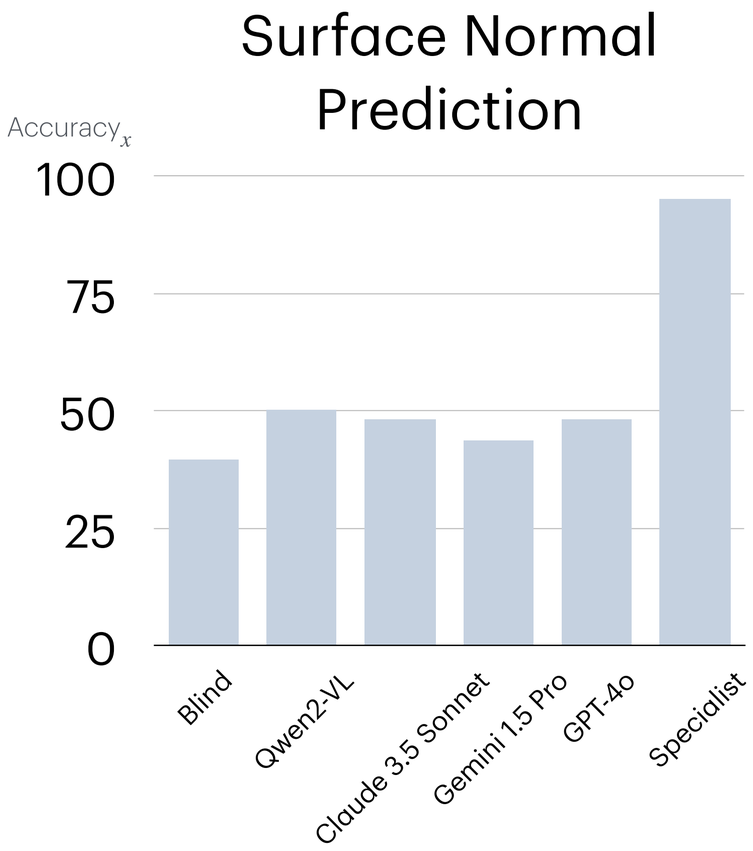
Experimental Results: Semantic Tasks



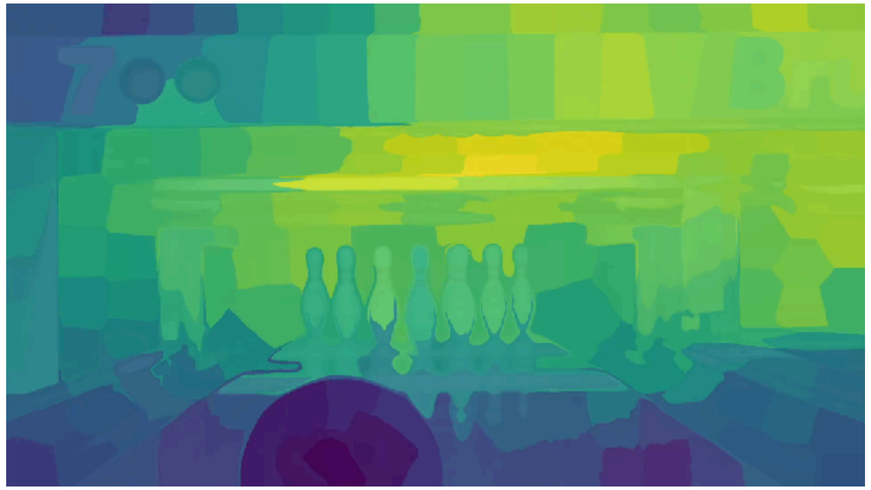


Experimental Results: Geometric Tasks

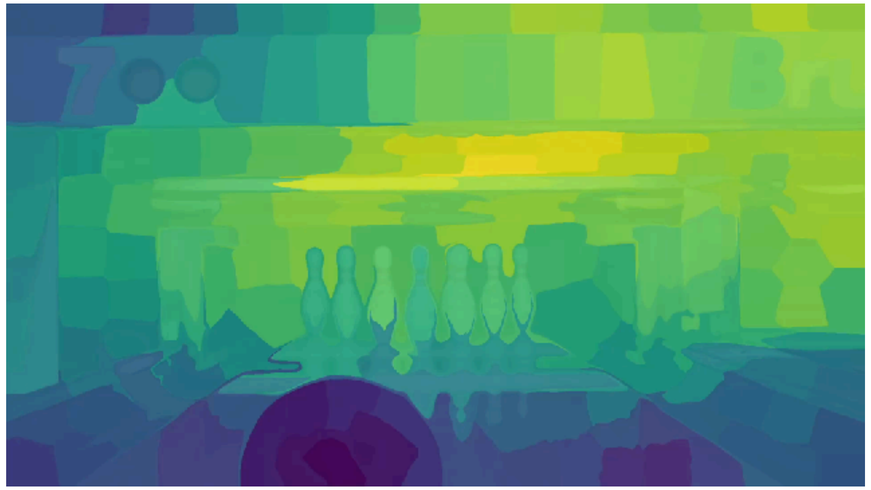


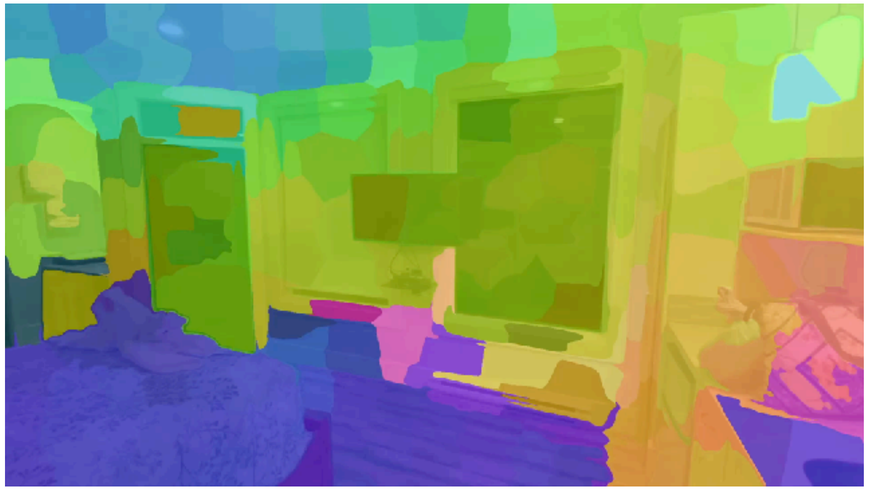


Experimental Results: Geometric Tasks

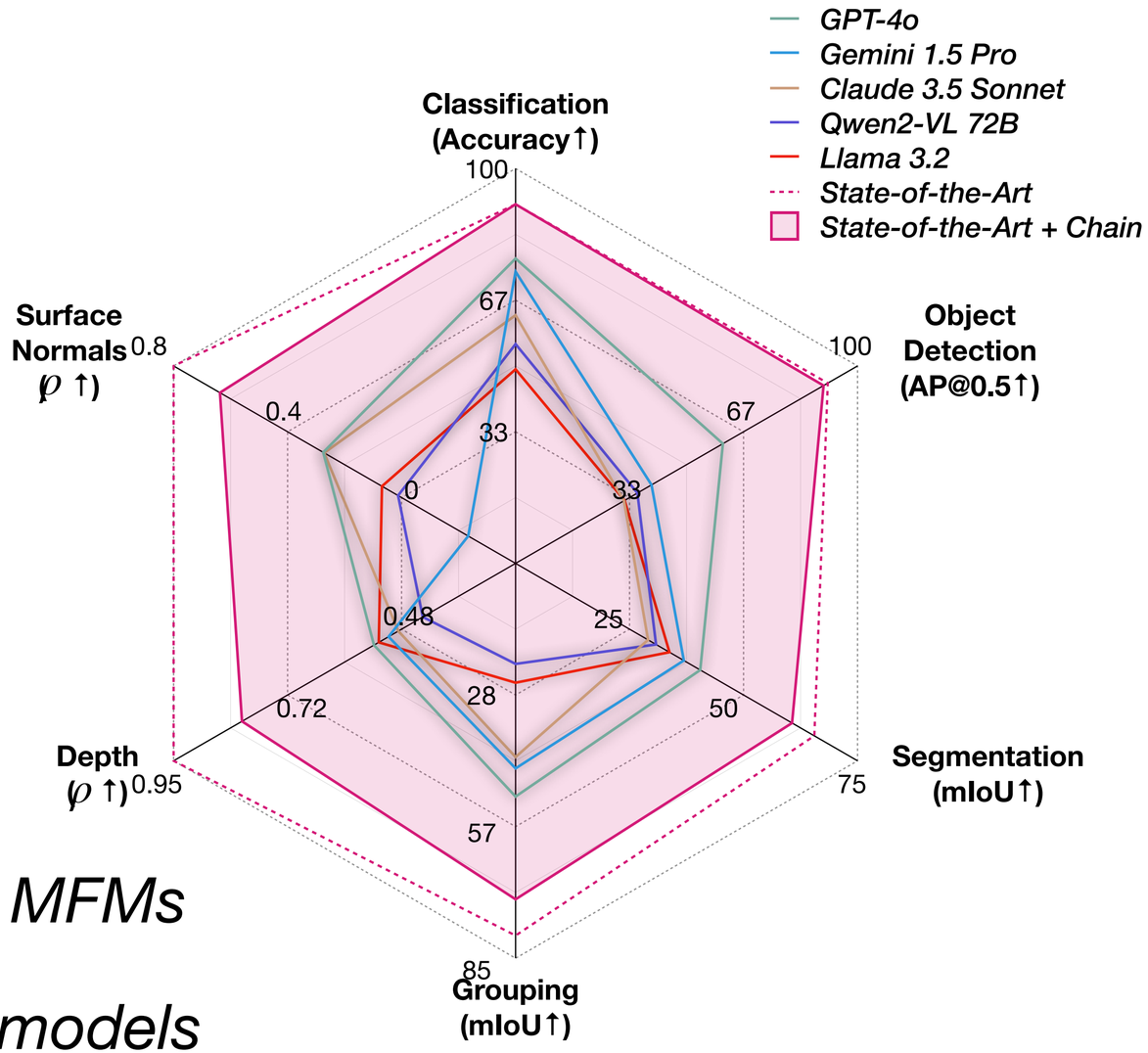


Experimental Results: Geometric Tasks





How Well Does GPT-4o Understand Vision? Evaluating Multimodal Foundation Models on Standard Computer Vision Tasks

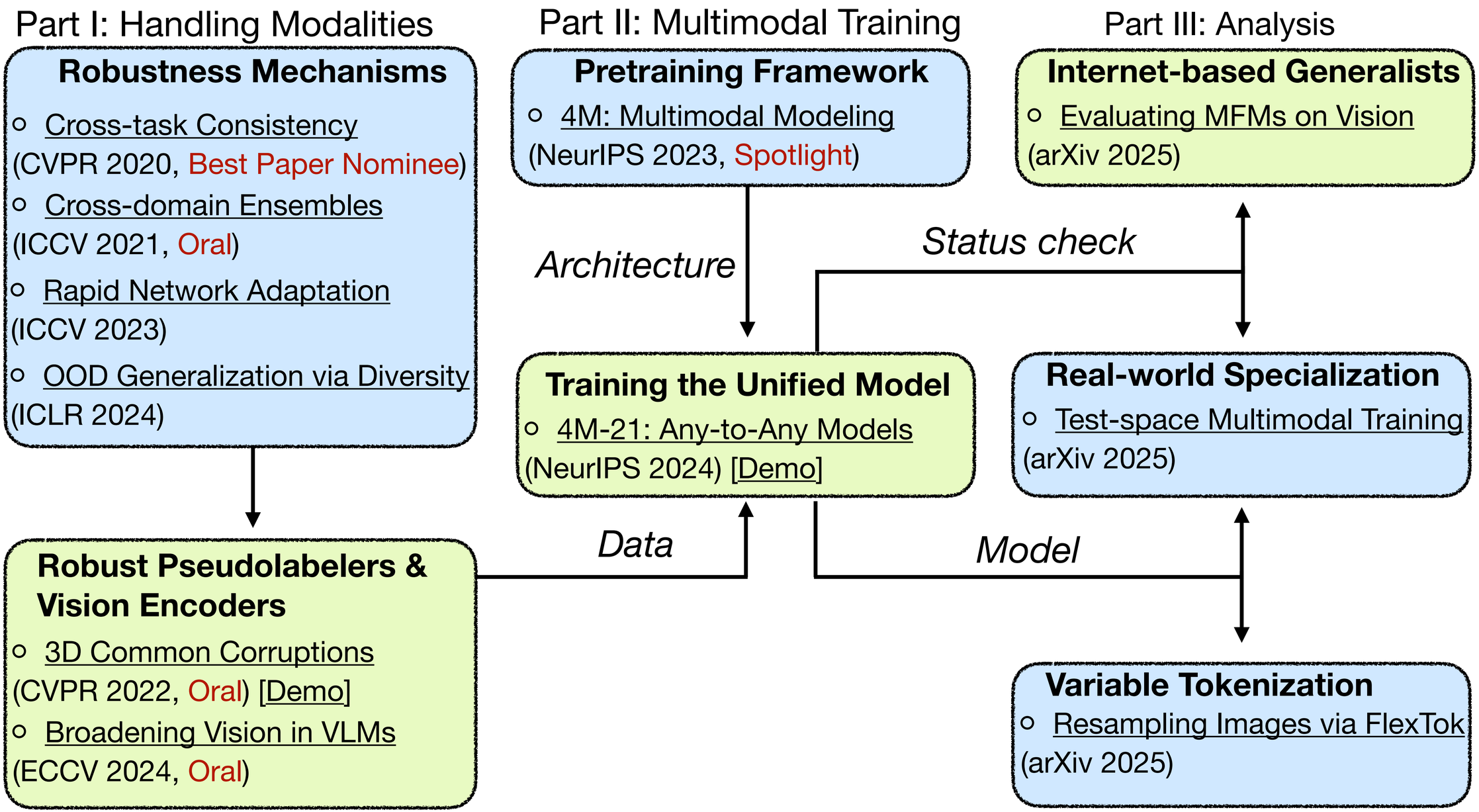


- *A systematic analysis of MFMs*
- *Comparison with vision models*

arXiv 2025

fm-vision-evals.epfl.ch

Publication map & Thesis outline



This thesis: Progress towards capable MFMs

What’s next?

Limitations & Future directions

Modeling Data Capabilities

Limitations & Future directions

Modeling

Architecture & Objectives

Pseudolabelers

Tokenization

Limitations & Future directions

Modeling	Data
Architecture & Objectives	**Temporal & 3D Data**
Pseudolabelers	**Specialization**
Tokenization	**Other Sensors**

Limitations & Future directions

Modeling	Data	Capabilities
Architecture & Objectives	**Temporal & 3D Data**	**Instruction Following**
Pseudolabelers	**Specialization**	**In-context Learning**
Tokenization	**Other Sensors**	**Adaptation**

Thank you to the jury members!

Lenka Zdeborová

Saining Xie

Josh Susskind

Caglar Gulcehre

Amir Zamir

Shout out to my amazing collaborators!

Roman Bachmann

David Mizrahi

Andrei Atanov

Ali Garjani

Rahul Ramachandran

Mingfei Gao

David Griffiths

Jiaming Hu

Petra Poklukar

Achin Kulshrestha

Teresa Yeo

Alessio Tonioni

Federico Tombari

Afshin Dehghan

Amir Zamir

EPFL Thank you for your attention! Google TUM

Publications during PhD

1. O. F. Kar, T. Yeo, A. Atanov, A. Zamir, "3D common corruptions and data augmentation." CVPR, 2022. (**Oral, top 4%**). Project page: Link

2. O. F. Kar, T. Yeo, A. Zamir, "3D common corruptions for object recognition." ICML Shift Happens Workshop, 2022. (**Invited**). Project page: Link

3. O. F. Kar, A. Tonioni, P. Poklukar, A. Kulshrestha, A. Zamir, F. Tombari, "BRAVE: Broadening the visual encoding of vision-language models." ECCV, 2024 (**Oral, top 2%**). Project page: Link

4. R. Bachmann*, **O. F. Kar***, D. Mizrahi*, A. Garjani, M. Gao, D. Griffiths, J. Hu, A. Dehghan, A. Zamir, "4M-21: An Any-to-Any Vision Model for Tens of Tasks and Modalities." NeurIPS, 2024. Project page: Link

5. R. Ramachandran, A. Garjani, R. Bachmann, A. Atanov*, **O. F. Kar***, Amir Zamir*, "How well does GPT-4o understand vision? Evaluating multimodal foundation models on standard computer vision tasks." In submission, 2025. Project page: Link

6. T. Yeo*, **O. F. Kar***, A. Zamir, "Robustness via cross-domain ensembles." ICCV, 2021. (**Oral, top 3%**). Project page: Link

7. R. Bachmann*, J. Allardice*, D. Mizrahi*, E. Fini, **O. F. Kar**, E. Amirloo, A. El-Nouby, A. Zamir, A. Dehghan, "FlexTok: Resampling Images into 1D Token Sequences of Flexible Length." In submission, 2025. Project page: Link

8. K. P. Singh, A. Garjani, M. U. Khattak, R. Singh, A. Atanov*, **O. F. Kar***, A. Zamir, "Test space training for specialized vision models." In submission, 2025.

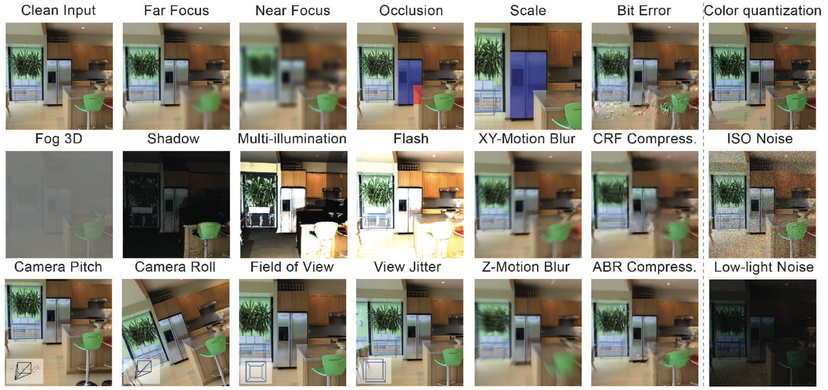
9. H. Benoit*, L. Jiang*, A. Atanov*, **O. F. Kar**, M. Rigotti, A. Zamir, "Unraveling the Key Components of OOD Generalization via Diversification." ICLR, 2024.

10. D. Mizrahi*, R. Bachmann*, **O. F. Kar**, T. Yeo, M. Gao, A. Dehghan, A. Zamir, "4M: Massively Multimodal Masked Modeling." NeurIPS, 2023 (**Spotlight, top 4%**). Project page: Link

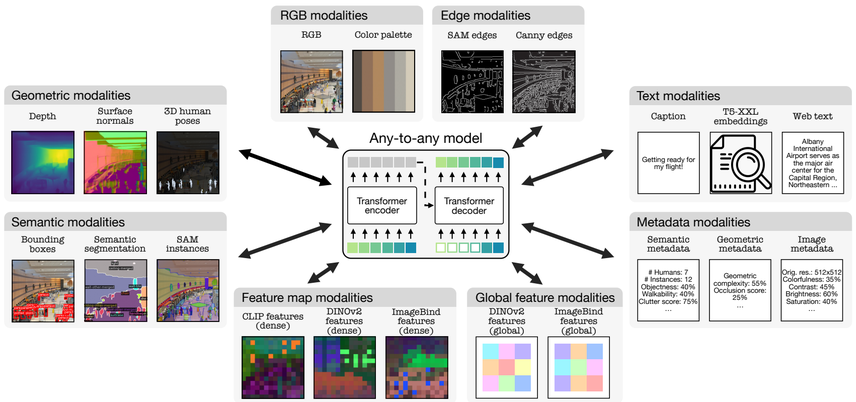
11. T. Yeo, **O. F. Kar**, Z. Sodagar, A. Zamir, "Rapid Network Adaptation: Learning to Adapt Neural Networks Using Test-Time Feedback." ICCV, 2023. Project page: Link

12. A. Zamir*, A. Sax*, T. Yeo, **O. F. Kar**, N. Cheerla, R. Suri, Z. Cao, J. Malik, L. Guibas, "Robust learning through cross-task consistency." arXiv, 2020. CVPR, 2020. (**Oral, best paper award nomination**). Project page: Link

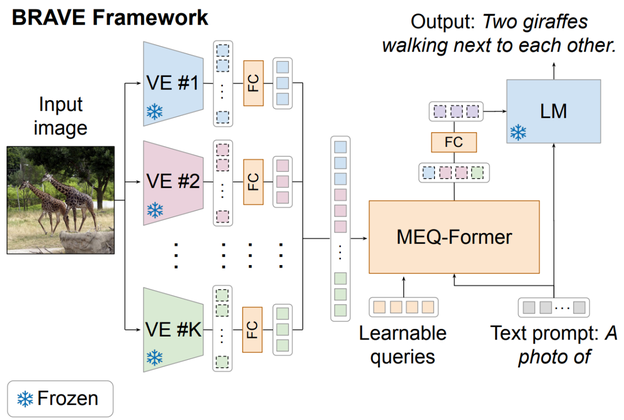
- 12 publications at CVPR, ECCV, ICCV, NeurIPS, ICLR, ICML + ongoing submissions
- 5 oral/spotlight presentations
- 600+ citations
- 2000+ stars on Github for code releases
- Live demos



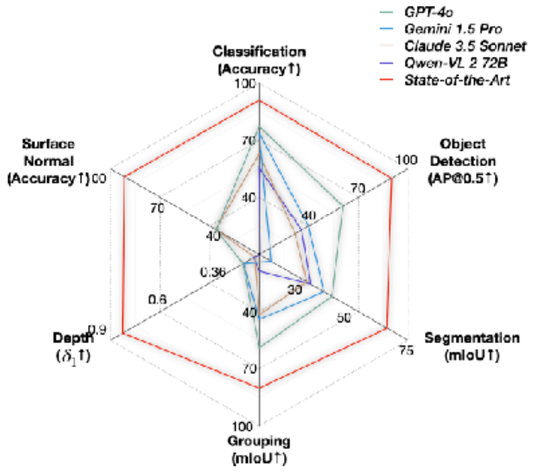
3D Common Corruptions



4M-21: Any-to-Any Models



Improving VLMs w/ BRAVE



Evaluating Vision in MFMs

Thank you for your attention!