UNIFIED-IO: A UNIFIED MODEL FOR VISION, LANGUAGE, AND MULTI-MODAL TASKS


Jiasen Lu


Christopher Clark


Rowan Zellers


Roozbeh Mottaghi

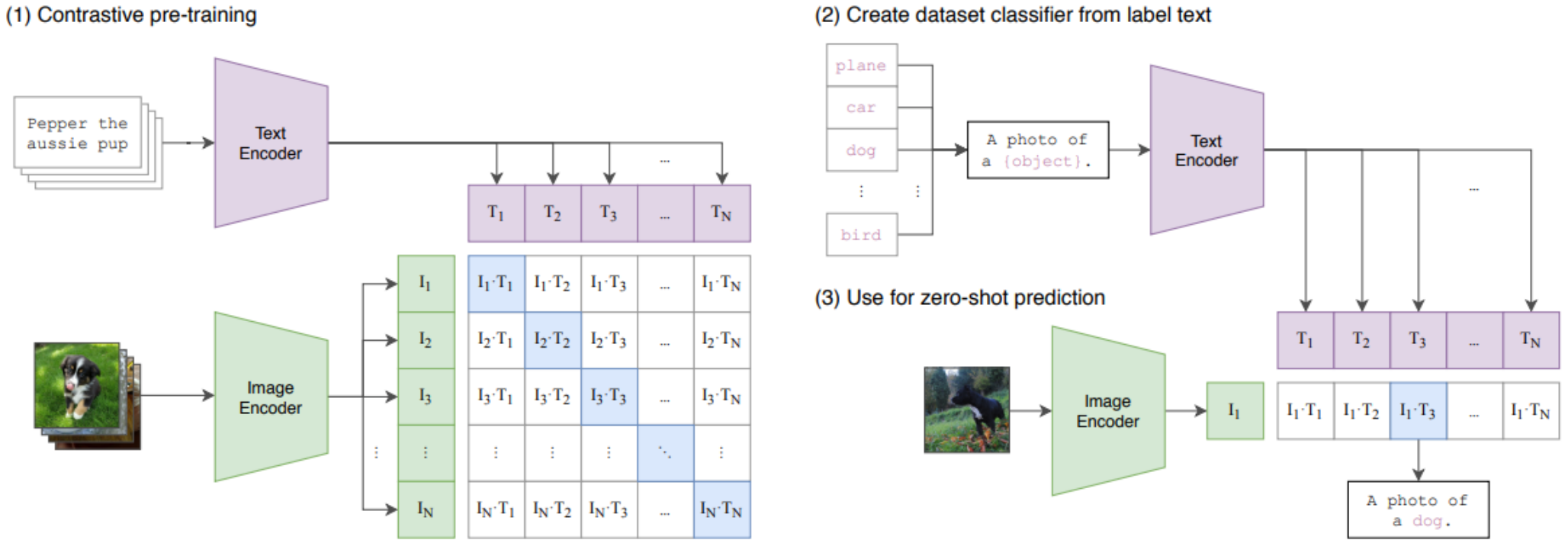

Ani Kembhavi

https://unified-io.allenai.org/

Single-Task Model vs. Unified Model

Single-Task Model	A set of parameters that optimize for single tasks.
Multi-Task Model	Shared parameters + task specific heads for multiple tasks.
Unified Model	Shared parameters + shared heads for multiple tasks.

Single-Task Model for Vision



CLIP: Connecting Text and Images [Radford et.al. 2021]

Single-Task Model for Vision

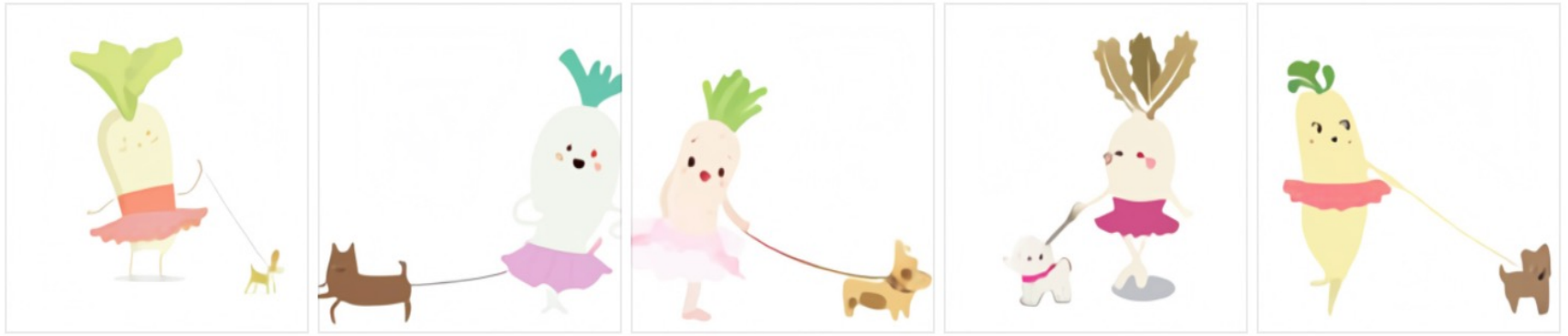
Image classifications

Semantic Segmentations

CLIP: Connecting Text and Images [Radford et.al. 2021]

Single-Task Model for Vision

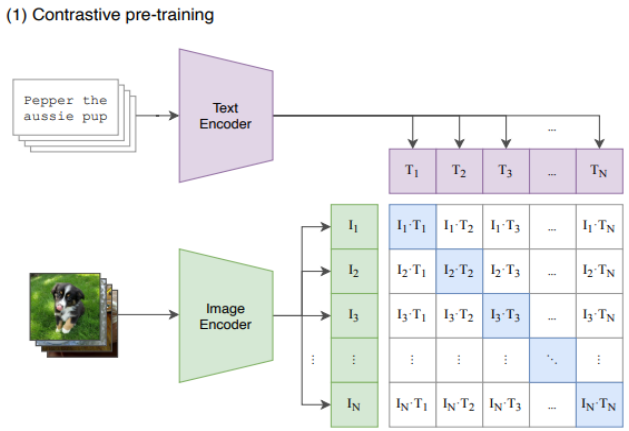
an illustration of a baby daikon radish in a tutu walking a dog



Edit prompt or view more images↓

Zero-Shot Text-to-Image Generation (DALLE) [Ramesh et.al. 2021]

Single-Task Model for Vision



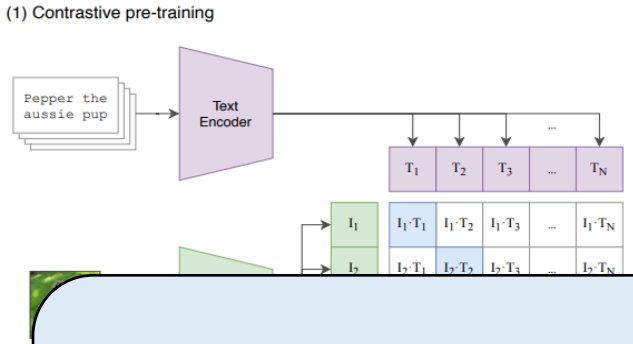
CLIP: Connecting Text and Images [Radford et.al. 2021]

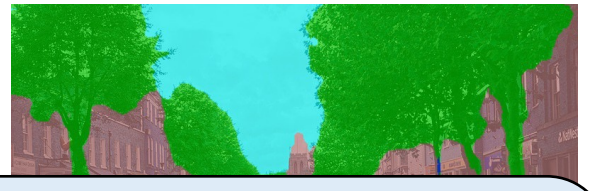
Semantic Segmentations

Zero-Shot Text-to-Image Generation (DALLE) [Ramesh et.al. 2021]

Image classifications

Single-Task Model for Vision





CLIP:

If the model can learn how to generate images, they should be able to solve any semantic vision tasks.

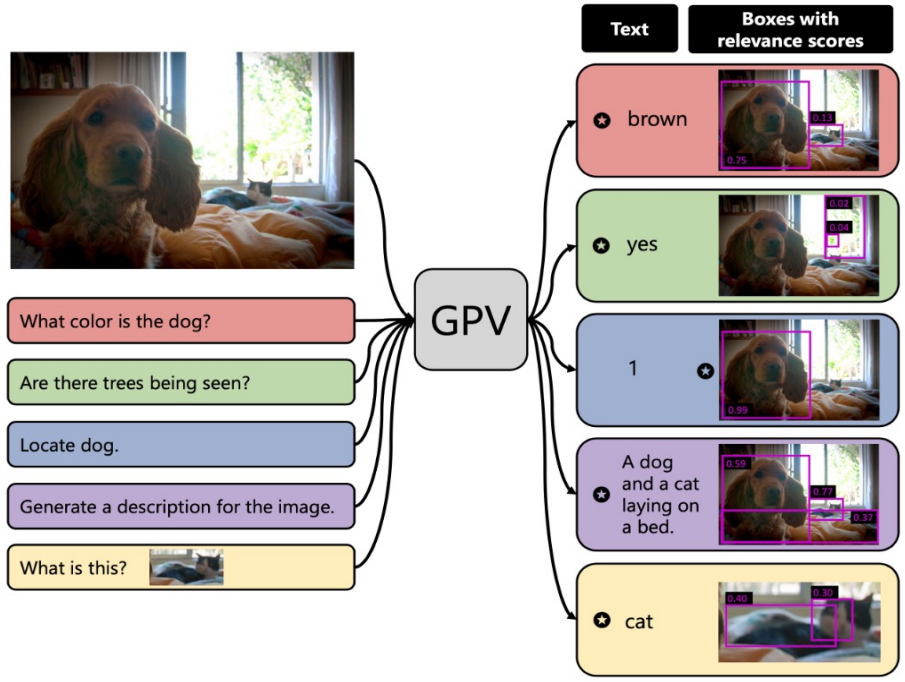
an illustration of a baby daikon radish in a tutu walking a dog

Edit prompt or view more images↓

Zero-Shot Text-to-Image Generation (DALLE) [Ramesh et.al. 2021]

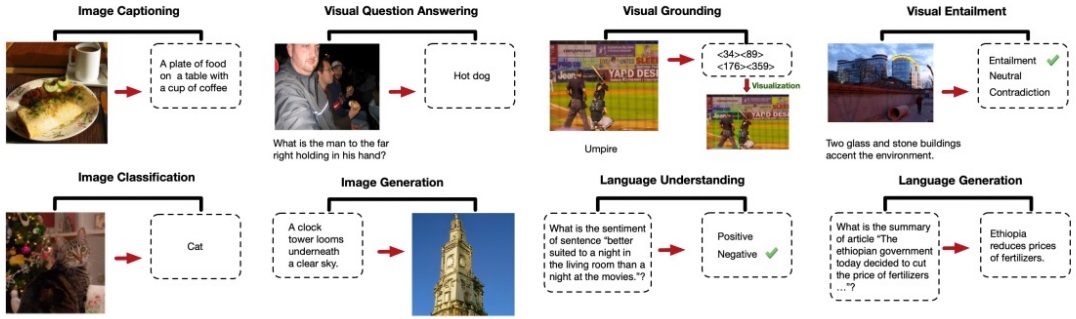
Image classifications

Prior Work



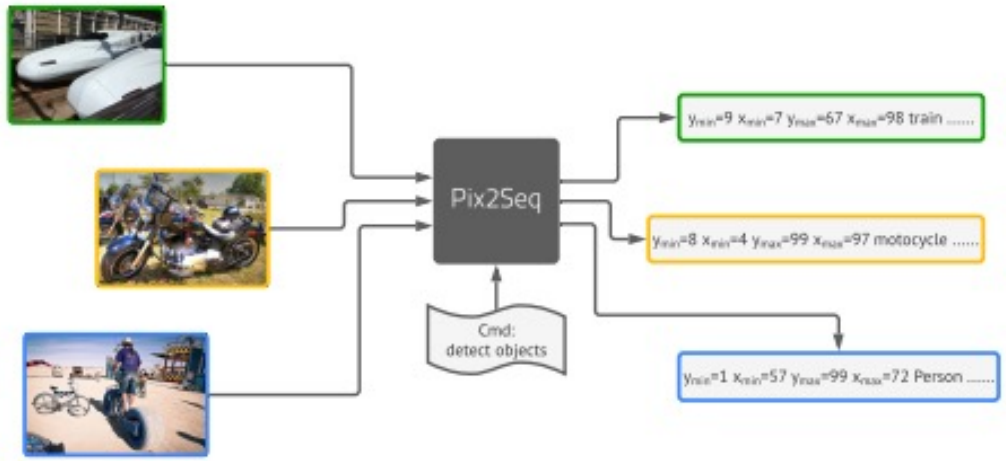
GPV [Gupta et.al. 2021]

Perceiver-IO [Jaegle et.al. 2021]

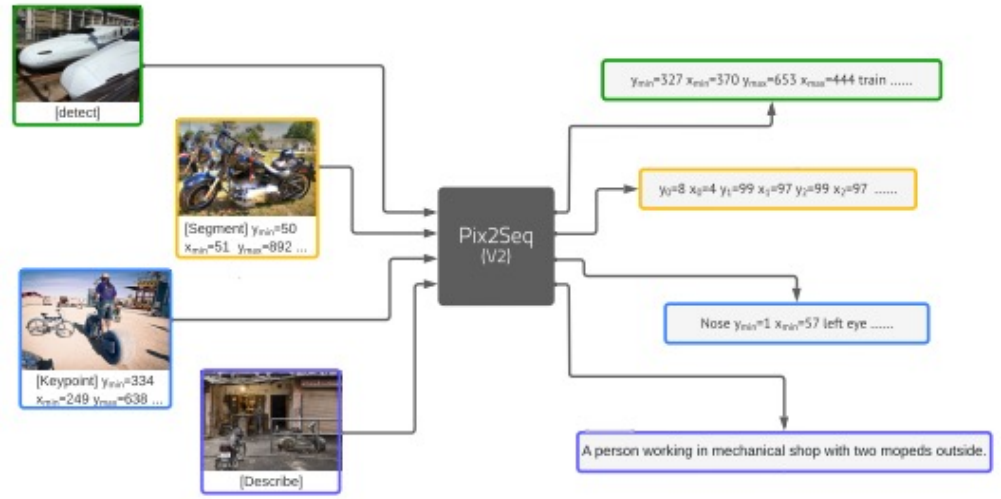


OFA [Wang et.al. 2022]

Prior Work

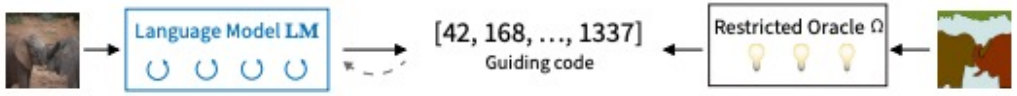


Pixel2seq [Chen et.al. 2021]



Pixel2seq v2 [Chen et.al. 2022]

(a) **Stage I** training: we train the base model f, which is guided by the code produced by the *restricted oracle* model Ω. The oracle has access to the ground-truth label, but is only allowed to communicate with f by passing a short discrete sequence, which we call a *guiding code*.

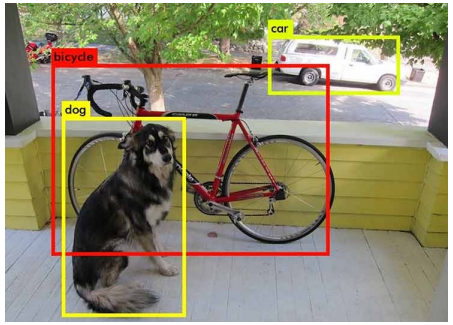


(b) **Stage II** training: we train a *language model* (LM) to output a *guiding code* by learning to mimic the oracle, but using only the image input.

UViM [Kolesnikov et.al. 2022]

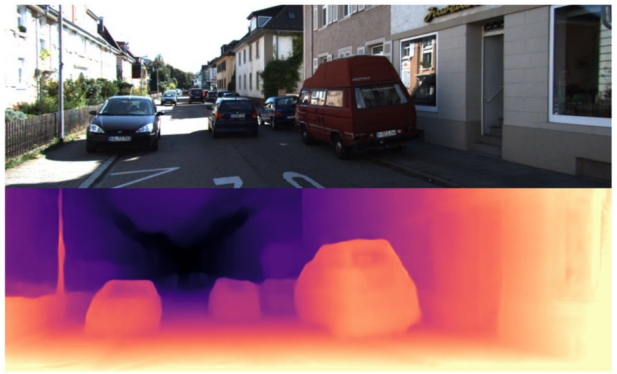
Vision has diverse output format

Image classifications



Object detections

Semantic Segmentations



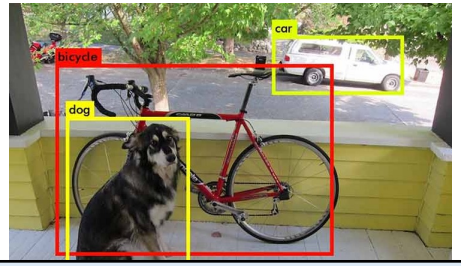
Depth Estimation

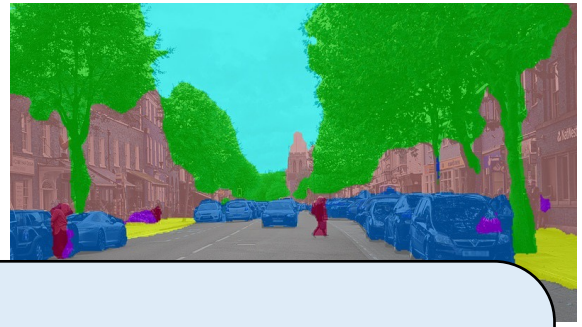
Pose Estimation

& MORE

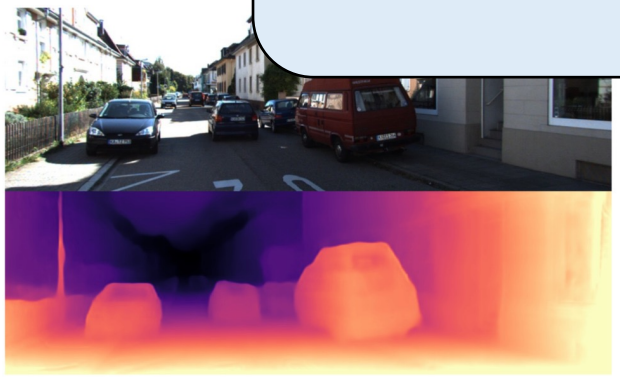
Vision has diverse output format

Image cl



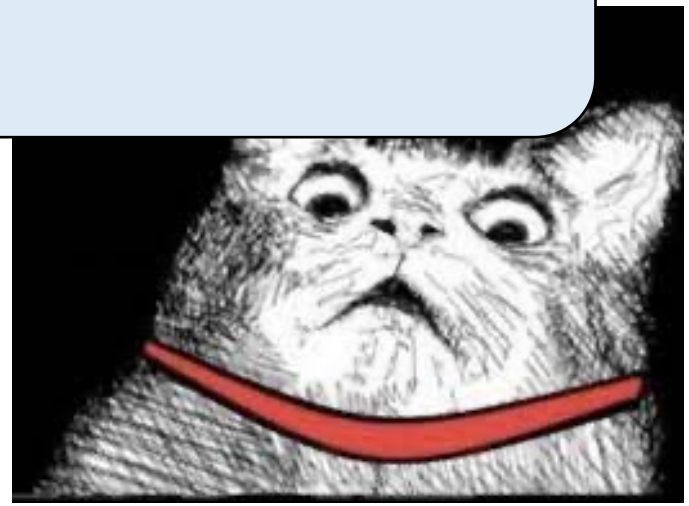


Convert diverse vision inputs/outputs into sequences.



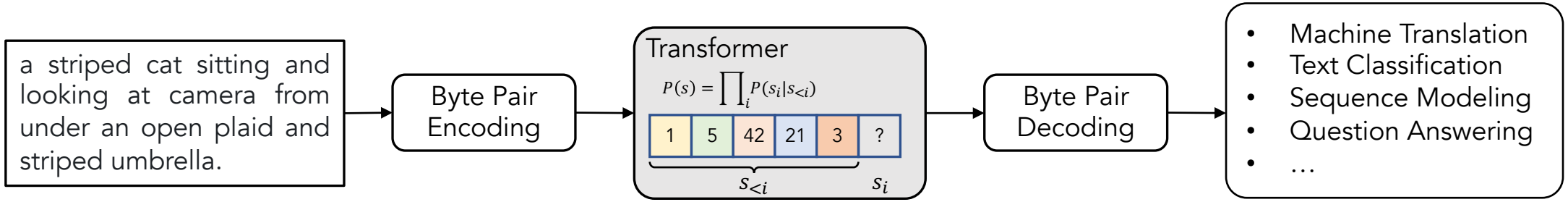
Depth Estimation

Pose Estimation



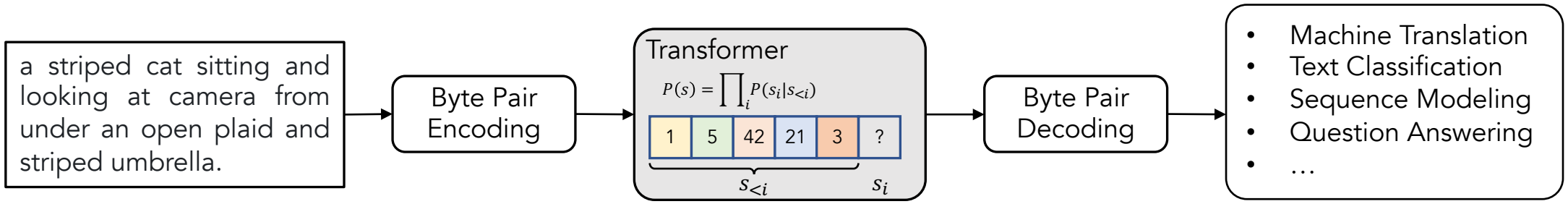
Unified IO

- **NLP Task Pipeline**

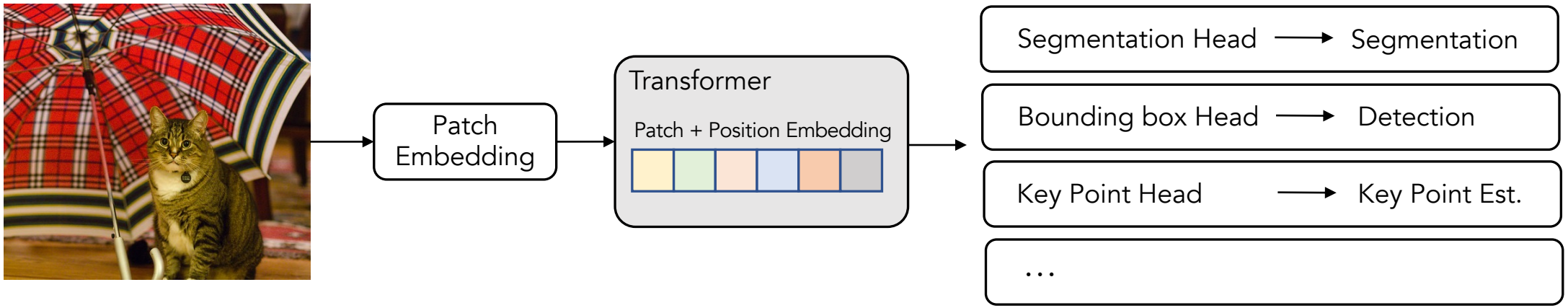


Unified IO

- **NLP Task Pipeline**

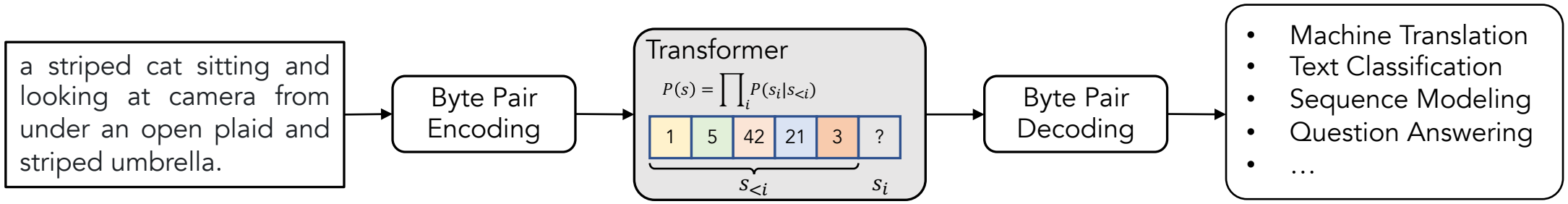


- **Vision Task Pipeline**

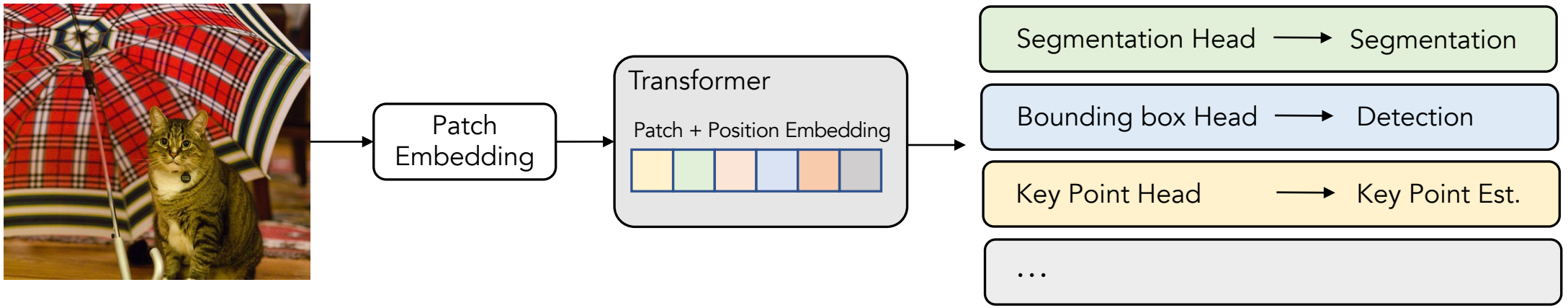


Unified IO

- **NLP Task Pipeline**

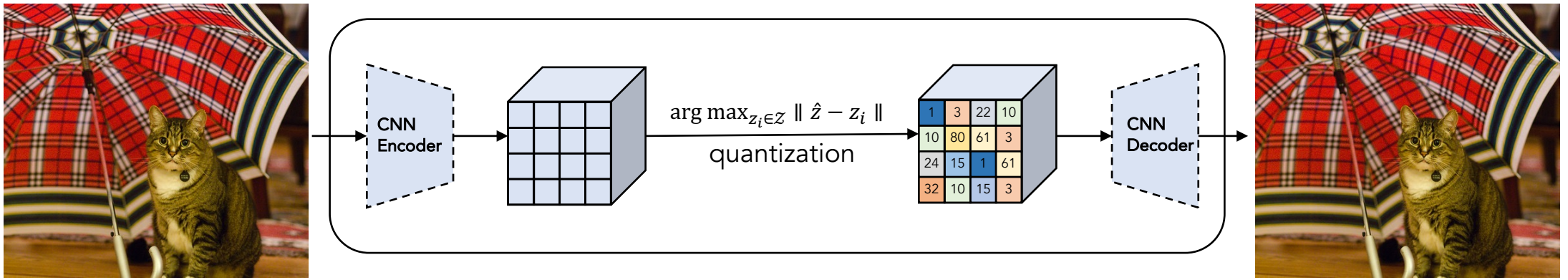


- **Vision Task Pipeline**



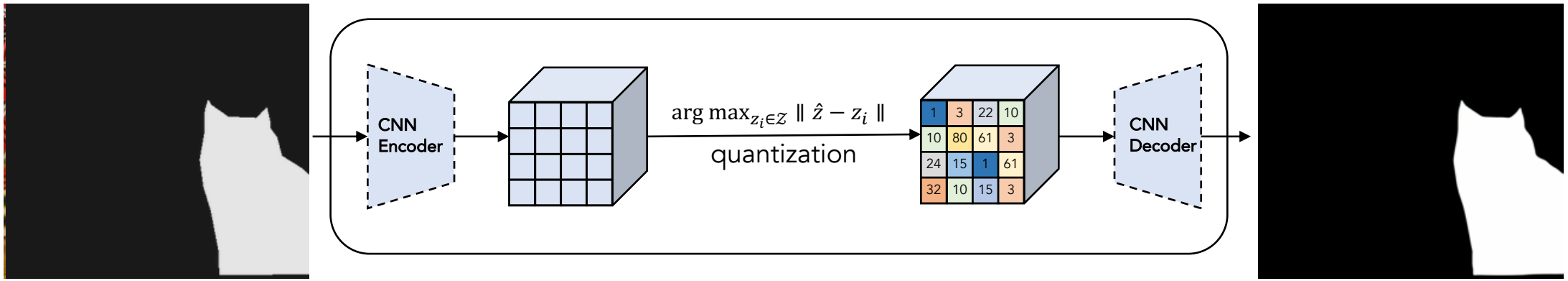
Unified IO

- Image Quantization Using VQ-VAE



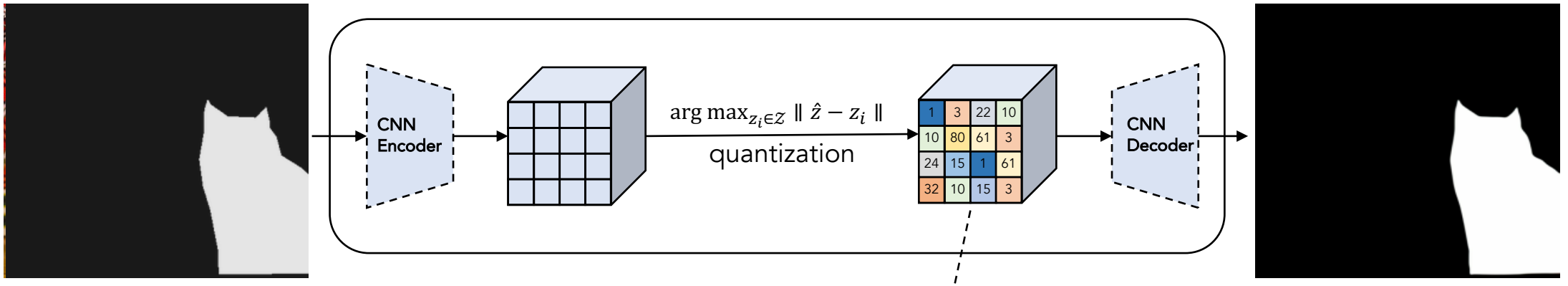
Unified IO

- Image Quantization Using VQ-VAE



Unified IO

- Image Quantization Using VQ-VAE



- Image Segmentation Using Unified Decoding

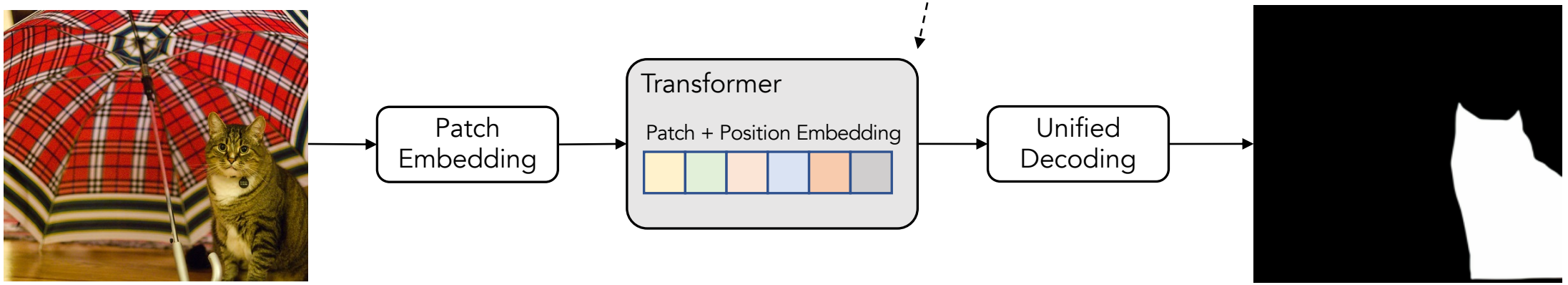


Image Output Quantization

- **Task with 2D structured outputs.**

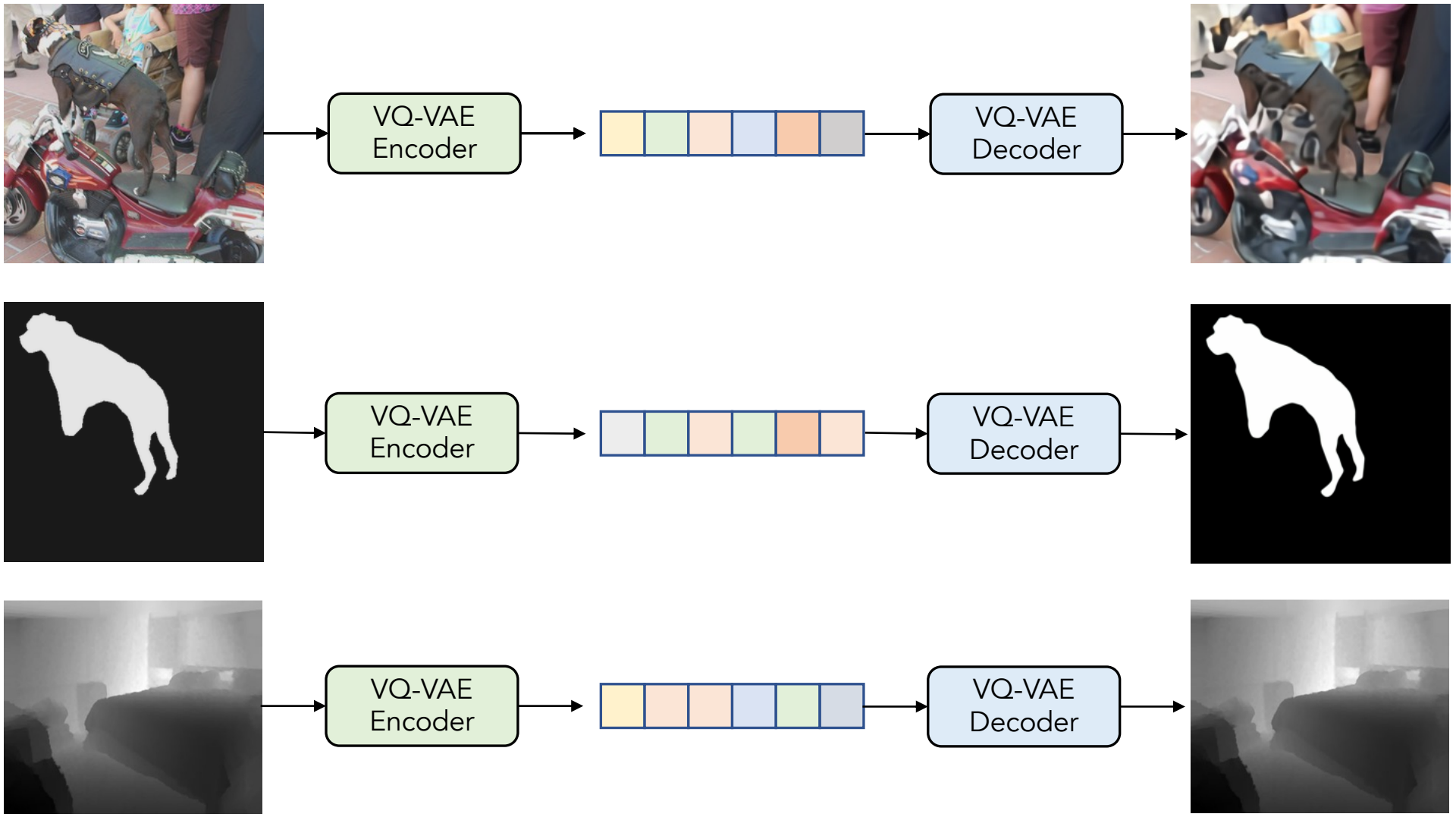
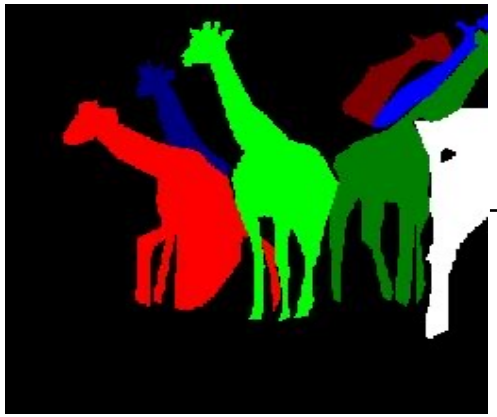
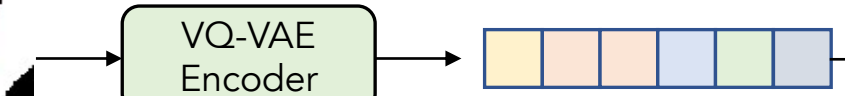


Image Output Quantization

- **Task with 2D structured outputs.**

What is the segmentation of " giraffe " ?





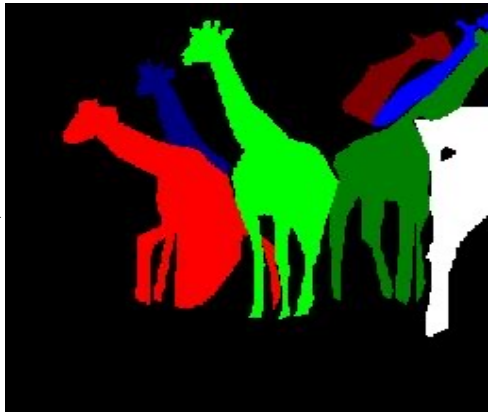


Image Output Quantization

- **Task with 1D structured outputs**

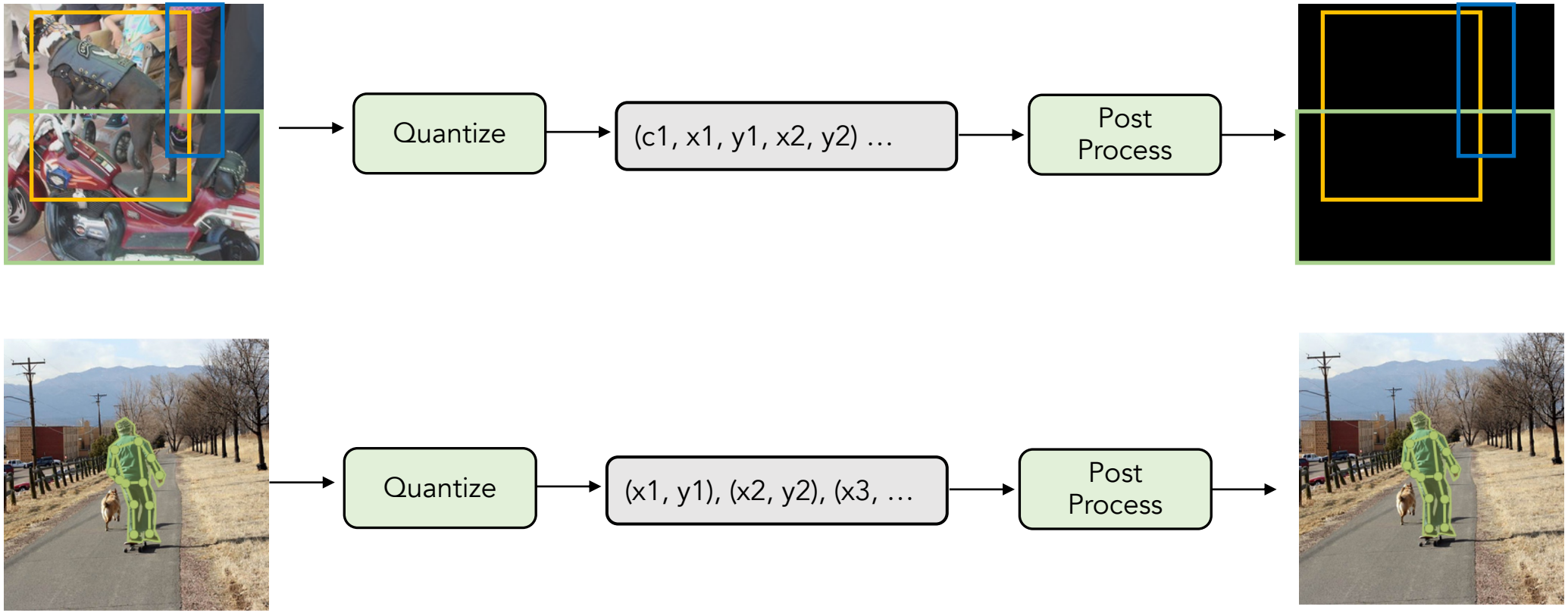


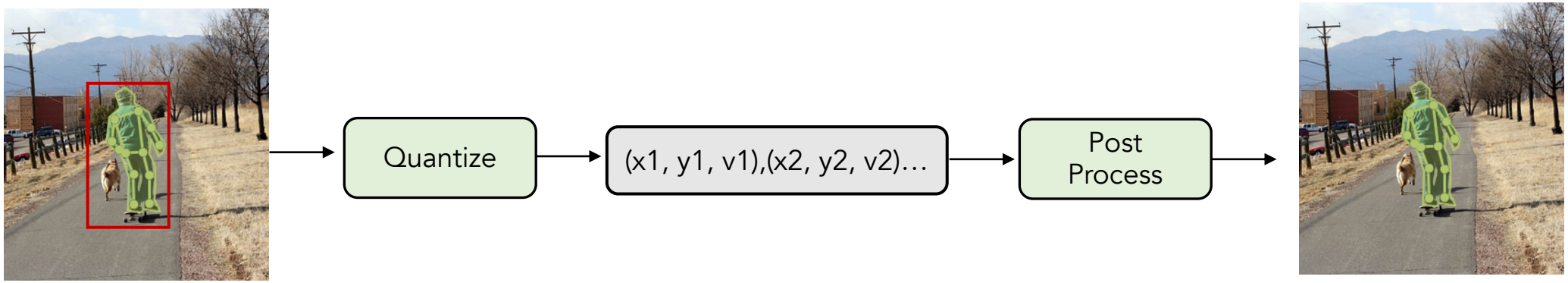
Image Output Quantization

- **Task with 1D structured outputs**

  Sequence length: 17 x 3 = 51 per person

  Take bounding box as input and detect one person each time.

  human pose estimation --> detection + single human pose estimation



Text Input for Different Tasks

- **Prompt driven task specification**

Categorization

CLASSIFICATION IN CONTEXT

What is the category of the region "loc389 loc560 loc684 loc684"?

horse

TRUTH | PREDICTION

horse

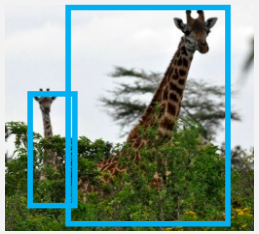
Localization

OBJECT LOCALIZATION

Which regions does the text "giraffe" describe ?

loc431 loc202 loc827 loc335 giraffe loc109 loc337 loc934 loc954 giraffe

TRUTH | PREDICTION



** text outputs omitted for brevity*

VQA

VISUAL QUESTION ANSWERING

How many signs are there ?

2

TRUTH | PREDICTION

2

Text Input for Different Tasks

- **Prompt driven task specification**

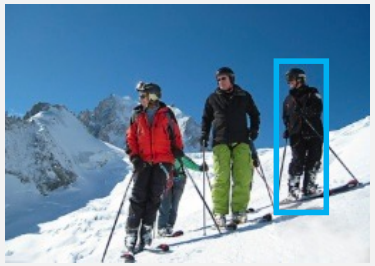
Refer Expression

REFERRING EXPRESSIONS

Which region does the text "man in solid black" describe ?

loc270 loc856 loc640 loc982

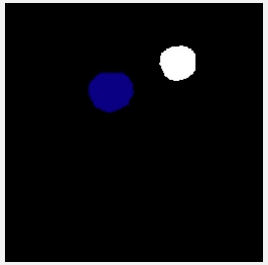
TRUTH | PREDICTION



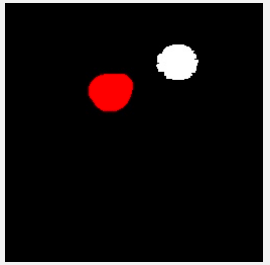
Segmentation

OBJECT SEGMENTATION

What is the segmentation of "apple" ?



TRUTH | PREDICTION



Keypoints

KEYPOINT ESTMATION

Find the human joints in the region "loc260 loc375 loc726 loc545".

loc307 loc487 loc299 loc499 loc295 loc481 loc315 loc507 loc305 loc455 loc369 loc517 loc359 loc413 loc457 loc527 loc429 loc387 loc423 loc529 ….

TRUTH | PREDICTION

** text outputs omitted for brevity*

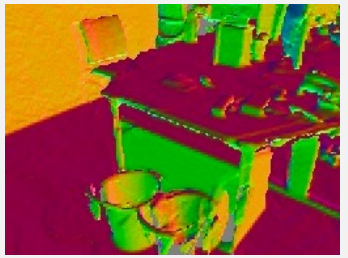
Text Input for Different Tasks

- **Prompt driven task specification**

Surface Normal

SURFACE NORMAL ESTIMATION

What is the surface normal of the image ?



TRUTH | PREDICTION

Image Generation

IMAGE GENERATION

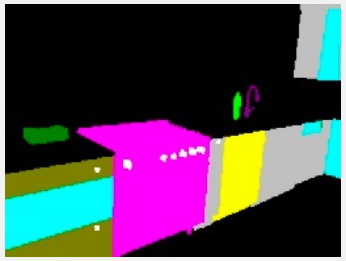
What is the complete image? Text: "a white sink in a small tiled bathroom"

TRUTH | PREDICTION

Seg based Image Generation

SEG BASED IMAGE GENERATION

What is the complete image? Segmentation color: "white: knob, silver: cupboard, olive: drawer, lime …"



TRUTH | PREDICTION

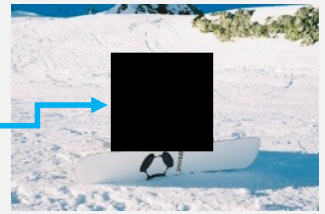
Text Input for Different Tasks

- **Prompt driven task specification**

Image Inpating

IMAGE INPAINTING

Fill in the blank region "loc257 loc425 loc575 loc758" person"?



TRUTH | PREDICTION

Depth Prediction

DEPTH ESTIMATION

What is the depth map of the image ?

TRUTH | PREDICTION

NLP Tasks

TEXT CLASSIFICATION

context: Swansea striker Lee Trundle has negotiated a lucrative image-rights deal with the League One club. Lee Trundle is in business with the League One club.
question: Does this sentence entail the following sentence?

Yes it entails

TRUTH | PREDICTION

Yes it entails

Model

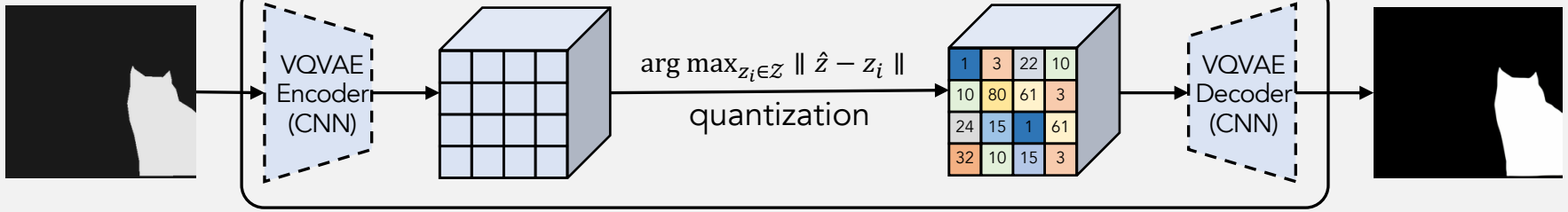


Image serialization using a VQ-VAE

Model

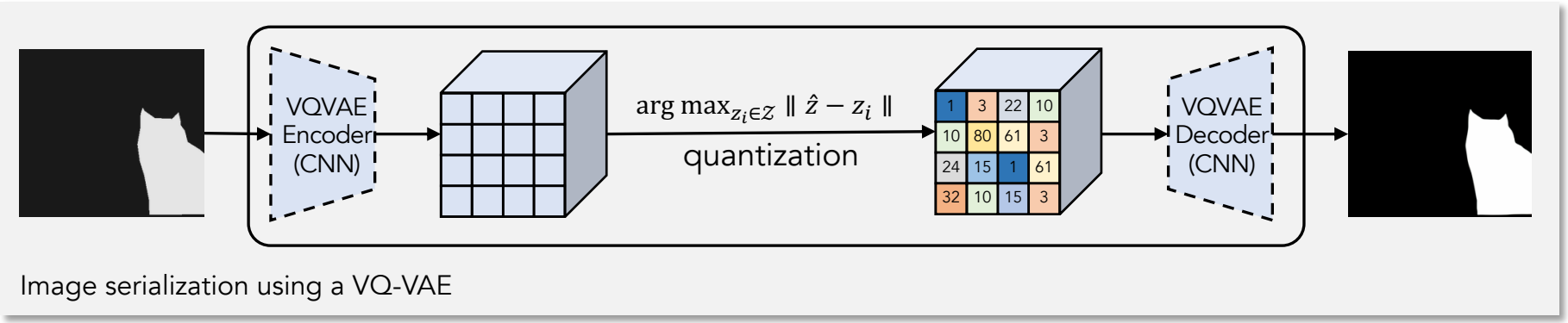


Image serialization using a VQ-VAE

Segment the cat

What is the color of the cat?

What is the depth map of the image?

Localize the cat

Model

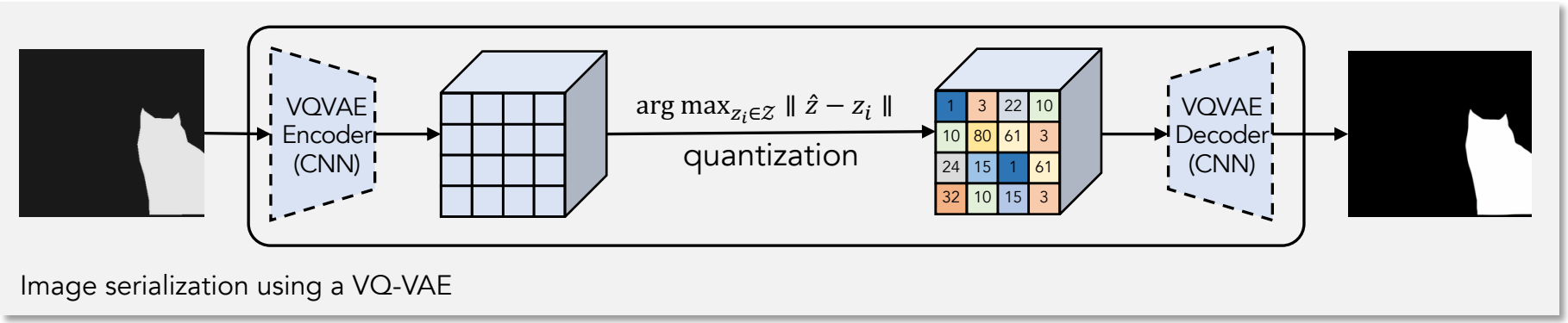
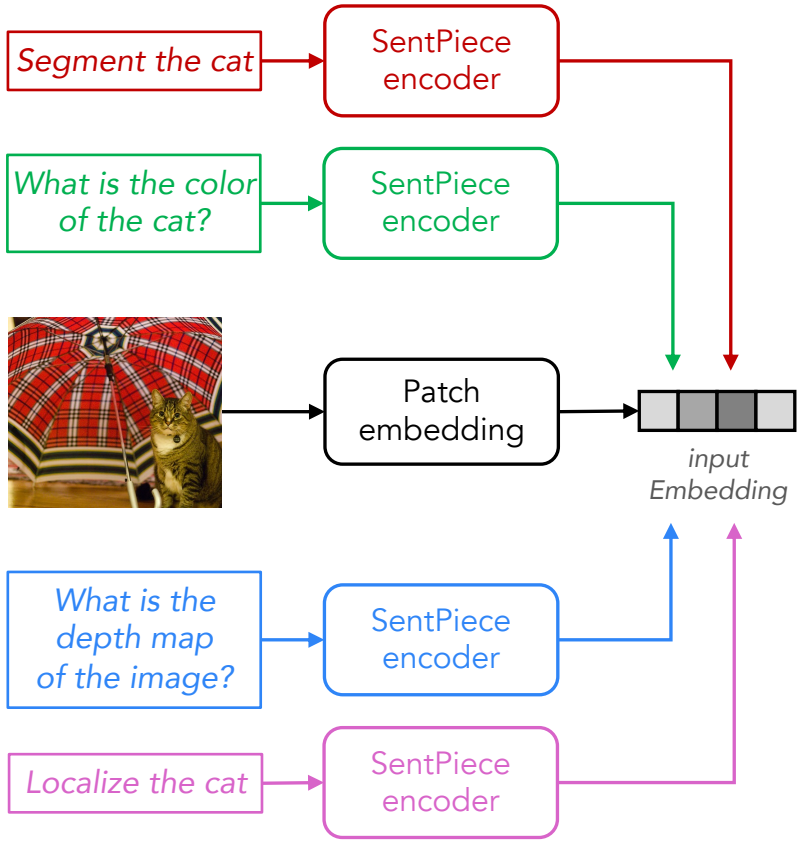


Image serialization using a VQ-VAE



Model

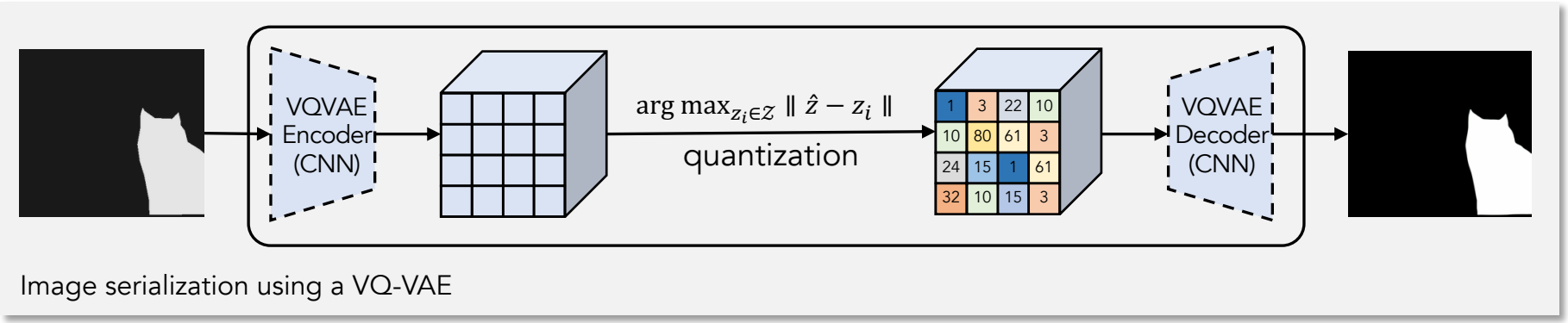
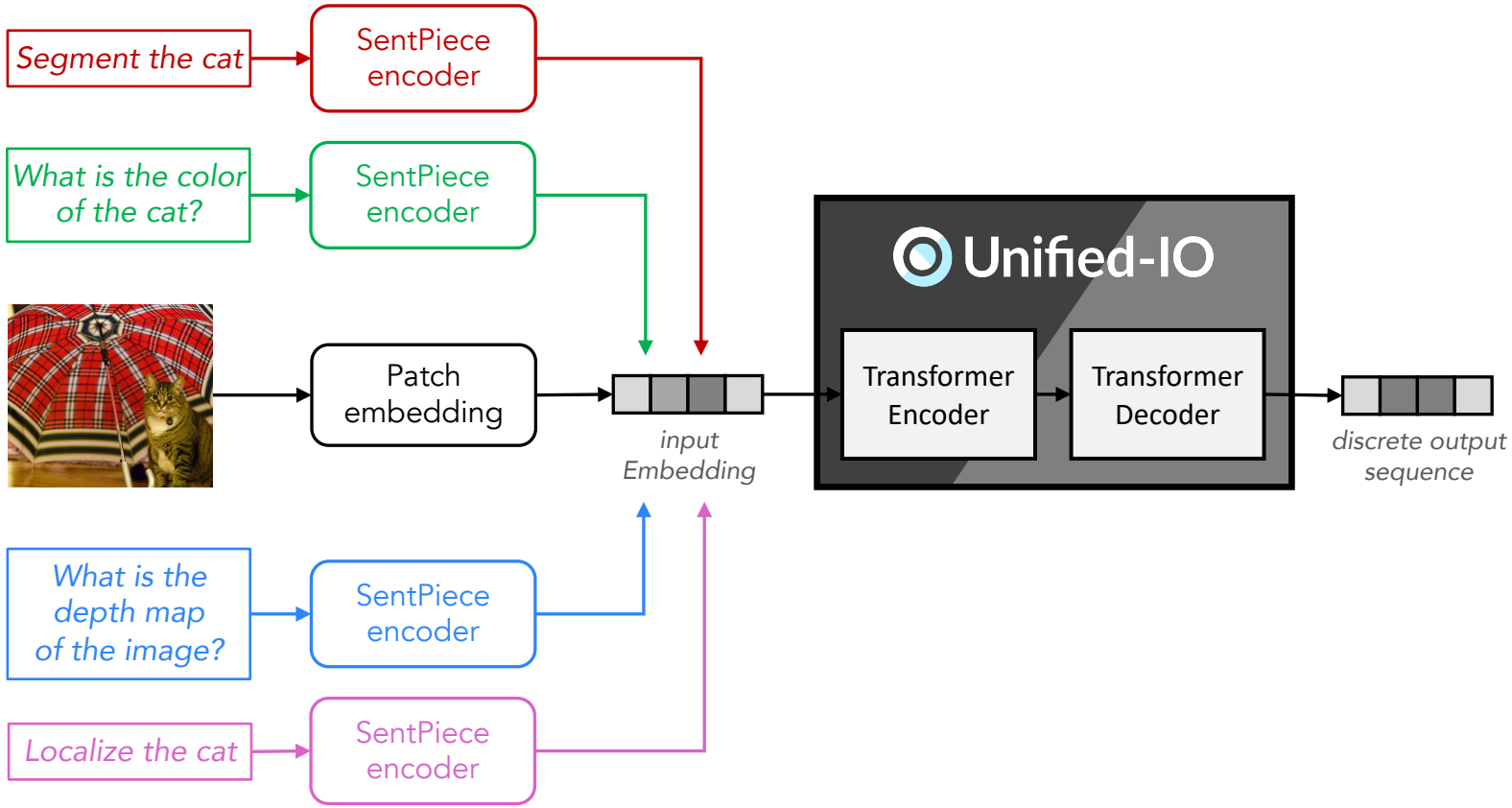


Image serialization using a VQ-VAE



Model

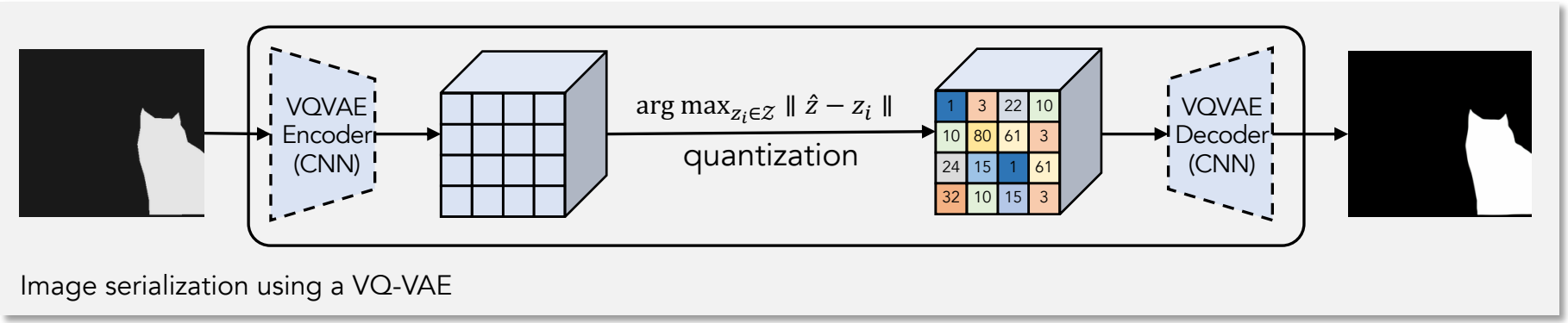
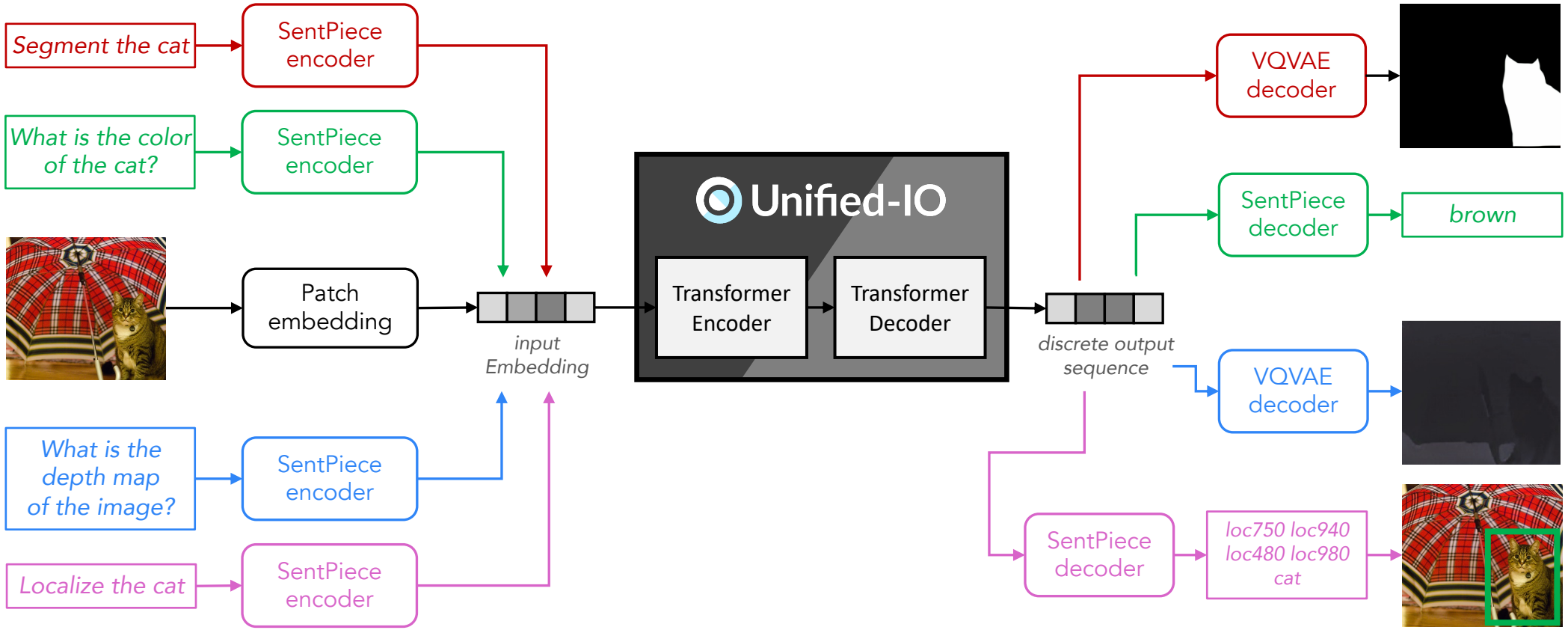
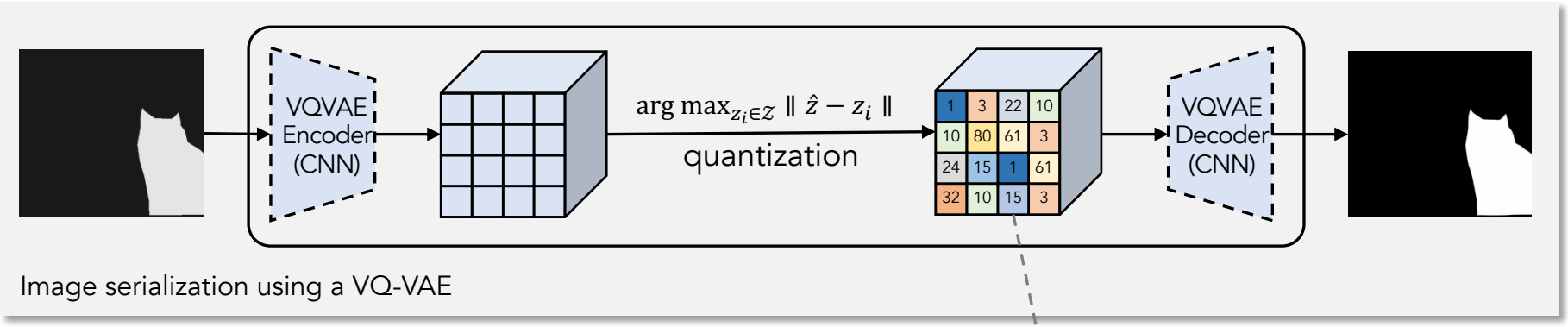


Image serialization using a VQ-VAE



Model



Segment the cat

SentPiece encoder

What is the color of the cat?

SentPiece encoder

Patch embedding

What is the depth map of the image?

SentPiece encoder

Localize the cat

SentPiece encoder

input Embedding

Unified-IO

Transformer Encoder

Transformer Decoder

Discrete sequence serves as the target

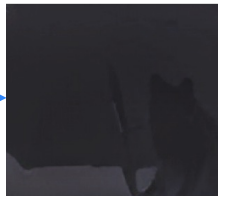
discrete output sequence

VQVAE decoder

SentPiece decoder

brown

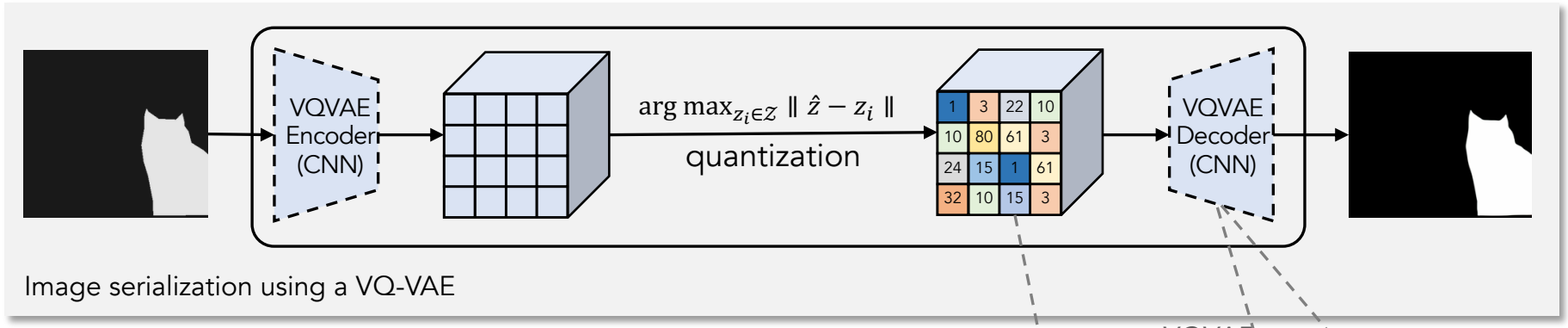
VQVAE decoder

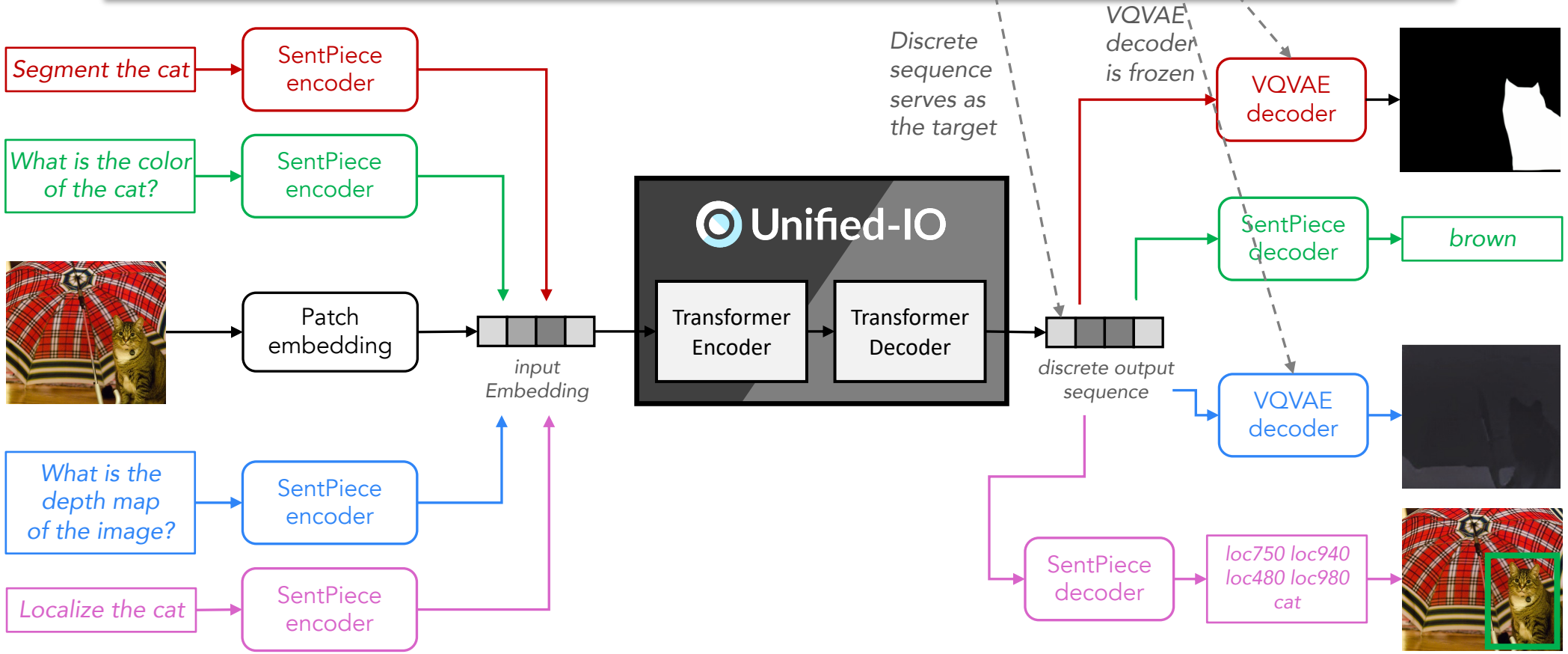


SentPiece decoder

loc750 loc940 loc480 loc980 cat

Model


VQVAE Encoder (CNN)
$\arg\max_{z_i \in \mathcal{Z}} \| \hat{z} - z_i \|$
quantization
1 3 22 10
10 80 61 3
24 15 1 61
32 10 15 3
VQVAE Decoder (CNN)
Image serialization using a VQ-VAE


Segment the cat
SentPiece encoder
What is the color of the cat?
SentPiece encoder
Patch embedding
What is the depth map of the image?
SentPiece encoder
Localize the cat
SentPiece encoder
input Embedding
Unified-IO
Transformer Encoder
Transformer Decoder
Discrete sequence serves as the target
VQVAE decoder is frozen
discrete output sequence
VQVAE decoder
SentPiece decoder
brown
VQVAE decoder
SentPiece decoder
loc750 loc940 loc480 loc980 cat

Model Details

- Follow T5 implementation with minimum modification.
- Use pretrained VQ-GAN from Imagenet.
- Use patch encoding for image inputs (no CNN involved)
- Absolute position encoding is important for vision – injecting the position embedding at input
- Relative position encoding within text and image.

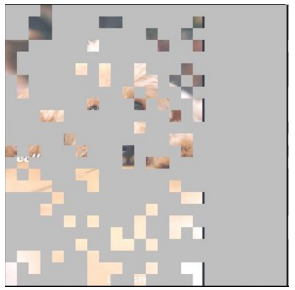
```
[ 5,  4,  4,  3,  2,  1,  0,  1,  2,  3,  4,  4,  5,  5],
```

Relative encoding for the text

```
[[45, 44, 44, 43, 42, 41, 40, 41, 42, 43, 44, 44, 45, 45],
 [37, 36, 36, 35, 34, 33, 32, 33, 34, 35, 36, 36, 37, 37],
 [37, 36, 36, 35, 34, 33, 32, 33, 34, 35, 36, 36, 37, 37],
 [29, 28, 28, 27, 26, 25, 24, 25, 26, 27, 28, 28, 29, 29],
 [21, 20, 20, 19, 18, 17, 16, 17, 18, 19, 20, 20, 21, 21],
 [13, 12, 12, 11, 10,  9,  8,  9, 10, 11, 12, 12, 13, 13],
 [ 5,  4,  4,  3,  2,  1,  0,  1,  2,  3,  4,  4,  5,  5],
 [13, 12, 12, 11, 10,  9,  8,  9, 10, 11, 12, 12, 13, 13],
 [21, 20, 20, 19, 18, 17, 16, 17, 18, 19, 20, 20, 21, 21],
 [29, 28, 28, 27, 26, 25, 24, 25, 26, 27, 28, 28, 29, 29],
 [37, 36, 36, 35, 34, 33, 32, 33, 34, 35, 36, 36, 37, 37],
 [37, 36, 36, 35, 34, 33, 32, 33, 34, 35, 36, 36, 37, 37],
 [45, 44, 44, 43, 42, 41, 40, 41, 42, 43, 44, 44, 45, 45],
 [45, 44, 44, 43, 42, 41, 40, 41, 42, 43, 44, 44, 45, 45]]
```

Relative encoding for the Image

Objective

- **Pre-Training**
  - Text span denoising (15%): An image of a <M> is lying on the <M>. ⇨ <M> dog <M> ground.
  - Mask image denoising (75%):  ⇨ 
- **Multi-Task Training**
  - Jointly train with 80 vision, vision language and language datasets/sets.

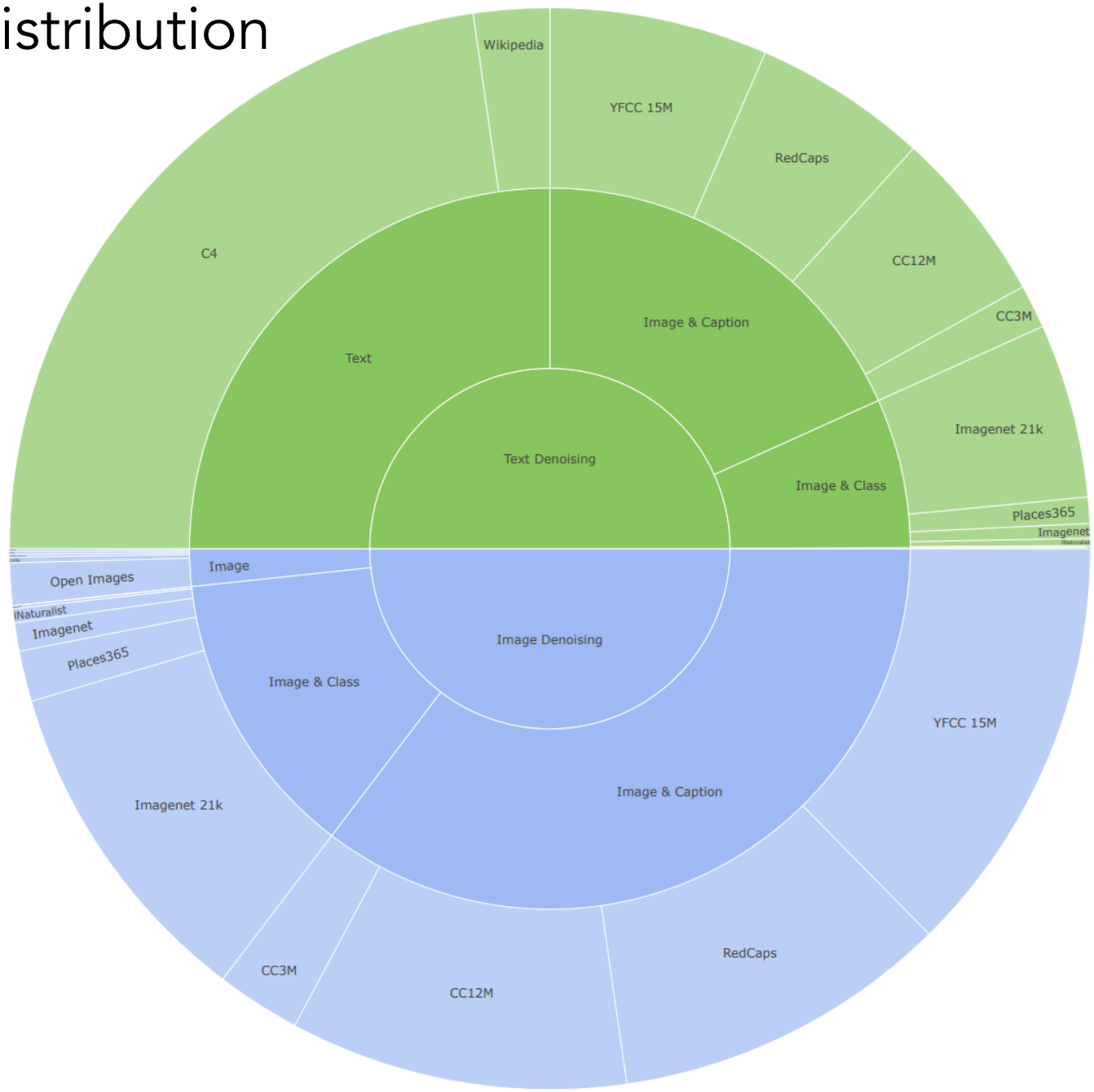
Tasks

Vision Tasks	V&L Tasks	Language Tasks
Image Classification	Image Captioning	GLUE
Object Detection	Visual Question Answering	Commonsense QA
Semantic Segmentation	Image Generation	SNLI
Depth Estimation	Refer Expression	SQUAD
Image Inpainting	VCR	SWAG
Pose Estimation	Region Captioning	Openbook QA
Relationship Detection	Situation Recognition	Fever

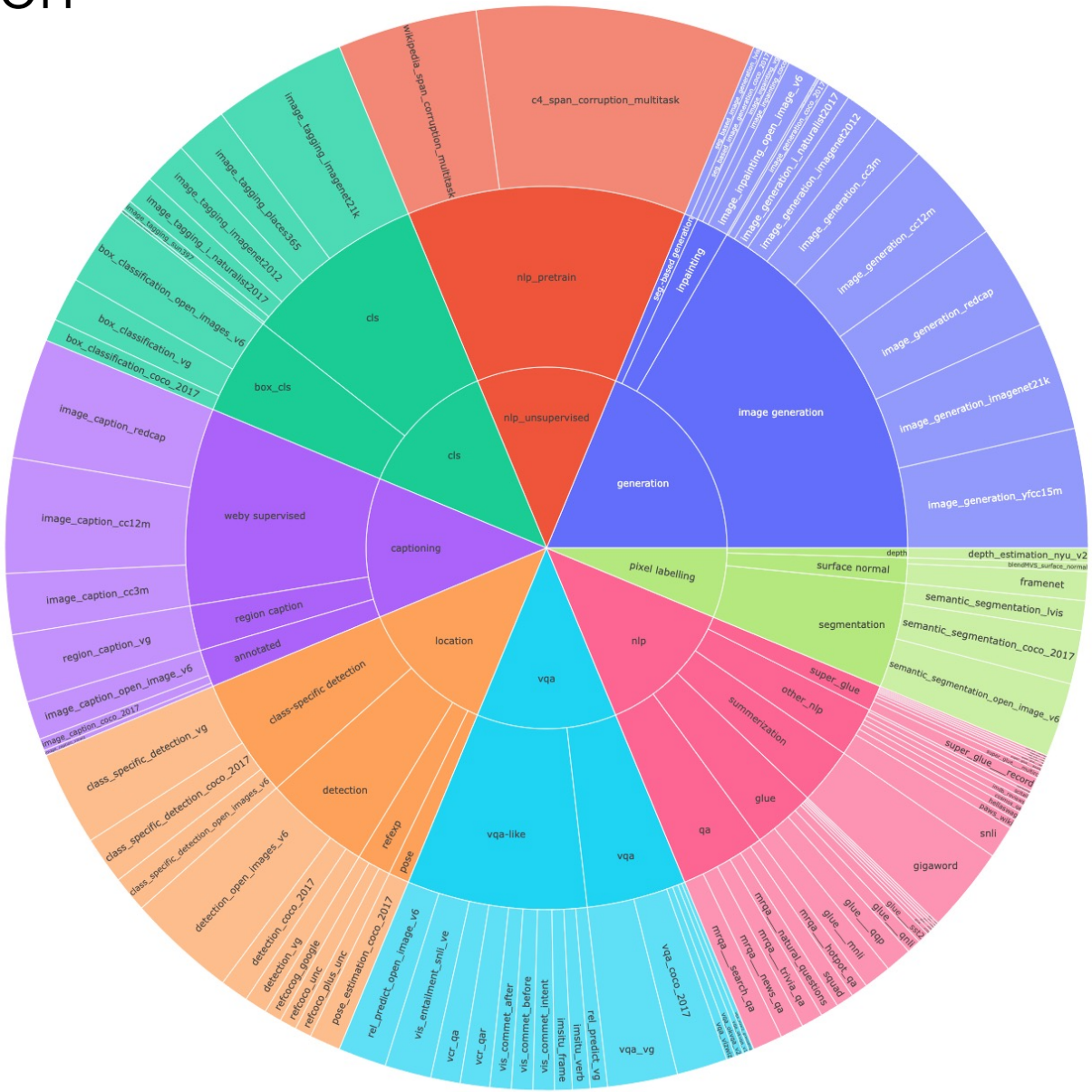
Tasks

	Example Source	Size				Input Modalities				Output Modalities			
		Datasets	Size	Percent	Rate	Text	Image	Sparse	Dense	Text	Image	Sparse	Dense
Image Synthesis		**14**	**56m**	**43.0**	**18.7**	✓	✓	✓	✓	-	✓	-	-
Image Synthesis from Text	*RedCaps*	9	55m	41.9	16.7	✓	-	-	-	-	✓	-	-
Image Inpainting	*VG*	3	1.2m	0.9	1.5	✓	✓	✓	-	-	✓	-	-
Image Synthesis from Seg.	*LVIS*	2	220k	0.2	0.6	✓	-	-	✓	-	✓	-	-
Sparse Labelling		**10**	**8.2m**	**6.3**	**12.5**	✓	✓	✓	-	-	-	✓	-
Object Detection	*Open Images*	3	1.9m	1.5	3.6	-	✓	-	-	-	-	✓	-
Object Localization	*VG*	3	6m	4.6	7.1	✓	✓	-	-	-	-	✓	-
Keypoint Estimation	*COCO*	1	140k	0.1	0.7	-	✓	✓	-	-	-	✓	-
Referring Expression	*RefCoco*	3	130k	0.1	1.1	✓	✓	-	-	-	-	✓	-
Dense Labelling		**6**	**2.4m**	**1.8**	**6.2**	✓	✓	-	-	-	-	-	✓
Depth Estimation	*NYU Depth*	1	48k	0.1	0.4	-	✓	-	-	-	-	-	✓
Surface Normal Estimation	*Framenet*	2	210k	0.2	1.1	-	✓	-	-	-	-	-	✓
Object Segmentation	*LVIS*	3	2.1m	1.6	4.7	✓	✓	-	-	-	-	-	✓
Image Classification		**9**	**22m**	**16.8**	**12.5**	-	✓	✓	-	✓	-	-	-
Image Classification	*ImageNet*	6	16m	12.2	8.1	✓	✓	-	-	✓	-	-	-
Object Categorization	*COCO*	3	6m	4.6	4.4	-	✓	✓	-	✓	-	-	-
Image Captioning		**7**	**31m**	**23.7**	**12.5**	-	✓	✓	-	✓	-	-	-
Webly Supervised Captioning	*CC12M*	3	26m	19.7	8.8	-	✓	-	-	✓	-	-	-
Supervised Captioning	*VizWiz*	3	1.4m	1.1	1.7	-	✓	-	-	✓	-	-	-
Region Captioning	*VG*	1	3.8m	2.9	2.0	-	✓	✓	-	✓	-	-	-
Vision & Language		**16**	**4m**	**3.0**	**12.5**	✓	✓	✓	-	✓	-	-	✓
Visual Question Answering	*VQA 2.0*	13	3.3m	2.5	10.4	✓	✓	✓	-	✓	-	-	-
Relationship Detection	*VG*	2	640k	0.5	1.9	-	✓	✓	-	✓	-	-	-
Grounded VQA	*VizWiz*	1	6.5k	0.1	0.1	✓	✓	-	-	✓	-	-	✓
NLP		**31**	**7.1m**	**5.4**	**12.5**	✓	-	-	-	✓	-	-	-
Text Classification	*MNLI*	17	1.6m	1.2	4.8	✓	-	-	-	✓	-	-	-
Question Answering	*SQuAD*	13	1.7m	1.3	5.2	✓	-	-	-	✓	-	-	-
Text Summarization	*Gigaword*	1	3.8m	2.9	2.5	✓	-	-	-	✓	-	-	-
Language Modelling		**2**	-	-	**12.5**	✓	-	-	-	✓	-	-	-
Masked Language Modelling	*C4*	2	-	-	12.5	✓	-	-	-	✓	-	-	-
All Tasks		**95**	**130m**	**100**	**100**	✓	✓	✓	✓	✓	✓	✓	✓

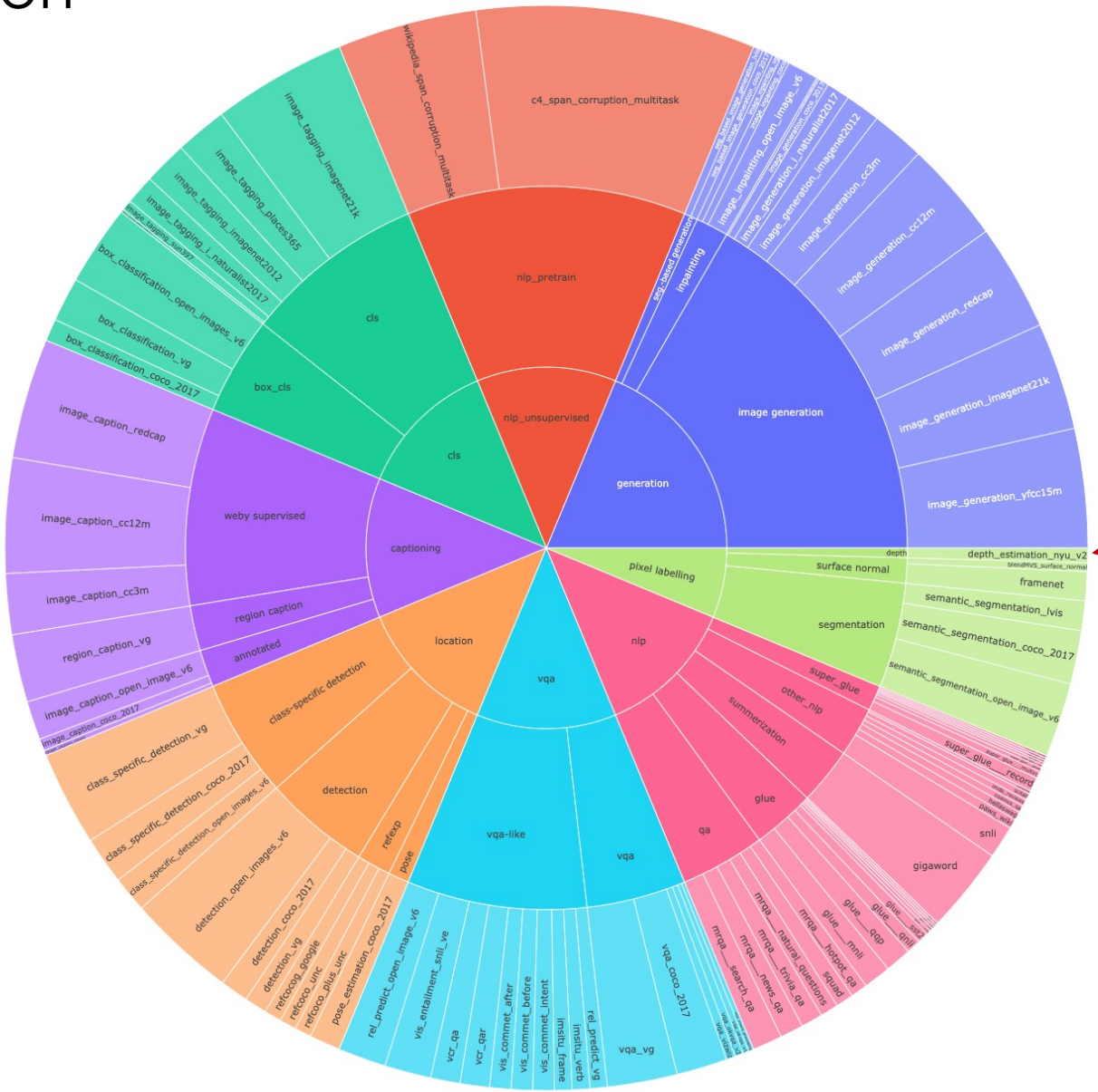
Pre-training Distribution



Tasks Distribution

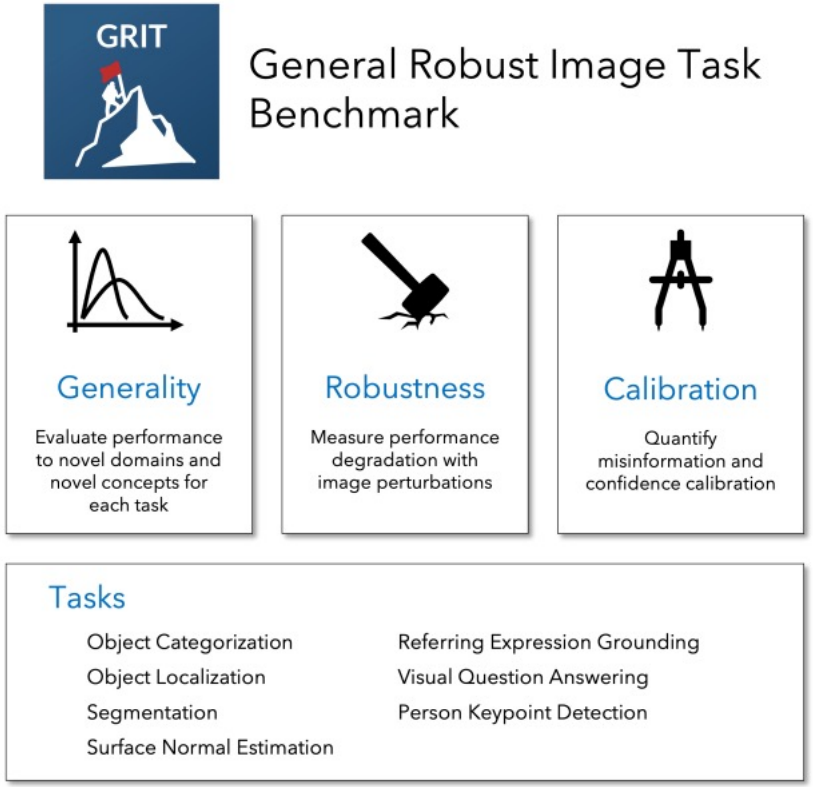


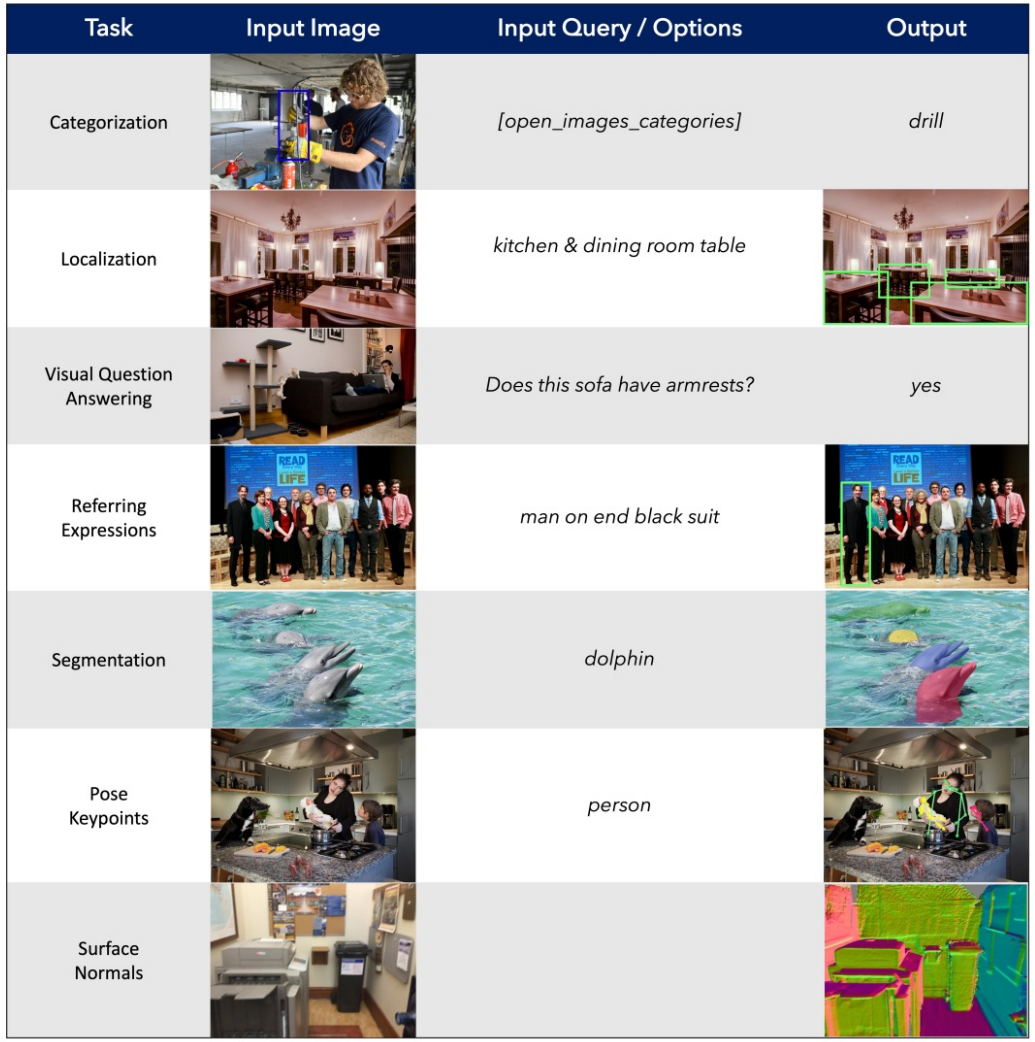
Tasks Distribution



Depth data

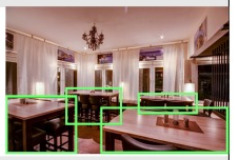
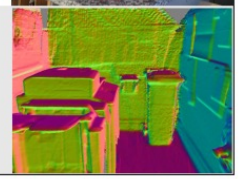
Evaluation





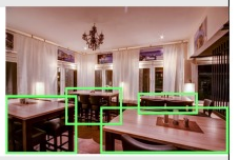
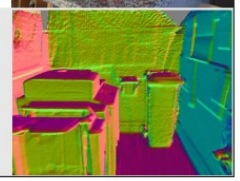
GRIT: General Robust Image Task Benchmark [Gupta et.al. 2022]

GRIT requires diverse skills

Task	Input Image	Input Query / Options	Output
Categorization		*[open_images_categories]*	*drill*
Localization		*kitchen & dining room table*	
Visual Question Answering		*Does this sofa have armrests?*	*yes*
Referring Expressions		*man on end black suit*	
Segmentation		*dolphin*	
Pose Keypoints		*person*	
Surface Normals			

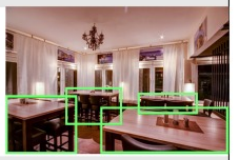
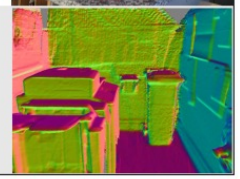
Inputs	Outputs
Bounding box	*Class name*
Class name	*Bounding box*
Questions	*Answers*
Phrases	*Bounding box*
Class name	*Segmentation*
	Joints + visibility
	Surface normal

GRIT requires diverse skills

Task	Input Image	Input Query / Options	Output
Categorization		*[open_images_categories]*	*drill*
Localization		*kitchen & dining room table*	
Visual Question Answering		*Does this sofa have armrests?*	*yes*
Referring Expressions		*man on end black suit*	
Segmentation		*dolphin*	
Pose Keypoints		*person*	
Surface Normals			

Inputs	Outputs
Bounding box	*Class name*
Class name	*Bounding box*
Questions	*Answers*
Phrases	*Bounding box*
Class name	*Segmentation*
	Joints + visibility
	Surface normal

GRIT requires diverse skills

Task	Input Image	Input Query / Options	Output	Inputs	Outputs
Categorization		*[open_images_categories]*	*drill*	*Bounding box*	*Class name*
Localization		*kitchen & dining room table*		*Class name*	*Bounding box*
Visual Question Answering		*Does this sofa have armrests?*	*yes*	*Questions*	*Answers*
Referring Expressions		*man on end black suit*		*Phrases*	*Bounding box*
Segmentation		*dolphin*		*Class name*	*Segmentation*
Pose Keypoints		*person*			*Joints + visibility*
Surface Normals					*Surface normal*

Results

		Categorization		Localization		VQA		Refexp		Segmentation		Keypoint		Normal		All	
		ablation	test	ablation	test	ablation	test	ablation	test	ablation	test	ablation	test	ablation	test	ablation	test
0	NLL-AngMF [3]	-	-	-	-	-	-	-	-	-	-	-	-	**49.6**	**50.5**	7.2	7.1
1	Mask R-CNN [29]	-	-	44.7	45.1	-	-	-	-	26.2	26.2	**70.8**	**70.6**	-	-	20.2	20.3
2	GPV-1 [26]	33.2	33.2	42.8	42.7	50.6	49.8	25.8	26.8	-	-	-	-	-	-	21.8	21.8
3	CLIP [56]	48.1	-	-	-	-	-	-	-	-	-	-	-	-	-	6.9	-
4	$\text{OFA}_{\text{LARGE}}$ [73]	22.6	-	-	-	72.4	-	61.7	-	-	-	-	-	-	-	22.4	-
5	GPV-2 [36]	54.7	55.1	53.6	53.6	63.5	63.2	51.5	52.1	-	-	-	-	-	-	31.9	32.0
6	$\text{UNIFIED-IO}_{\text{SMALL}}$	42.6	-	50.4	-	52.9	-	51.1	-	40.7	-	46.5	-	33.5	-	45.4	-
7	$\text{UNIFIED-IO}_{\text{BASE}}$	53.1	-	59.7	-	63.0	-	68.3	-	49.3	-	60.2	-	37.5	-	55.9	-
8	$\text{UNIFIED-IO}_{\text{LARGE}}$	57.0	-	64.2	-	67.4	-	74.1	-	54.0	-	67.6	-	40.2	-	57.0	-
9	$\text{UNIFIED-IO}_{\text{XL}}$	**61.7**	**60.8**	**67.0**	**67.1**	**74.5**	**74.5**	**78.6**	**78.9**	**56.3**	**56.5**	68.1	67.7	45.0	44.3	**64.5**	**64.3**

Unified IO on GRIT ablation and test benchmark

Results

		Categorization		Localization		VQA		Refexp		Segmentation		Keypoint		Normal		All	
		ablation	test	ablation	test	ablation	test	ablation	test	ablation	test	ablation	test	ablation	test	ablation	test
0	NLL-AngMF [3]	-	-	-	-	-	-	-	-	-	-	-	-	**49.6**	**50.5**	7.2	7.1
1	Mask R-CNN [29]	-	-	44.7	45.1	-	-	-	-	26.2	26.2	**70.8**	**70.6**	-	-	20.2	20.3
2	GPV-1 [26]	33.2	33.2	42.8	42.7	50.6	49.8	25.8	26.8	-	-	-	-	-	-	21.8	21.8
3	CLIP [56]	48.1	-	-	-	-	-	-	-	-	-	-	-	-	-	6.9	-
4	OFA$_{\text{LARGE}}$ [73]	22.6	-	-	-	72.4	-	61.7	-	-	-	-	-	-	-	22.4	-
5	GPV-2 [36]	54.7	55.1	53.6	53.6	63.5	63.2	51.5	52.1	-	-	-	-	-	-	31.9	32.0
6	UNIFIED-IO$_{\text{SMALL}}$	42.6	-	50.4	-	52.9	-	51.1	-	40.7	-	46.5	-	33.5	-	45.4	-
7	UNIFIED-IO$_{\text{BASE}}$	53.1	-	59.7	-	63.0	-	68.3	-	49.3	-	60.2	-	37.5	-	55.9	-
8	UNIFIED-IO$_{\text{LARGE}}$	57.0	-	64.2	-	67.4	-	74.1	-	54.0	-	67.6	-	40.2	-	57.0	-
9	UNIFIED-IO$_{\text{XL}}$	**61.7**	**60.8**	**67.0**	**67.1**	**74.5**	**74.5**	**78.6**	**78.9**	**56.3**	**56.5**	68.1	67.7	45.0	44.3	**64.5**	**64.3**

Unified IO on GRIT ablation and test benchmark

Results

		Categorization		Localization		VQA		Refexp		Segmentation		Keypoint		Normal		All	
		ablation	test	ablation	test	ablation	test	ablation	test	ablation	test	ablation	test	ablation	test	ablation	test
0	NLL-AngMF [3]	-	-	-	-	-	-	-	-	-	-	-	-	**49.6**	**50.5**	7.2	7.1
1	Mask R-CNN [29]	-	-	44.7	45.1	-	-	-	-	26.2	26.2	**70.8**	**70.6**	-	-	20.2	20.3
2	GPV-1 [26]	33.2	33.2	42.8	42.7	50.6	49.8	25.8	26.8	-	-	-	-	-	-	21.8	21.8
3	CLIP [56]	48.1	-	-	-	-	-	-	-	-	-	-	-	-	-	6.9	-
4	OFA$_{\text{LARGE}}$ [73]	22.6	-	-	-	72.4	-	61.7	-	-	-	-	-	-	-	22.4	-
5	GPV-2 [36]	54.7	55.1	53.6	53.6	63.5	63.2	51.5	52.1	-	-	-	-	-	-	31.9	32.0
6	UNIFIED-IO$_{\text{SMALL}}$	42.6	-	50.4	-	52.9	-	51.1	-	40.7	-	46.5	-	33.5	-	45.4	-
7	UNIFIED-IO$_{\text{BASE}}$	53.1	-	59.7	-	63.0	-	68.3	-	49.3	-	60.2	-	37.5	-	55.9	-
8	UNIFIED-IO$_{\text{LARGE}}$	57.0	-	64.2	-	67.4	-	74.1	-	54.0	-	67.6	-	40.2	-	57.0	-
9	UNIFIED-IO$_{\text{XL}}$	**61.7**	**60.8**	**67.0**	**67.1**	**74.5**	**74.5**	**78.6**	**78.9**	**56.3**	**56.5**	68.1	67.7	45.0	44.3	**64.5**	**64.3**

Unified IO on GRIT ablation and test benchmark

Results

		Categorization		Localization		VQA		Refexp		Segmentation		Keypoint		Normal		All	
		ablation	test	ablation	test	ablation	test	ablation	test	ablation	test	ablation	test	ablation	test	ablation	test
0	NLL-AngMF [3]	-	-	-	-	-	-	-	-	-	-	-	-	**49.6**	**50.5**	7.2	7.1
1	Mask R-CNN [29]	-	-	44.7	45.1	-	-	-	-	26.2	26.2	**70.8**	**70.6**	-	-	20.2	20.3
2	GPV-1 [26]	33.2	33.2	42.8	42.7	50.6	49.8	25.8	26.8	-	-	-	-	-	-	21.8	21.8
3	CLIP [56]	48.1	-	-	-	-	-	-	-	-	-	-	-	-	-	6.9	-
4	OFA$_{\text{LARGE}}$ [73]	22.6	-	-	-	72.4	-	61.7	-	-	-	-	-	-	-	22.4	-
5	GPV-2 [36]	54.7	55.1	53.6	53.6	63.5	63.2	51.5	52.1	-	-	-	-	-	-	31.9	32.0
6	UNIFIED-IO$_{\text{SMALL}}$	42.6	-	50.4	-	52.9	-	51.1	-	40.7	-	46.5	-	33.5	-	45.4	-
7	UNIFIED-IO$_{\text{BASE}}$	53.1	-	59.7	-	63.0	-	68.3	-	49.3	-	60.2	-	37.5	-	55.9	-
8	UNIFIED-IO$_{\text{LARGE}}$	57.0	-	64.2	-	67.4	-	74.1	-	54.0	-	67.6	-	40.2	-	57.0	-
9	UNIFIED-IO$_{\text{XL}}$	**61.7**	**60.8**	**67.0**	**67.1**	**74.5**	**74.5**	**78.6**	**78.9**	**56.3**	**56.5**	68.1	67.7	45.0	44.3	**64.5**	**64.3**

Unified IO on GRIT ablation and test benchmark

Results

		Categorization		Localization		VQA		Refexp		Segmentation		Keypoint		Normal		All	
		ablation	test	ablation	test	ablation	test	ablation	test	ablation	test	ablation	test	ablation	test	ablation	test
0	NLL-AngMF [3]	-	-	-	-	-	-	-	-	-	-	-	-	**49.6**	**50.5**	7.2	7.1
1	Mask R-CNN [29]	-	-	44.7	45.1	-	-	-	-	26.2	26.2	**70.8**	**70.6**	-	-	20.2	20.3
2	GPV-1 [26]	33.2	33.2	42.8	42.7	50.6	49.8	25.8	26.8	-	-	-	-	-	-	21.8	21.8
3	CLIP [56]	48.1	-	-	-	-	-	-	-	-	-	-	-	-	-	6.9	-
4	OFA$_{\text{LARGE}}$ [73]	22.6	-	-	-	72.4	-	61.7	-	-	-	-	-	-	-	22.4	-
5	GPV-2 [36]	54.7	55.1	53.6	53.6	63.5	63.2	51.5	52.1	-	-	-	-	-	-	31.9	32.0
6	UNIFIED-IO$_{\text{SMALL}}$	42.6	-	50.4	-	52.9	-	51.1	-	40.7	-	46.5	-	33.5	-	45.4	-
7	UNIFIED-IO$_{\text{BASE}}$	53.1	-	59.7	-	63.0	-	68.3	-	49.3	-	60.2	-	37.5	-	55.9	-
8	UNIFIED-IO$_{\text{LARGE}}$	57.0	-	64.2	-	67.4	-	74.1	-	54.0	-	67.6	-	40.2	-	57.0	-
9	UNIFIED-IO$_{\text{XL}}$	**61.7**	**60.8**	**67.0**	**67.1**	**74.5**	**74.5**	**78.6**	**78.9**	**56.3**	**56.5**	68.1	67.7	45.0	44.3	**64.5**	**64.3**

Unified IO on GRIT ablation and test benchmark

Results

		Categorization		Localization		VQA		Refexp		Segmentation		Keypoint		Normal		All	
		ablation	test	ablation	test	ablation	test	ablation	test	ablation	test	ablation	test	ablation	test	ablation	test
0	NLL-AngMF [3]	-	-	-	-	-	-	-	-	-	-	-	-	**49.6**	**50.5**	7.2	7.1
1	Mask R-CNN [29]	-	-	44.7	45.1	-	-	-	-	26.2	26.2	**70.8**	**70.6**	-	-	20.2	20.3
2	GPV-1 [26]	33.2	33.2	42.8	42.7	50.6	49.8	25.8	26.8	-	-	-	-	-	-	21.8	21.8
3	CLIP [56]	48.1	-	-	-	-	-	-	-	-	-	-	-	-	-	6.9	-
4	OFA$_{\text{LARGE}}$ [73]	22.6	-	-	-	72.4	-	61.7	-	-	-	-	-	-	-	22.4	-
5	GPV-2 [36]	54.7	55.1	53.6	53.6	63.5	63.2	51.5	52.1	-	-	-	-	-	-	31.9	32.0
6	UNIFIED-IO$_{\text{SMALL}}$	42.6	-	50.4	-	52.9	-	51.1	-	40.7	-	46.5	-	33.5	-	45.4	-
7	UNIFIED-IO$_{\text{BASE}}$	53.1	-	59.7	-	63.0	-	68.3	-	49.3	-	60.2	-	37.5	-	55.9	-
8	UNIFIED-IO$_{\text{LARGE}}$	57.0	-	64.2	-	67.4	-	74.1	-	54.0	-	67.6	-	40.2	-	57.0	-
9	UNIFIED-IO$_{\text{XL}}$	**61.7**	**60.8**	**67.0**	**67.1**	**74.5**	**74.5**	**78.6**	**78.9**	**56.3**	**56.5**	68.1	67.7	45.0	44.3	**64.5**	**64.3**

Unified IO on GRIT ablation and test benchmark

Results

	NYUv2	*ImageNet*	*Place365*	*VQAv2*	*OkVQA*	*A-OkVQA*	*VizWizQA*	*VizWizGround*	*Swig*	*SNLI-VE*	*VisComet*	*Nocaps*	*COCO*	*COCO*	*MRPC*	*BoolQ*	*SciTail*
Split	val	val	val	test-dev	test	test	test-dev	test-std	test	val	val	val	val	test	val	val	test
Metric	RMSE	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	IOU	Acc.	Acc.	CIDEr	CIDEr	CIDEr	CIDEr	F1	Acc	Acc
Unified SOTA	UViM	-	-	-	Flamingo	-	Flamingo	-	-	-	-	-	-	-	T5	PaLM	-
	0.467	-	-	-	57.8	-	49.8	-	-	-	-	-	-	-	92.20	92.2	-
UNIFIED-IO$_{\text{SMALL}}$	0.649	42.8	38.2	57.7	31.0	24.3	42.4	35.5	17.3	76.5	-	45.1	80.1	-	84.9	65.9	87.4
UNIFIED-IO$_{\text{BASE}}$	0.469	63.3	43.2	61.8	37.8	28.5	45.8	50.0	29.7	85.6	-	66.9	104.0	-	87.9	70.8	90.8
UNIFIED-IO$_{\text{LARGE}}$	0.402	71.8	50.5	67.8	42.7	33.4	47.7	54.7	40.4	86.1	-	87.2	117.5	-	87.5	73.1	93.1
UNIFIED-IO$_{\text{XL}}$	0.385	79.1	53.2	77.9	54.0	45.2	57.4	65.0	49.8	91.1	21.2	100.0	126.8	122.3	89.2	79.7	95.7
Single or fine-tuned SOTA	BinsFormer	CoCa	MAE	CoCa	KAT	GPV2	Flamingo	MAC-Caps	JSL	OFA	SVT	CoCa	-	OFA	Turing NLR	ST-MOE	DeBERTa
	0.330	91.00	60.3	82.3	54.4	38.1	65.7	27.3	39.6	91.0	18.3	122.4	-	145.3	93.8	92.4	97.7

Unified IO on other tasks

Visualization – image synthesis

IMAGE SYNTHESIS

What is the complete image? Text

INPUT

small personal pizza with bacon and spinach

many large kites flying in the sky

the train is on the tracks in the station

a beach area with black birds flying over it

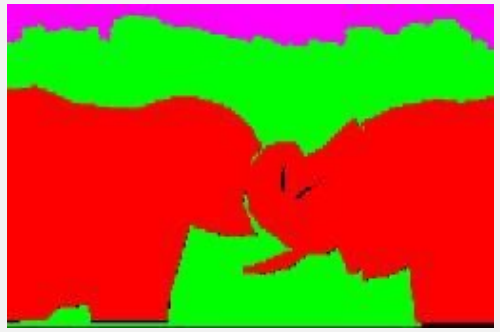
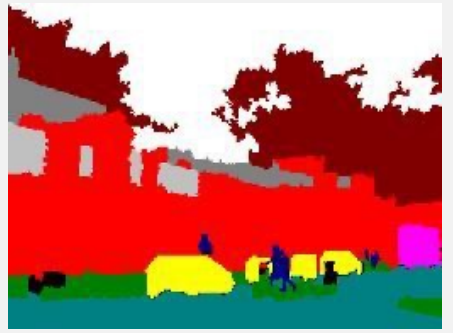
PREDICTION

Visualization – image synthesis

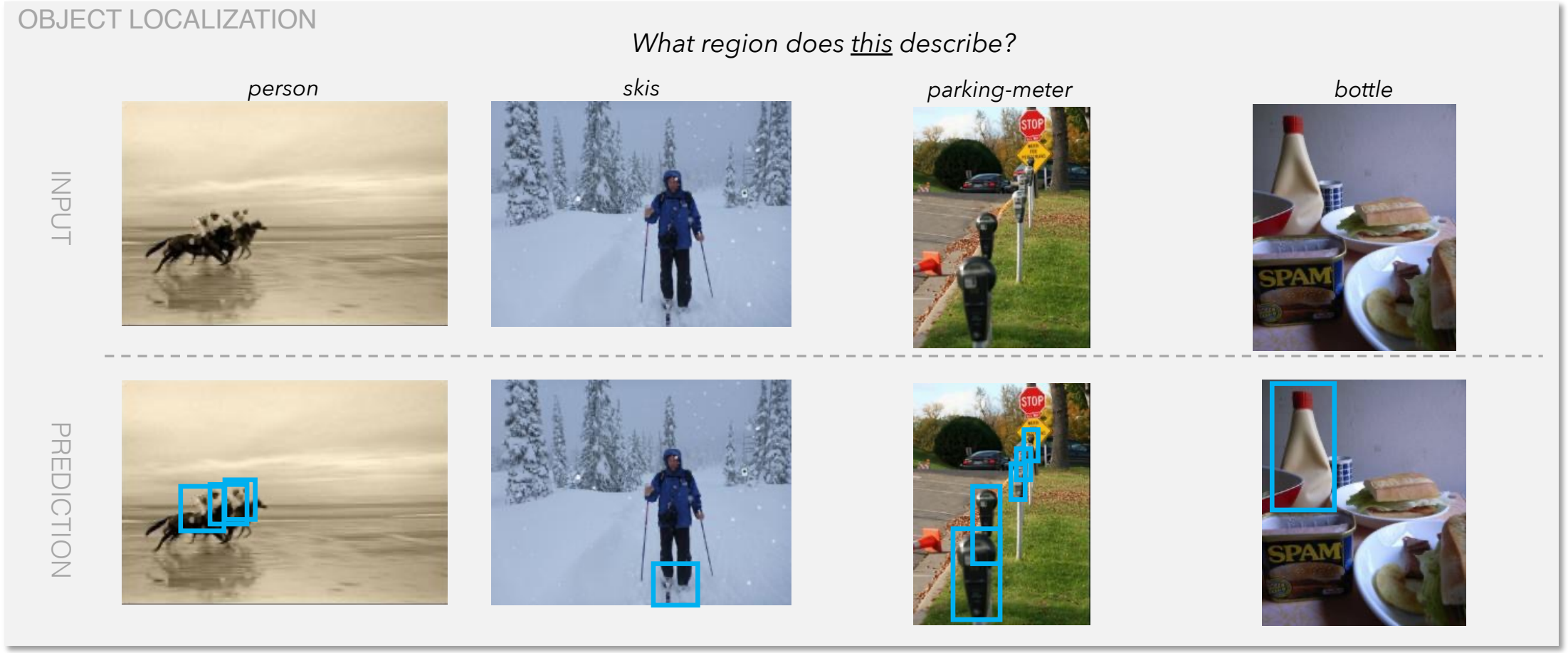
Visualization – image synthesis

IMAGE GENERATION FROM SEGMENTATION

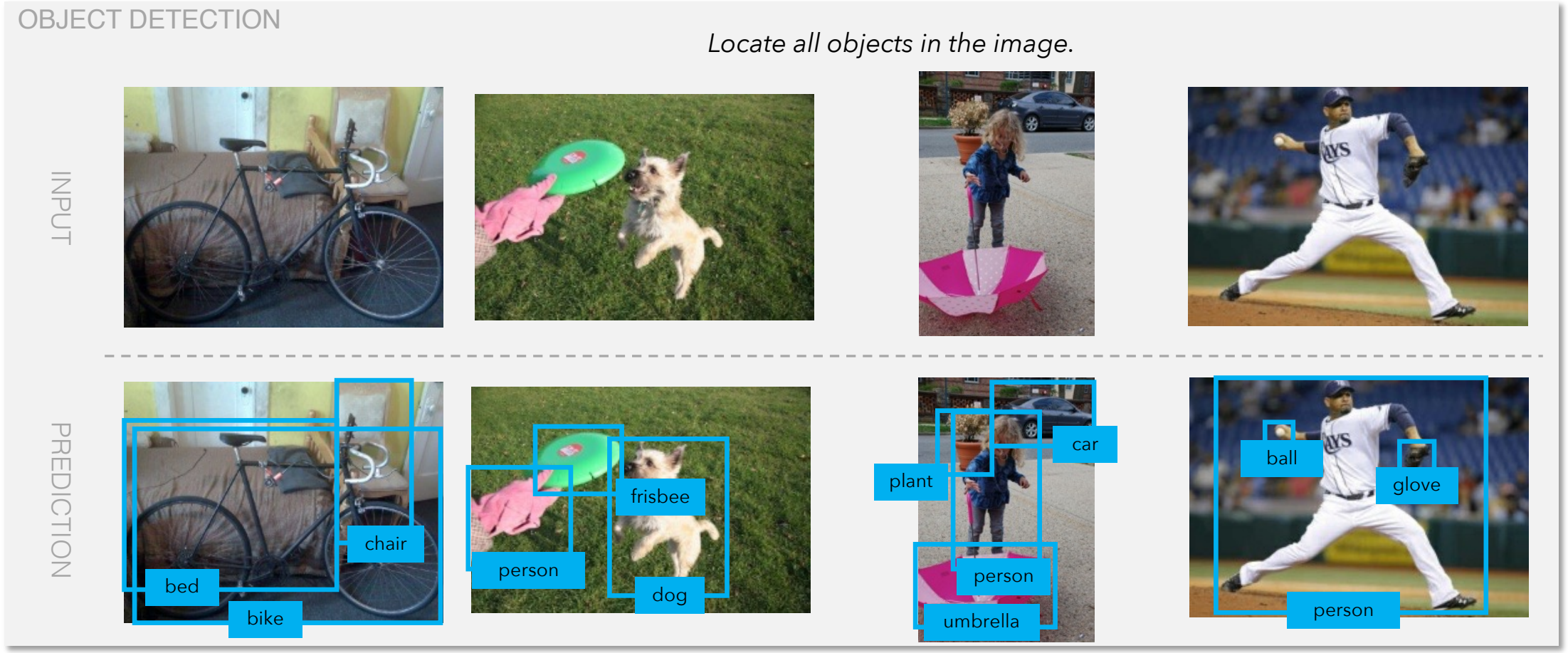
What is the complete image? Segmentation color

	fuchsia: tree, lime: dirt….	*yellow: field, aqua: baseball…*	*lime: building, navy: wall…*	*white: tree, red: building…*
INPUT				
PREDICTION				

Visualization – sparse labeling

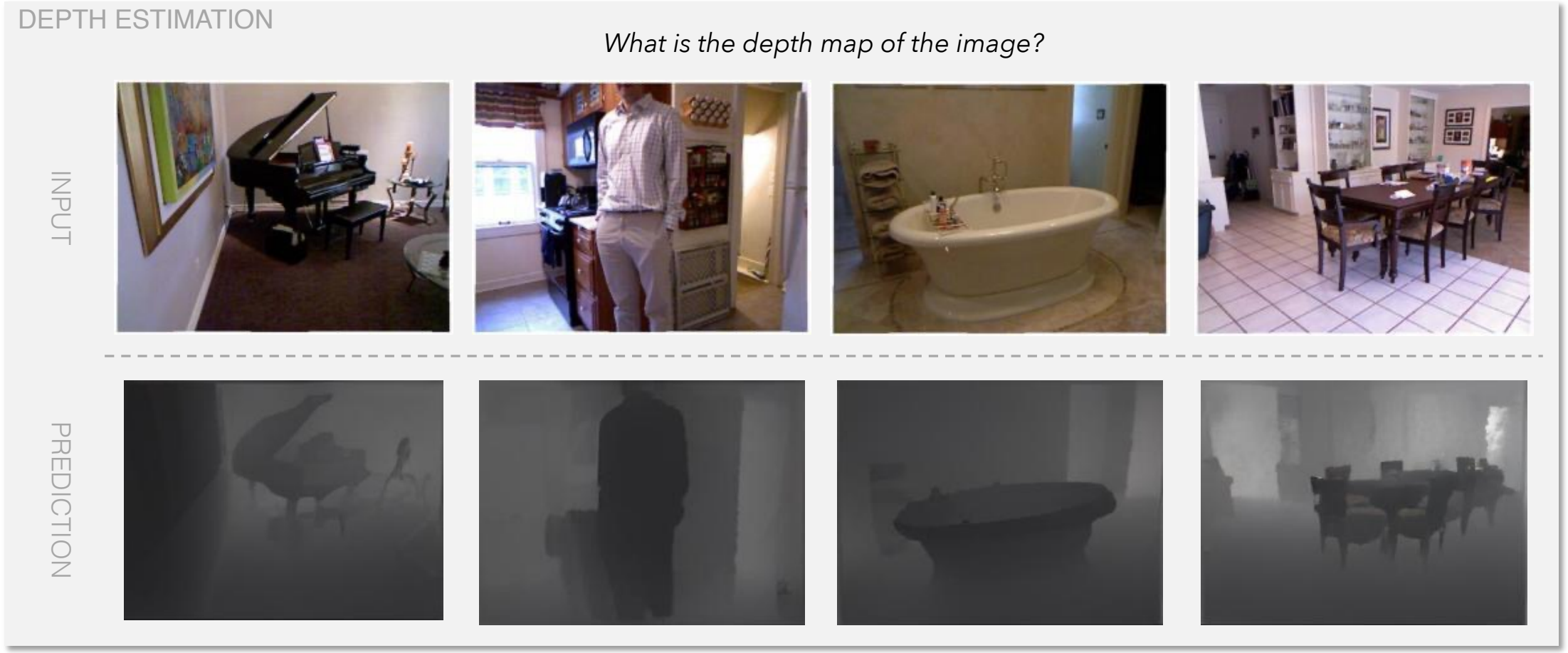


Visualization – sparse labeling



Visualization – sparse labeling

Visualization – dense labeling

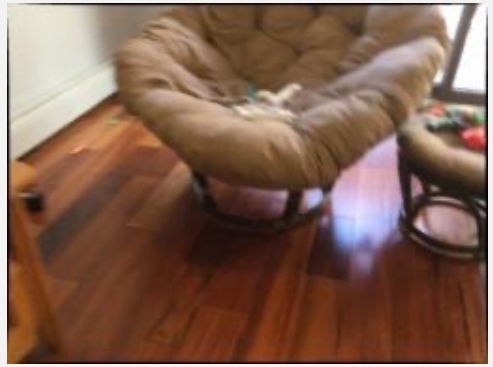


Visualization – dense labeling

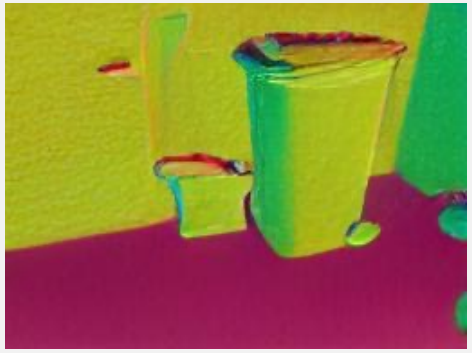
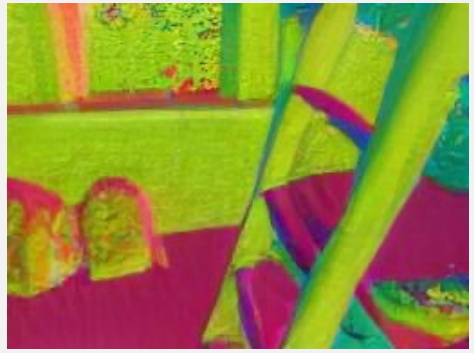
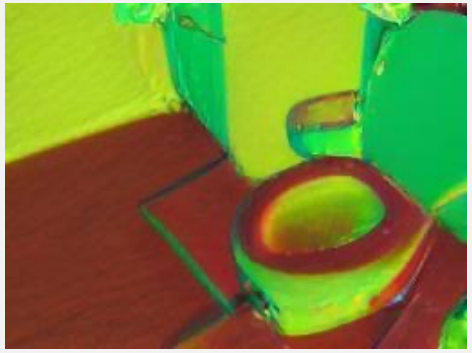
SURFACE NORMAL ESTIMATION

What is the surface normal of the image?

INPUT




PREDICTION

Visualization – dense labeling

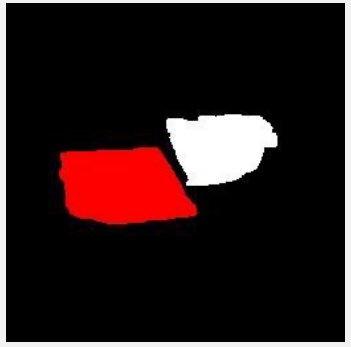
OBJECT SEGMENTATION

What is the segmentation of this?

INPUT

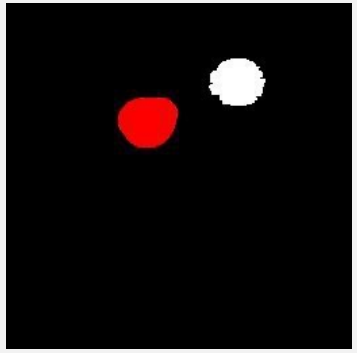
PREDICTION

pizza



bed

apple



grass

Referring expressions using different prompts

	Prompt	Refexp Score
0	Which region does the text " REFEXP " describe ?	78.9
1	Which region does the text "REFEXP" describe?	76.7
2	Which region matches the text " REFEXP " ?	77.4
3	Locate the " REFEXP " .	65.6
4	Which region can be described as " REFEXP " ?	64.8
5	Locate the region described by " REFEXP " .	43.2
6	Where is the " REFEXP " ?	41.5
7	Where is the "REFEXP"?	0.1

Summary

- Propose unified IO which is the first framework that can handle massive vision, vision – language, and language tasks.
- We treat 2D image tasks as condition image generation tasks.
- We use pre-trained VQ-GAN to convert images into discrete sequences.
- We will release the Code + pre-trained model.

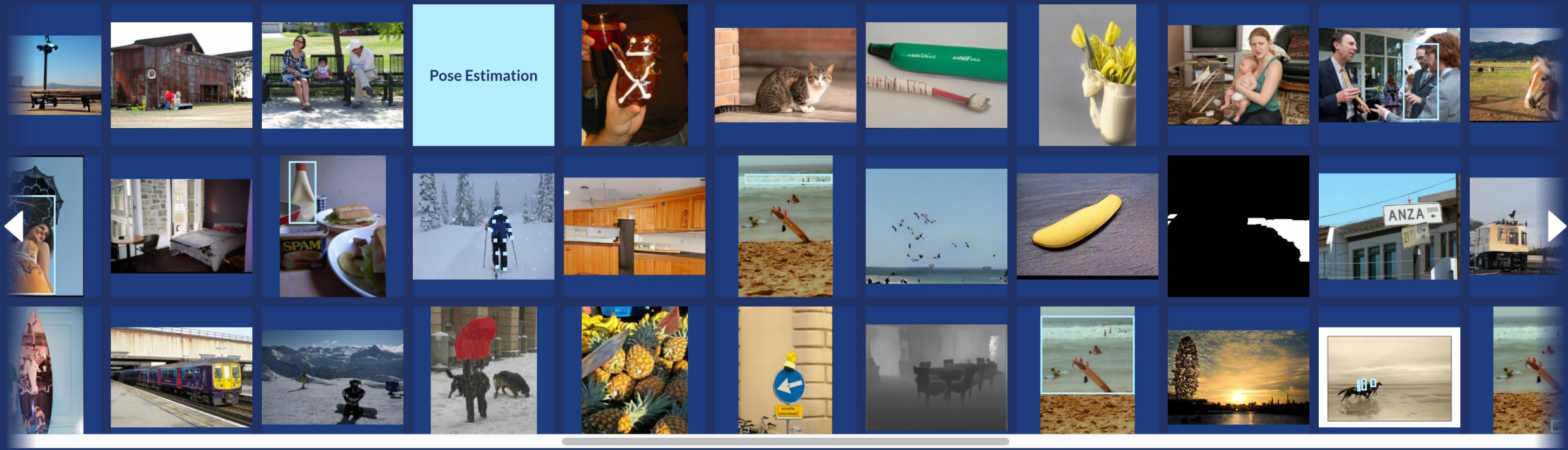
Acknowledgement

- This research was made possible with cloud TPUs from Google's TPU Research Cloud (TRC).
- Google Jax and T5X team -- amazing libraries.
- AI2 Reviz Team – Great demo pages.

Unified-IO

Share to:

A new general-purpose model with **unprecedented breadth**, Unified-IO can perform a wide array of **visual** and **linguistic** tasks.



https://unified-io.allenai.org

AI2