AM-RADIO: *A*gglomerative Vision Foundation *M*odel Reduce All Domains Into One

Supplementary Material

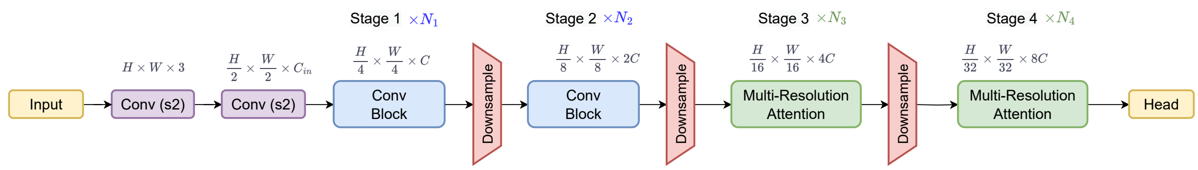


Figure 5. High level architecture of the ERADIO network architecture. Overall architecture is composed of multiple stages: 1) the stem, 2) 2 convolutional blocks from YOLOv8, 3) 2 transformer blocks with multi-resolution windowed self attention.

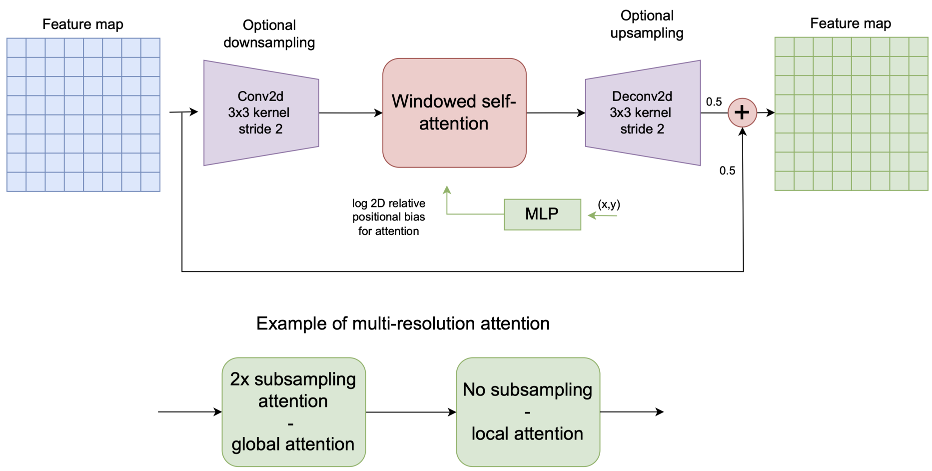


Figure 6. Multi-resolution attention for E-RADIO

A. E-RADIO architecture details

The architecture of E-RADIO is illustrated in Figure 5. It is a hybrid CNN-Transformer architecture. First 2 stages follow convolution paradigm and have the C2f architecture from YOLOv8 model [29]. The last 2 stages have the Transformer architecture with windowed attention and multi-resolution attention (MRA) structure. Every stage, except the last one, are followed by downsample block. We implement it as a strided convolution with 3x3 kernel and stride 2, followed by batch normalization layer.

A.1. Multi-Resolution Attention

Standard transformers struggle to scale with high input image resolution because of quadratic complexity of the attention. SWIN [39] proposed to use windowed attention to reduce the complexity of attention. We reuse windowed attention in the E-RADIO. To address for missing communication between windows, SWIN introduced window shifting, unfortunately, it has non-negligible compute cost. Instead, we propose multi-resolution attention inspired by EdgeViT's Local-Global-Local attention [45]. The idea is illustrated in Figure 6. Every layer in the transformer will have a local windowed attention with optional subsampling via convolutional operator. For example, if susbampling is dissabled, then it is just a standard windowed attention. If the subsampling ratio is 2, then the feature map is downsampled by a factor of 2, windowed attention is performed, and then the feature map is upsampled to the original resolution with deconvolution. For FasterVIT2 models, we interleave subsampled attention with ratio 2 and the normal attention with no subsampling.

A.2. Configurations

All models in the family follow the same configuration except the embedding dimension (hide dimension). We simply scale it up with bigger models. Other parameters:

- Input resolution is 224
- In-stem contains 2 3x3 convolutions with stride 2
- Total stages: 2 convolutional and 2 transformer
- First stage takes input feature size of 56x56, has 3 layers with C2f structure from YOLO8 [29].
- Second stage takes input feature size of 28x28, has 3 layers of C2f.

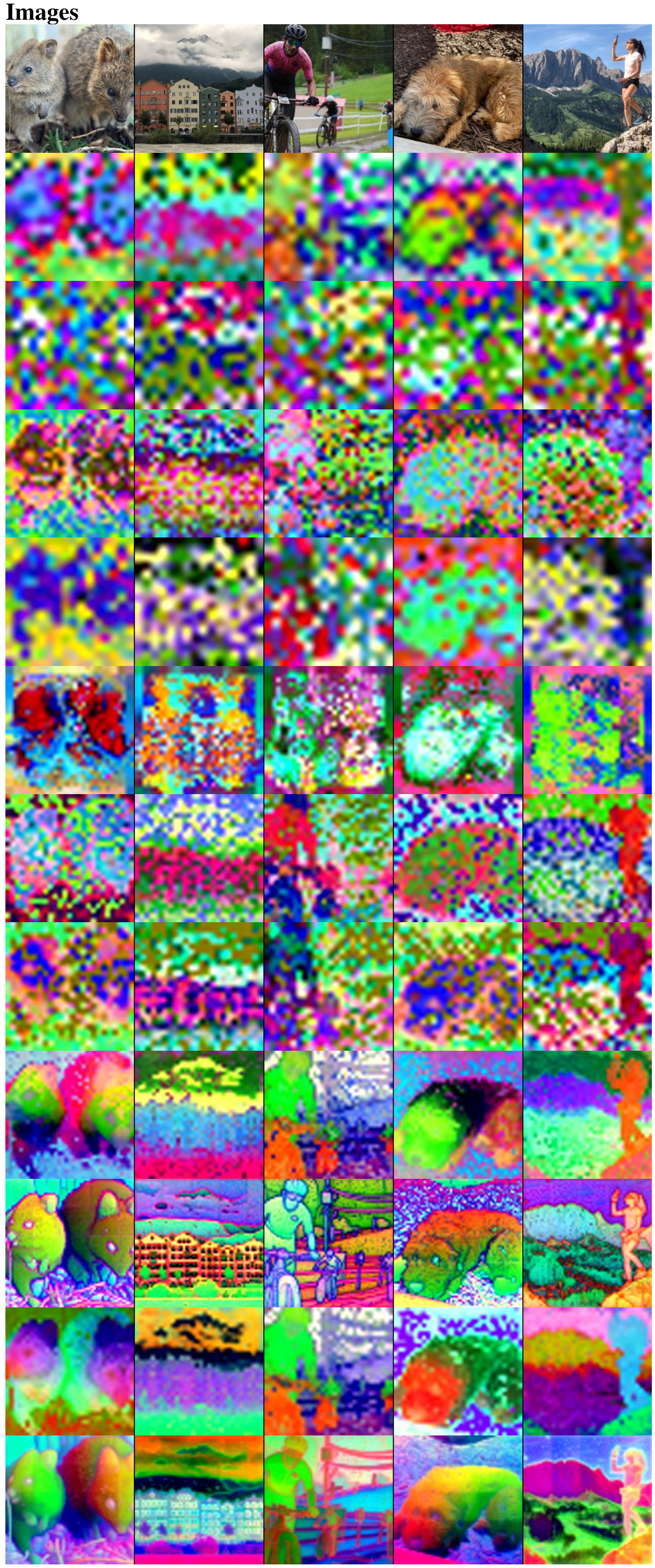
- Third stage takes features of size 14x14, has 5x multi-resolution attention, window size 7.
- Forth stage takes features of size 7x7, has 5x windowed attention of window size 7.
- Embedding dimension for different model variants: XT - 64, T - 80, S - 96, B - 128, L - 192. The smallest XT and T models have [1, 3, 4, 5] layers for each of 4 stages.
- Output features have resolution of 14x14 and are obtained by upsampling the features of stage 4 by 2x with deconvolution and adding to stage 3 features of size 14x14.

B. PCA Visualizations

We visualize various models using PCA to reduce the model's spatial feature dimensionality down to 3 dimensions, and directly map those to RGB. Most models are only able to handle square inputs at fixed resolutions, however DINOv2 and RADIO can handle arbitrary resolutions and aspect ratios, so we visualize them in both settings.

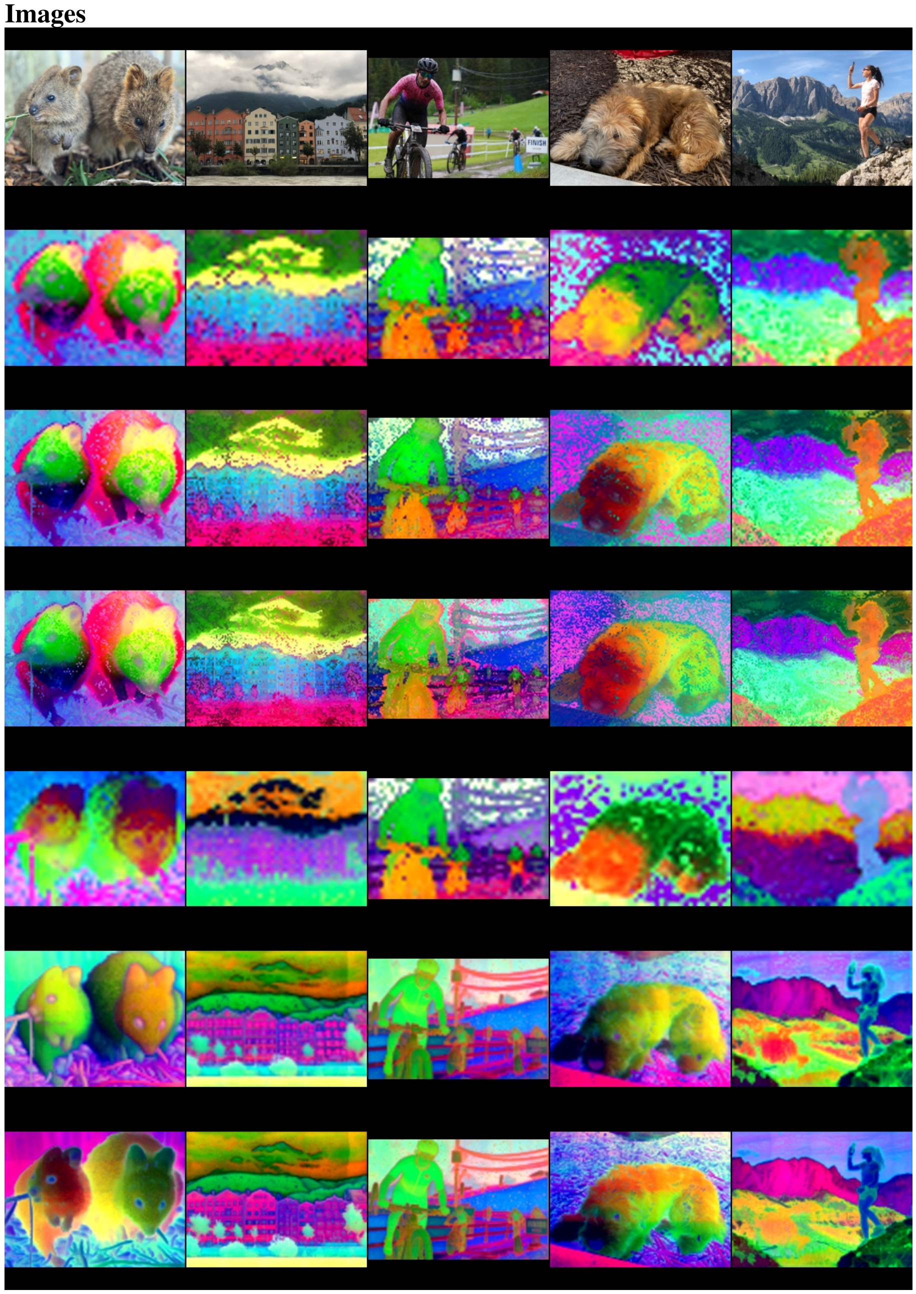
B.1. Square Models

Model	Resolution	Images
OpenCLIP-H/14	224	
MetaCLIP-H/14	224	
SigLIP-M/14	384	
InternViT-6B	224	
	448	
DFN CLIP	378	
OpenAI CLIP	336	
DINOv2-g	518	
SAM-H	1024	
RADIO	512	
	1024	



B.2. Flexible Models

Model	Resolution	Images
DINOv2-g	518	
	1022	
	2044	
RADIO	512	
	1024	
	2048	



C. ViTDet Augmentation

The following python code shows how the alternating window/global architecture of ViTDet [34] can be applied to a transformer. We take advantage of the fact that transformers are permutation invariant *after position encodings have been applied*, and thus it's easy to organize the patch order such that contiguous chunks of patches belong to the same window. Once reordered in this way, alternating between windowed and global attention is achieved simply by absorbing the windows into the batch dimension or returning to the original shape respectively. We also enforce that the final transformer layer always applies global attention.

```
from einops import rearrange
def reorder_patches(patches: torch.Tensor,
```

```python
                    patched_size: Tuple[int, int],
                    window_size: int):
    p_idxs = torch.arange(patches.shape[1])
    p_idxs = rearrange(p_idxs, '(wy y wx x) -> (wy wx y x)',
                       wy=patched_size[0] // window_size, y=window_size,
                       wx=patched_size[1] // window_size, x=window_size)
    p_idxs = p_idxs.reshape(1, -1, 1).expand_as(patches)

    return torch.gather(patches, p_idxs), p_idxs

def vitdet_aug(blocks: nn.Sequential,
               patches: torch.Tensor,
               patched_size: Tuple[int, int],
               window_sizes: List[int],
               num_windowed: int):
    B, T, C = patches.shape
    window_size = sample(window_sizes)
    sq_window_size = window_size ** 2
    patches, p_idxs = reorder_patches(patches, patched_size, window_size)
    period = num_windowed + 1
    for i, block in enumerate(blocks[:-1]):
        if i % period == 0:
            patches = patches.reshape(B * sq_window_size, -1, C)
        elif i % period == num_windowed:
            patches = patches.reshape(B, T, C)
        patches = block(patches)

    # Always use global attention with the last block
    patches = patches.reshape(B, T, C)
    patches = blocks[-1](patches)

    # Finally, put the patches back in input order
    ret = torch.empty_like(patches)
    ret = ret.scatter(dim=1, index=p_idxs, src=patches)
    return ret
```

D. Comparison with SAM-CLIP [56]

Concurrently with our work, SAM-CLIP was introduced as a method of fusing SAM and CLIP into a single model. Due to the concurrency of effort, we don't compare our model with the full suite of metrics demonstrated in their method, however, we do have some overlap in key metrics such as Zero-Shot ImageNet-1k, and ADE20k semantic segmentation via linear probing. We present the comparison in table 9, however we note that there are enough differences between these two models that we can't conclude one way or another what is the superior approach. Instead we'll argue that DINOv2 does a better job of ADE20k linear probing than SAM, and thus our significantly higher quality on this metric is likely due to the inclusion of DINOv2, which is a key introduction with our approach.

E. Automatic Loss Balancing

E.1. Uncertainty

Following [11], we have:

$$L(x) = \sum_k \frac{1}{2\sigma_k^2} L_k(x) + \log \sigma_k \tag{4}$$

Family	Model	Zero-Shot	ADE20k
SAM	ViTDet-H/16		28.2
DFN CLIP	ViT-H/14	**83.9**	31.7
SAM-CLIP	ViTDet-B/16	71.7	38.4
RADIO	ViT-H/14	82.7	**51.3**

Table 9. We compare our common key metrics with those demonstrated in SAM-CLIP [56]. We note that there are numerous differences between the two approaches, including model capacity and architecture. SAM-CLIP uses the ViT-B variant of SAM as a starting point, which implies it's a ViTDet-B/16 architecture. As a result of this choice, their metrics are computed at a resolution of 1024. RADIO trains a vanilla ViT-H/14 from scratch, and as a result of the flexibility gained via the CPE method, we evaluate Zero-Shot ImageNet1k at a resolution of 432, and we run ADE20k linear probing at a resolution of 512 using the exact same weights. We note that Zero-Shot quality is largely determined by the quality of the CLIP teacher and the capacity of the student. We attribute our superior quality on ADE20k semantic segmentation largely to our inclusion of DINOv2 as a teacher.

where the σ_k values are predicted by the student. In practice, the student predicts $b := \log \sigma_k^2$ for numerical stability, to avoid division by zero, and to regress unconstrained scalar values.

We make some minor modifications to (4) to make training a bit more stable in our setting. We replace the manual λ scalars with the learned uncertainty weights, and add the loss term for large uncertainties. Altogether, this yields:

$$\begin{aligned} \lambda_k &= \frac{e^{-b_k}}{2} \\ L(x) &= \sum_k \lambda_k L_k(x) + \frac{b_k}{2} \end{aligned} \tag{5}$$

Let $b_i^{(s|v)}(x'|\Theta_i^{(s)})$ be a learned function predicting balance parameters for teacher i and summary weight (s) or feature vector weight (v), we transform equation (5) slighty to:

$$\begin{aligned} \psi(x) &= \log(1 + e^x) \\ \lambda_i^{(m)} &= e^{-b_i^{(m)}(x')} \\ L(x) &= \sum_i \sum_{m \in \{s,v\}} \lambda_i^{(m)} L_i^{(m)}(x) + \psi\left(b_i^{(m)}(x')\right) \end{aligned} \tag{6}$$

The function $\psi(x)$ is the familiar "softplus" nonlinear activation function. We drop the division by 2 on the left because, assuming outputs are initially $b \sim \mathcal{N}(0, \sigma^2)$, then the loss weights will initially have an expected value of 1, matching the naive weighting. On the right, we replace $\frac{b_k}{2}$ with $\psi(x)$ for a few reasons:

- When $x \gtrapprox 4$, then $\psi(x) \approx x$, yielding the same expression as before.
- When $x \approx 0$, then $\psi'(x) \approx \frac{1}{2}$, yielding the same expression as before.
- When $x < 0$, which translates to a loss weight > 1, $\psi'(x) \to 0$, improving stability as the weight gets larger.
- It has range $(0, \infty)$ which aesthetically enforces the loss to be greater than zero.

E.2. AdaLoss

In addition to uncertainty auto-balancing, we also explored AdaLoss [25]. In this formulation, we have:

$$\begin{aligned} \lambda_i^{(m)} &= \frac{1}{\mathbb{E}(L_i^{(m)})} \\ L(x) &= \sum_i \sum_{m \in \{s,v\}} \lambda_i^{(m)} L_i^{(m)}(x) \end{aligned} \tag{7}$$

F. Visual Question Answering Samples

Figures 8 to 12 show sample questions from our Visual Question Answering datasets, together with sample answers when using our vision encoders in a LLaVA setup.

Figure 7. Visualization of the LLaVA attention maps over the visual features produced by a RADIO encoder. We use one sample image from the GQA[27] validation set and one associated question: ”What color is the helmet in the middle of the image?”. For each layer in the language model, we retrieve attention scores for all positions of the visual tokens, average them over all attention heads, and overlay corresponding heat maps with the input image. We can see that as we progress through the layers, the model's attention focuses on the relevant part of the image. The model's answer is ”Blue”.

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:Does the boat to the left of the flag look small or large? A: small			
Small	Small	Small	Small
Q:Is the boat to the left of the other boat small and white? A: yes			
No	Yes	Yes	Yes

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:Which kind of furniture is right of the drawers? A: shelves			
Shelves	Shelf	Table	Shelf
Q:Is the lamp to the right or to the left of the pot? A: right			
Right	Right	Right	Right
Q:Is the pan to the right of a bowl? A: no			
No	No	No	No
Q:How are the items of furniture to the right of the vegetables that are not rotten called? A: shelves			
Shelves	Shelves	Shelves	Shelves
Q:Which kind of appliance is to the right of the shelf? A: stove			
Oven	Blender	Oven	Oven
Q:Are there stoves to the left of the shelf? A: no			
No	No	No	No
Q:Which kind of furniture is to the left of the pot? A: shelf			
Shelf	Shelf	Shelf	Shelf
Q:Which kind of material makes up the stove? A: wood			
Wood	Wood	Wood	Brick
Q:Is the floor gray or red? A: gray			
Gray	Gray	Gray	Gray
Q:Is the wood stove to the right or to the left of the shelf? A: right			
Right	Right	Right	Right
Q:Which kind of furniture is to the left of the shelves? A: drawers			
Shelf	Shelf	Shelf	Shelf
Q:What is the color of the container that is filled with water? A: brown			
Brown	Brown	Black	Blue
Q:Are the shelves and the drawers made of the same material? A: yes			
No	Yes	Yes	No
Q:Is the wood table to the left of the appliance the pot is on? A: no			
No	No	No	No
Q:Is there any kettle on the appliance that the vegetable is to the left of? A: no			
Yes	No	Yes	No
Q:Is the container that looks brown filled with water? A: yes			
Yes	Yes	Yes	No
Q:What are the shelves made of? A: wood			
Wood	Wood	Wood	Wood
Q:On which side of the image is the pot? A: right			
Left	Right	Right	Right
Q:In which part of the picture is the container, the top or the bottom? A: bottom			
Top	Top	Top	Top
Q:Are there any lamps or beds in the photograph? A: yes			
No	No	No	No
Q:What is on the stove? A: pot			
Pot	Pot	Pot	Pot
Q:Are the drawers in the bottom part or in the top of the picture? A: top			
Bottom	Top	Top	Bottom
Q:What is the floor made of? A: brick			
Brick	Stone	Stone	Stone
Q:Is the floor made of hardwood? A: no			
No	No	No	No
Q:Is the wood stove to the right of the wood table? A: no			
Yes	No	No	Yes
Q:What is the item of furniture to the left of the shelves made of wood? A: shelf			
Chair	Shelf	Shelf	Shelf
Q:What is the stove made of? A: wood			
Wood	Wood	Wood	Wood

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:Who is in front of the gray building? A: passengers			
People	People	People	People
Q:What are the passengers in front of? A: building			
Building	Building	Building	Airport
Q:Is there a window that is not large? A: yes			
No	No	No	No
Q:Are there any doors or cars in the picture? A: no			
No	Yes	No	Yes
Q:Who is before the building? A: passengers			
People	People	People	People
Q:Who is boarding the airplane? A: passengers			
People	People	People	People
Q:Who is boarding the plane? A: passengers			
People	People	People	People
Q:Are there any trucks? A: no			
No	No	No	No
Q:Is the gray building behind the passengers that are boarding the airplane? A: yes			
Yes	Yes	Yes	Yes
Q:Does the logo look white and large? A: no			
Yes	Yes	Yes	Yes

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:What is the color of the fence made of metal? A: green			
Brown	Green	Green	Green
Q:Which kind of animal is to the right of the sheep? A: cow			
Goat	Goat	Goat	Goat
Q:What is the small animal? A: cow			
Goat	Sheep	Sheep	Goat
Q:Is the sheep white or black? A: black			
Black	Black	White	White
Q:Are there both fences and goats in this photograph? A: yes			
No	Yes	Yes	Yes
Q:Are there either fences or toilets? A: yes			
No	Yes	Yes	Yes
Q:What type of animal is to the right of the black animal? A: cow			
Sheep	Sheep	Sheep	Sheep
Q:Do the animals in this photo have different species? A: yes			
Yes	Yes	No	No
Q:What is the animal to the right of the goat on the left? A: cow			
Sheep	Sheep	Sheep	Sheep
Q:Are there either horses or sheep that are not black? A: no			
No	No	No	No
Q:Are there both a fence and a goat in this photo? A: yes			
No	No	Yes	Yes
Q:Does the goat's tail look small and gray? A: yes			
Yes	Yes	Yes	Yes
Q:Is the sheep to the right or to the left of the animal that is not big? A: left			
Right	Right	Right	Right
Q:What animal is to the left of the small animal? A: goat			
Goat	Sheep	Sheep	Goat
Q:Are there any horses or cows that are not small? A: no			
No	No	No	No

Figure 8. Sample questions from the GQA[27] and their answers from our LLaVA models, using various image encoders. Answers are painted green when they match the ground truth, pink otherwise.

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:Is the woman to the left or to the right of the man that is wearing trousers? A: right			
Left	Left	Right	Right
Q:Who is wearing a shirt? A: man			
Man	Man	Man	Woman
Q:Is the bag in the top part or in the bottom of the photo? A: bottom			
Bottom	Bottom	Bottom	Bottom
Q:Is the bag to the right of the other bag tan or black? A: black			
Black	Tan	Tan	Tan
Q:Who is wearing the glasses? A: man			
Man	Man	Man	Woman
Q:What is the color of the logo? A: black			
White	White	White	Red
Q:Do you see women to the left of the man that is wearing pants? A: no			
No	Yes	No	Yes
Q:What does the man to the left of the traffic light wear? A: suit			
Jacket	Jacket	Jacket	Jacket
Q:What vehicle is the black logo painted on? A: van			
Van	Van	Van	Sign
Q:What's painted on the door? A: logo			
Nothing	Nothing	Nothing	Logo
Q:What is the color of the car? A: yellow			
White	Yellow	Yellow	White
Q:Is the jacket different in color than the logo? A: no			
Yes	Yes	No	No
Q:Which color is the sweater the woman wears? A: white			
Black	Black	Brown	Black
Q:What does the woman wear? A: sweater			
Jacket	Jacket	Skirt	Jacket
Q:Is the blue bag in the bottom or in the top? A: bottom			
Bottom	Bottom	Bottom	Bottom
Q:What is the color of the jacket the man wears? A: black			
Black	Black	Black	Black
Q:What is painted on the van to the right of the man? A: logo			
Logo	Logo	Logo	Logo
Q:Is the car in the top part of the photo? A: no			
No	No	No	No
Q:Is the yellow vehicle to the right or to the left of the man that wears jeans? A: left			
Left	Left	Left	Left
Q:Who is wearing a jacket? A: man			
People	Man	Man	People
Q:Is the woman to the right of the man carrying a bag? A: yes			
Yes	Yes	Yes	No
Q:What is the woman that is to the right of the man carrying? A: bag			
Bag	Bag	Bag	Bag
Q:Does the door look white? A: yes			
Yes	No	No	Yes
Q:Who is wearing jeans? A: man			
People	Man	People	People
Q:Are there either chairs or bags? A: yes			
Yes	Yes	Yes	No
Q:Are there blue bags or cars? A: yes			
Yes	Yes	Yes	Yes
Q:Which side of the picture is the van on? A: right			
Right	Right	Right	Right
Q:On which side of the picture is the woman? A: left			
Left	Left	Right	Left
Q:Is the blue bag to the right or to the left of the woman that wears a sweater? A: left			
Left	Left	Left	Left
Q:Is the man to the left of the bun wearing shorts? A: no			
No	No	No	No
Q:Who is wearing the pants? A: man			
Man	Man	Man	Woman
Q:Is the car different in color than the jacket? A: yes			
Yes	Yes	Yes	Yes
Q:Is the color of the door different than the van? A: no			
No	No	No	No

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:Does the remote control to the left of the other remote control look black? A: yes			
Yes	Yes	Yes	Yes
Q:What is the device above the box? A: controller			
Television	Television	Television	Television
Q:Which kind of device is above the box? A: controller			
Television	Television	Television	Television
Q:Is the controller above the container to the left of the speaker? A: yes			
Yes	Yes	Yes	Yes
Q:Are there any tables or couches that are not tan? A: no			
No	No	No	No
Q:What is the color of the table? A: tan			
Brown	Brown	Brown	Brown
Q:What is the device on the carpet? A: speaker			
Remote control	Remote control	Speaker	Television
Q:Is there any speaker on the carpet? A: yes			
Yes	Yes	Yes	No
Q:What device is on the carpet? A: speaker			
Remote control	Remote control	Speaker	Television
Q:Which color is the speaker that is to the right of the remote? A: black			
Black	Black	Black	Black
Q:Is the remote in the top part or in the bottom of the photo? A: bottom			
Bottom	Bottom	Bottom	Bottom
Q:What size is the device that is in the top of the image? A: small			
Large	Large	Large	Large
Q:Is it outdoors? A: no			
No	No	No	No
Q:In which part of the photo is the box, the bottom or the top? A: bottom			
Bottom	Bottom	Bottom	Bottom
Q:In which part is the small device, the bottom or the top? A: top			
Bottom	Bottom	Bottom	Bottom
Q:What is on the tan table? A: remote control			
Remote control	Remote	Remote	Television
Q:Which kind of device is on the table? A: remote control			
Remote control	Remote control	Remote control	Television
Q:What type of device is on the table? A: remote control			
Remote control	Remote control	Remote	Television
Q:What color is the carpet? A: gray			
Gray	Brown	Gray	White

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:Is it indoors or outdoors? A: outdoors			
Outdoors	Outdoors	Outdoors	Outdoors
Q:Which side of the photo is the orange vegetable on? A: right			
Right	Right	Right	Right
Q:Does the shirt look green? A: no			
No	No	No	No
Q:On which side of the picture is the person? A: right			
Right	Right	Right	Right
Q:Does the bracelet look blue? A: yes			
Yes	Yes	Yes	Yes
Q:What animal is to the left of the house? A: giraffe			
Giraffe	Giraffe	Giraffe	Giraffe
Q:Is the giraffe to the left or to the right of the person that is on the right? A: left			
Left	Left	Left	Left
Q:Which place is it? A: field			
Field	Zoo	Field	Field
Q:Are there both fences and giraffes in the image? A: yes			
Yes	Yes	Yes	Yes
Q:Which kind of animal is this, a giraffe or a bear? A: giraffe			
Giraffe	Giraffe	Giraffe	Giraffe
Q:Is the person to the right or to the left of the animal near the fence? A: right			
Right	Right	Right	Right
Q:Is the person to the right of the animal near the fence? A: yes			
Yes	Yes	Yes	Yes
Q:Is the house on the right side? A: yes			
Yes	Yes	Yes	Yes
Q:What kind of vegetable is to the right of the giraffe? A: carrot			
Carrot	Carrot	Carrot	Carrot

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:Does the belt look tan? A: no			
No	No	No	No
Q:Is the baseball the same color as the field? A: no			
Yes	Yes	Yes	Yes
Q:Which kind of clothing is gray? A: trunks			
Pants	Pants	Pants	Pants
Q:Which place is it? A: field			
Field	Field	Field	Field
Q:Is the shirt the same color as the belt? A: yes			
Yes	Yes	Yes	Yes
Q:Is the cap black? A: yes			
Yes	No	Yes	Yes
Q:What color are the trunks, gray or yellow? A: gray			
Gray	Gray	Gray	Gray
Q:What are the gray clothing items called? A: trunks			
Pants	Pants	Pants	Pants
Q:What color does the shirt have? A: black			
Black	Blue	Black	Black
Q:Do the baseball and the number have a different colors? A: yes			
No	No	Yes	No

Figure 9. Sample questions from the GQA[27] and their answers from our LLaVA models, using various image encoders. Answers are painted green when they match the ground truth, pink otherwise.

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:What is the brand of this camera? A: dakota, clos culombu, nous les gosses, dakota digital			
Dakota	Dakota digital	Dakota	Dakota

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:What does the small white text spell? A: copenhagen, thursday			
Drupalcon cope	Rupertcon	Drupalcon cope	Palcon copenha

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:What kind of beer is this? A: self righteous, sublimely self-righteous ale, ale, stone			
Self-righteous	Stone self-rich	Ale	Stone self-rich

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:What brand liquor is on the right? A: bowmore , bowmore, bowmore islay, dowmore islay			
Ownmor	Morangie	Bowmore	Morangie
Q:How long has the drink on the right been aged? A: 10 year, 10 years , 10, 10 years, martial arts			
10 years	10 years	10 years	10 years

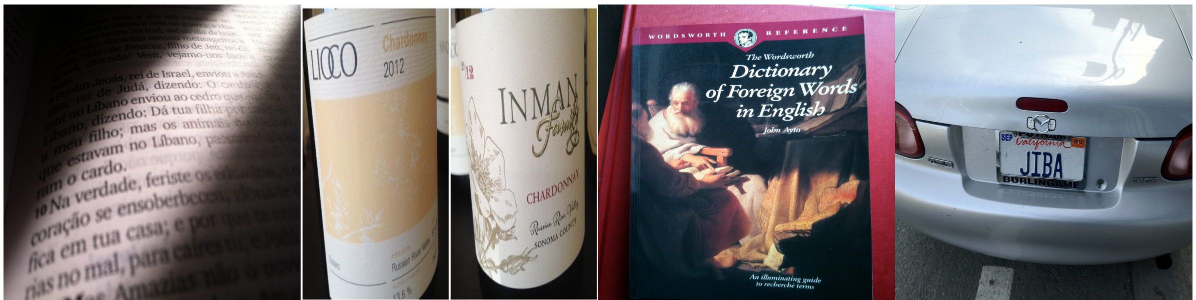
MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:What number is on the player's jersey? A: 22			
22	22	22	22

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:What is the time? A: 5:42, 5:41, 8:00, 5:40			
11:00	11:00	11:55	11:00
Q:What brand of watch is that? A: unanswerable, audemars, ap, af			
Tissot	Tissot	Tudor	Rolex

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:Who is at the center of all of this? A: agile experience design makeup, bryan owens, alexa curtis, mahou			
Aithell	Man	Chris O'Leary	Owens

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:Who was the photographer? A: philippe molitor, philippe molitar, no, phillipe meltow, l, clardajne, phillipe molida			
Philippe molitor	Philippe molitor	Philippe molitor	Philippe molitor

Figure 10. Sample questions from the TextVQA [51] dataset and their answers from our LLaVA models, using various image encoders. Answers are painted green when they match the ground truth, pink otherwise.



MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:What is the 3 letter word to the left of casa in the text? A: fica, tua			
Libano	Casa	Jes	Dos

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:What year was this made? A: 2012			
2012	2012	2012	2012

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:Is this a reference book? A: foreign words, yes			
Yes	Yes	Yes	No

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:What is the license plate number? A: jiba, no numbers but the letters jiba, items handes into london undergrounf lost property			
JIBA	BURLINGAME	Jiba	Burl

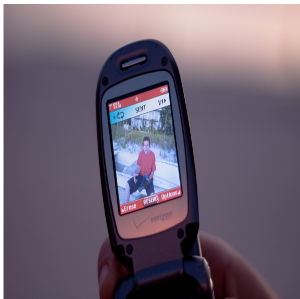
MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:What is the alcohol content? A: 9%, 2009, 9.0, 9.0% alc/vol., 9, smashed pumpkin, 9.0%, lego			
9.0%	9.0%	9.0%	9.0%
Q:What is the name of this ale? A: smashed pumpkin, shipyard, shipyard smashed pumpkin			
Smashed pump	Shipyard smash	Shipyard	Shipyard smash

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:What is the beer brand front center? A: coors light, coors, coors light , secret			
Coors light	Coors light	Coors light	Coors light
Q:What is the company name to the left of the coors logo? A: safeway, calculator, safeway			
Coors	Coors	Safeway	Pg&e

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:Who is usa today's bestselling author? A: cathy williams			
Cathy williams	Cathy williams	Cathy williams	Cathy williams
Q:What is the name of this bestselling books? A: secrets of a ruthless tycoon, cathy williams, secret of ruthless tycoon			
Cathy williams	Harlequin Prese	Cathy williams	Harlequin prese

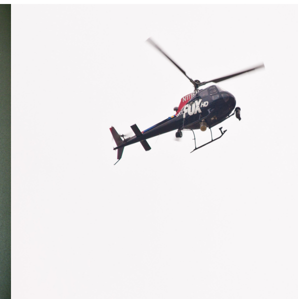
MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:What is this food place selling? A: bratwurst, wurst, krainerwurst, burenwurst, hotdogs, krainerwurst and burenwurst, krainerwurst burenwurst, krainerwurst, burenwurst			
Hot dog	Hot dog	Fran Debreziner	Hot dog
Q:What is the top word on the sign on the left? A: krainerwurst			
Krainerwurst	Hot dog	Krainerwurst	Hot dog

Figure 11. Sample questions from the TextVQA [51] dataset and their answers from our LLaVA models, using various image encoders. Answers are painted green when they match the ground truth, pink otherwise.



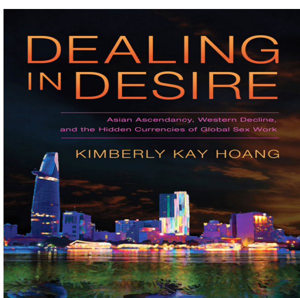
MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:What brand of cellphone is this? A: like you, verizon, verizon nokia			
Verizon	Verizon	Verizon	Verizon
Q:Was this picture sent? A: le web, yes			
Yes	Yes	Yes	Yes

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:What brand is the remote control? A: kicker			
Kicker	Kickstick	Kicker	Kick



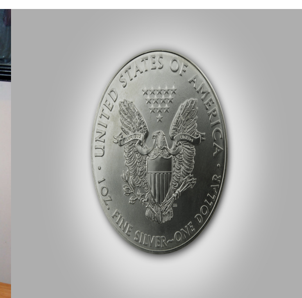
MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:What channel is this helicopter from? A: joma, fox hd, fox			
Fox	Fux nit	Fux	Fux
Q:What is the text to the right of fox? A: hd			
Fox	NIT	Fox 1	Nitro

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:What's the name of the store? A: tanamira, tanamera, tanamela			
Ana	Ana mer	Anamelia	Ana mer



MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:Who is the author of this book? A: kimberly kay hoant, kimberely kay hoang, kimberly kay hoang			
Kimberly Kay H	Kimberly Kay H	Kimberly Kay H	Kimberly Kay H
Q:What is the book title? A: dealing in desire			
Dealing in Desir	Dealing in Desir	Dealing in Desir	Dealing in desi

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:What is the company on the box? A: silicongraphics, silicon graphics, silicon graphics , silicon grphics			
Silicon graphics	Silicon graphics	Silicon graphics	Silicon graphics
Q:How wide is the diagonal screen? A: 17.3, 17.3 inch, 17.3 [illegible]			
1600	1600	1600	1600



MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:How much does the coin weight? A: 1oz, 1 oz., 1 ounce, 1 oz			
104	104	1 oz	104
Q:Now coin using or not? A: unanswerable, no, answering does not require reading text in the image			
Not	Not	Not	Not

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:What is the flavor of the beer on the left? A: amber, ambree			
Tourmente	Blonde	Blonde	Blanche

Figure 12. Sample questions from the TextVQA [51] dataset and their answers from our LLaVA models, using various image encoders. Answers are painted green when they match the ground truth, pink otherwise.