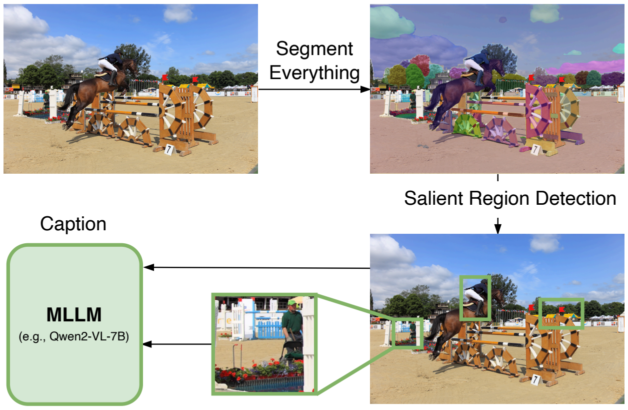


Figure 10. **Curation of bounding boxes and captions of salient regions in the pre-training data.** For each high-resolution image, we segment all the masks, detect salient regions with small or dense masks, and use an MLLM to generate captions about the local regions.

A. Details of PS3 Pre-Training Data Curation

A.1. General Overview

The curation of each component of the pre-training data is detailed below.

High-resolution images. We collect 75M images with 1K - 4K resolution. These include 38M natural images from DataComp [11] and SA-1B [15] and 37M document images from IDL [2] and PDFA [27].

Local captions and bounding boxes of salient regions for natural images. We propose a pipeline of first detecting salient regions and then generating local captions (Figure 10). For saliency detection, inspired by recent work on segmenting everything in an image [15, 42, 51], we first use EfficientViT-SAM [51] to generate segmentation masks of the whole image, and then select local regions containing small or dense masks as salient regions. The intuition is that a local region should contain small or cluttered objects in order to have rich details. The saliency detection algorithm is explained in more details in the Appendix **??**. For local captions, we use an off-the-shelf MLLM (*e.g.*, Qwen2-VL [41]) as an captioner. Specifically, we zoom in and crop the local region, send it along with the global image to the MLLM, and let it describe the local region based on the global context. This results in 3 - 4 pairs of local captions and bounding boxes per image and 134M pairs in total, with an average box size around 400×400.

Local captions and bounding boxes for documents. Both IDL [2] and PDFA [27] provide bounding boxes and OCR results of sentences or words in each PDF. Therefore, we simply sample from these bounding boxes and use the corresponding OCR results as the local captions. We generate 148M pairs of boxes and captions with an average box size around 50×400.

Global captions. Our pre-training also uses global captions (see Section 2.3). We use the same MLLM captioner to generate global captions for natural images. For document images we do not use any global captions.

A.2. Details

The full pre-training data sources and statistics are listed in Table 3.

Table 3. **Data sources and statistics.** We collect in total 75M images with 1K - 4K resolution and 282M pairs of bounding boxes and detailed captions about salient local regions in the images.

Data Source	1K - 2K Res.			2K - 4K Res.		
	#Img	#Box	Avg. Box size	#Img	#Box	Avg. Box size
Natural images						
DataComp [11]	18M	54M	424×438	9M	36M	562×578
SA-1B [15]	-	-	-	11M	44M	302×312
Documents						
IDL [2]	12M	48M	28×286	7M	28M	30×330
PDFA [27]	12M	48M	80×461	6M	24M	84×569
Agg.	42M	150M	-	33M	132M	-

High-resolution images. The images consists of two types, natural images and documents. For natural images, we collect 18M images with 1K - 2K resolution and 20M images with 2K - 4K resolution from DataComp [11] and SA-1B [15]. For documents, we take all 37M PDF pages from IDL [2] and PDFA [27] and convert each page into image with DPI of 150, which normally results in images with resolution above 1.5K.

Local captions and bounding boxes of salient regions for natural images. In the saliency detection pipeline described in Section 2.1, we first use EfficientViT-SAM [51] to generate all the masks in each image, similar to the "segment everything" mode in the original SAM [15]. We use EfficientViT-SAM with model size of XL1. The arguments used for generating masks are listed in Table 4. Notably, `points_per_side` and `crop_n_layers` largely affect how dense and detailed the generated masks are.

After generating all the masks, we locate local bounding boxes of salient regions that contain small or dense masks. The detailed process is as follows: 1) We preset a set of boxes which are square, have the same sizes, and are uniformly distributed in the image with the distance between adjacent boxes equal to the size of the box. The box size is set depending on the size of the image. For example, for SAM, we set the box size as $1/5$ of the shortest side of the image. This results in a typical box size of 300 - 400 in a 2K resolution image. For each square box, we also preset two boxes at the same position with the same area as the square box but with aspect ratios of $1.5 : 1$ and $1 : 1.5$, respectively. 2) For each box, we calculate the saliency score of the box. The saliency score is the accumulation of the

Table 4. **Arguments for EfficientViT-SAM mask generation.**

Argument	Value
`points_per_side`	24
`points_per_batch`	128
`crop_n_layers`	1
`crop_n_points_downscale_factor`	1
`pred_iou_thresh`	0.6
`stability_score_thresh`	0.85
`min_mask_region_area`	0

scores contributed by each mask that has overlaps with the box. The contribution to the score from a mask is calculated as $\frac{1}{\max(\texttt{Area}(mask),\ 40\cdot 40)\ /\ \texttt{Area}(image)} \cdot \frac{\texttt{Area}(mask \cap box)}{\texttt{Area}(mask)}$, where the first term is larger when the area of the mask is smaller compared to the area of the whole image and the second term is larger when a larger portion of the mask resides inside the box. We then select the top-k boxes with the highest saliency scores. To encourage the selected boxes to cover more areas, we ensure no overlap between boxes.

Finally, after detecting the local salient boxes, we generate captions about the local region by sending two images, the local crop and the global image, to an MLLM and asking it to generate a caption about the local details given the global context. Here we use Qwen2-VL [41] as the MLLM since it has superior results to other open-source MLLMs and can handle multiple images. We set $\texttt{max\_pixels} = 256\cdot 28\cdot 28$ for Qwen2-VL to handle both the global image and the local crop. The prompt following the two images is: *The second image is a crop from the first image. Given the context of the first image, please describe the second image briefly. Make sure to cover all the objects and texts in the second image. Make sure to describe all the attributes, including color, shape, spatial relations, of each object in the second image. If there's no text in the second image, you don't need to mention there's no text. Please only describe the objects in the foreground and don't describe the background such as the sky or the weather. Please only describe the objects in the second image and don't describe the objects in the first image. Please use 1-2 sentences.*

Local captions and bounding boxes for documents. IDL provides bounding boxes and OCR results of each sentence, and PDFA provides the same labels for each word. For IDL, we randomly sample one sentence and use the bounding box as well as its OCR result as the caption. For PDFA, we sample 15 consecutive words and use the union of their bounding box and concatenate these words as the caption.

Global captions. We directly use Qwen2-VL to caption the whole image. The prompt is as follows: *Please describe the image briefly. Make sure to cover all the objects and texts in the image. Make sure to describe all the attributes, including color, shape, spatial relations, of each object in the image. If there's no text in the image, you don't need to mention there's not text. Please only describe the objects in the foreground and don't describe the background such as the sky or the weather. Please use 2-3 sentences.*

Examples of the pre-training data. See Figure 11-12.

B. Additional Details of PS3 Pre-Training Algorithm

Learning high-res visual representation. The model learns detailed high-res visual representation by optimizing the contrastive loss between visual features of the local regions and the text embedding of the local captions. As mentioned in Section 2.3, we mix global and local contrast to prevent degradation of the global visual features, *i.e.*, each batch contains both global image-caption pairs and local region-caption pairs. There are several additional details worth highlighting: **1) Pooling only tokens in the ground-truth boxes.** SigLIP uses attention pooling to compress all the output tokens into one for contrastive loss. When a box contains fewer patches than the pre-set k, the model will select patches outside the box as well. This results in aligning irrelevant visual features to the text embedding in contrastive loss. To avoid this, we constrain the attention pooling only to tokens inside the box. **2) Avoiding intra-image contrast.** Since we have multiple local box and captions for each high-res image, there is a chance that one batch contains multiple local regions from the same image. It can be problematic to contrast between different regions of the same image if those regions are visually similar [5]. We make sure each image only appears once in a batch to avoid intra-image contrast. We use either the global image-caption pair or one of the local region-caption pair for each image in the batch with the probability of 50% each. We find this empirically improve the average accuracy across benchmarks by a marginal 0.1%.

Learning top-down and bottom-up patch selection. As described in Section 2.3, PS3 learns patch selection with the supervision from labeled bounding boxes. The objective is a segmentation loss between the predicted selection score map and the ground truth map. Note that the selection score depends on the low-res features from Stage 1 of PS3, which means the patch selection loss has gradients on the features. We empirically find this degrades the quality of low-res features and during pre-training we detach the low-res features before predicting the selection score.

Hyperparameters of PS3 pre-training. During pre-training, since the data comes from different data sources (SAM, DataComp, IDL, and PDFA), we sample from each data source such that the probability each data source is sampled is the same for all data sources. Table 5 shows other hyperparameters.

Table 5. **Hyperparameters of PS3 pre-training.**

Argument	Value
#epochs	15
#samples each epoch	5e6
global batch size	8192 / 4096 (for 3780 resolution)
learning rate	5e-6
warmup iterations	1500
beta1	0.9
beta2	0.95
weight decay	3e-4

C. Additional Comments on Training MLLMs with PS3

Parallel training of next-token prediction. Since the high-res patch selection is dependent on the latent embedding of the previous tokens, we cannot train next-token prediction for every token in parallel. Therefore, in training, we first run LLM on the low-res vision features and the text tokens to get the last-layer embedding of the last token, use it to extract high-res vision features, and then run LLM on the full sequence again for next-token prediction. Although this requires running LLM twice, we empirically observe the additional computational cost is marginal since the first run is on a shorter sequence and it does not require gradient because it is only for obtaining the last-layer embedding which is detached before the patch selection. Note that during inference the next token prediction is sequential so we only need to run LLM once for each token.

D. Ablations and Analysis

D.1. Key Designs in Pre-Training Algorithm and Model Architecture

We conduct ablation studies on the pre-training algorithm designs (Section 2.3), PS3 model designs (Section 2.2), and MLLM model design (Section 3.1). We follow the same experimental setting as in Section 4 and use PS3 with max resolution of 1512 and 100% patch selection as the baseline. For the ablation of each design, we report the improvement of the baseline model on the average accuracy of the seven benchmarks compared to the model without the design. Results are shown in Table 6. Overall all the designs are helpful. Among the pre-training algorithm designs, we can see that it is crucial to select the patches in the ground-truth boxes during patch selection and pool only their corresponding tokens before calculating the contrastive loss. This is because otherwise the model will select irrelevant patches and contrast them with the local caption, leading to noisy pre-training and degrading the performance by over 8%. We also find avoiding contrast between local regions in the same image improves the performance by 5.3% which aligns with the observation in previous work [5]. The model architecture designs also help with the performance. Notably, the low-res KV cache allows the model to see the global context while extracting the local high-res features, which significantly improves the performance by 8.8%. Additionally, extracting features at multiple scales (*e.g.*, 756 and 1512) performs better than only extracting features at the largest scale (*e.g.*, 1512 only) as in AnyRes. Adding scale-aware positional embeddings in PS3 and additional vision positional embeddings when feeding PS3 features to LLM also improve the performance by a small margin.

D.2. Top-Down and Bottom-Up Patch Selection Matters

We compare the performance of using random, bottom-up, and top-down patch selection, as shown in Table 7. We report benchmark accuracy as well as the patch recall rate which evaluates how many patches out of the ground-truth region are selected at test time. This is evaluated on a test set split from the data used to train patch selection in MLLM. We can see that both bottom-up and top-down selection significantly improves the recall and the accuracy over random selection. For example, when selecting 44% patches, bottom-up selection improves the recall rate by 43.7% and the accuracy by 7.4% over random selection. Top-down selection further improves the recall rate by 3.8% and the accuracy by 1.3%. Interestingly, patch selection not only affects test time but training as well. When we train MLLM with 44% patches, even if we select all patches at test time, top-down selection is still better than the other two. The reason is likely that top-down selection provides more informative visual context for MLLM during training, which leads to less noisy training dynamics. Note that this model has performance comparable to training with 100% patches, which means that training with top-down patch selection improves training efficiency without hurting the performance too much.

Key Finding 5: Top-down and bottom-up patch selection provides more relevant visual information to MLLM, which does not only improve performance at test time but also improve the model optimization during training.

D.3. Trade-Off Between Different Image Scales

We find the optimal balance of patch selection between image scales varies for different tasks (Figure 13). For example, when using in total 729 (20%) high-res tokens, V*Bench performance peaks when selecting no patches (0%) at 756 scale and only patches (25%) at 1512 scale, while it performs the best on DocVQA when selecting 67% 756-scale patches and only 8% 1512-scale patches. Note that selecting more patches at 756 scale covers more regions of the image because one 756-scale patch represents larger area in the

Table 6. **Ablation of PS3 pre-training, model, and MLLM designs.** Δ is the change of the average performance on the seven benchmarks after adding the design.

Training and Model Design Choices	Δ
Pre-training algorithm designs (*Section 2.3*)	
- using ground truth selection score	+5.1
- pooling only tokens in ground-truth boxes	+3.7
- mixing global and local contrast	+1.0
- w/o intra-image contrast	+5.3
PS3 model designs (*Section 2.2*)	
- multi-scale feature extraction	+1.0
- scale-aware pos. emb.	+0.8
- low-res KV cache	+8.8
MLLM model design (*Section 3.1*)	
- additional vision pos. emb.	+0.8

Table 7. **Ablation of top-down and bottom-up patch selection.** *Select (Train)* and *Select (Test)* are the percentage of high-res patches PS3 selects at training and test time. *Recall* is the recall rate of how many patches in the ground-truth regions are selected at test time.

Patch Selection	Select (Train)	Select (Test)	Recall (Test)	Avg. Acc.
Random	44%	44%	43.7%	52.3
Bottom-up	44%	44%	87.4%	59.7 (+7.4)
Top-down	44%	44%	91.2%	61.0 (+8.7)
Random	44%	100%	100%	56.5
Bottom-up	44%	100%	100%	61.1 (+4.6)
Top-down	44%	100%	100%	61.9 (+5.4)
Top-down	100%	100%	100%	63.2

original image than a 1512-scale patch, but lose more details at the 1512 scale. This indicates V*Bench needs smaller coverage of image regions but requires more high-res information at 1512 resolution to perform well. DocVQA, on the other hand, needs more coverage of the image, which is probably because one usually needs to read the whole document to answer the questions. We also observe degraded performance on all benchmarks when selecting only tokens at 756 scale, which is because not all tokens at 756 scales are relevant to the question while the relevant information from 1512 scale is completely lost in this way.

D.4. SFT Data for High-Resolution Feature Alignment

We study the effect of high-resolution SFT data for MLLMs using PS3 as the vision encoder. Commonly-used image QA data normally contains only low-res images or lacks questions about details in high-res images, from which it is hard for MLLM to learn to utilize the high-res vision features. We hypothesize the high-res QA data in Section 3.2, though generated in a naive way, can help align the high-res visual features to the text space of LLM, thus improving the high-res perception capability. We conduct an ablation study in Table 8. We find that the high-res SFT data in general helps with the performance on resolution-sensitive benchmarks. The improvements are especially significant on natural images, *e.g.*, a 3.1% improvement on V*Bench. On the other hand, the performance on ChartQA and DocVQA does not change, which is probably because the current high-res SFT data does not contain document images.

D.5. Visualization of PS3 Visual Representations

We visualize the visual features in PS3 using principal component analysis (PCA) and compare with the baselines of S^2 and AnyRes that does not pre-train vision models at high resolutions. Results are shown in Figure 14. We visualize the features of SigLIP-S^2 and SigLIP-AnyRes at 3780×3780 resolution and the features from three different scales (756×756, 1512×1512, and 3780×3780) for PS3. We can see the feature map of SigLIP-S^2 is blurry compared to other models because S^2 concatenates both low-res and high-res features and the PCA visualization is partially dominated by the low-res components. The AnyRes feature map, which is equivalent to S^2 but only with the high-res features, shows noisy patterns. One possible reason is AnyRes splits the high-res image into small tiles and individually processes each, which means the features lack global context and are inconsistent when running PCA on the whole feature map. On the other hand, PS3 shows sharp and fine-grained features at 4K resolution, *e.g.*, at the edges of objects. We also observe PS3 features from different scales tend to group pixels at different scales. For example, the features at 756 and 1512 scales group large-scale objects or things such as the dock and the sky, while the features at 3780 resolution usually group small-scale patterns such as the texts on the banners.

D.6. Comparing PS3 to State-Of-The-Art Token Pruning Methods

We compare the efficiency of PS3 on high-res images with state-of-the-art token pruning methods for MLLMs including ToMe [3], FastV [6], and VisionZip [47]. The previous methods are heuristic-based, either merging visual tokens based on their similarities or pruning tokens based on the attention score in the vision encoder or LLM backbone. In contrast, PS3 adopts a learning-based approach to select important or relevant visual tokens. We test the models in both 1512 and 3780 resolution and try different token selection ratio for each resolution. All the baselines use PS3 as the vision encoder for a fair comparison but instead of using the PS3 patch selection module, they use PS3 to process the whole image and prune the tokens in their own way. The results are shown in Table 9. We can see that PS3 achieves similar efficiency on the LLM backbone when pruning the same number of tokens. However, PS3 significantly reduces the ViT la-

Table 8. **Ablation of MLLM SFT data for high-resolution feature alignment.** The high-resolution SFT data (*HR Data*) generally improves high-resolution perception, especially on natural images.

Vision Encoder	Max Res.	HR Data	Text VQA	Chart QA	Doc VQA	Info VQA	OCR Bench	V* Bench	Real World	*Avg*
PS3	1512	✗	68.8	71.2	79.6	39.6	535	60.9	62.9	*62.4*
PS3	1512	✓	69.3	71.1	79.4	41.3	534	64.0	63.8	*63.2*
			(+0.5)	(-0.1)	(-0.2)	(+1.7)	(-0.1)	(+3.1)	(+0.9)	*(+0.8)*

tency (*e.g.*, 0.286s → 0.096s when selecting 25% patches at 1512 resolution) while other methods retain the high ViT latency because PS3 only processes the selected patches while other methods only prune the tokens after processing all the patches. This advantage is especially important for 4K resolution images where the ViT latency dominates the LLM latency. For the same reason, the previous methods all run out of memory on 4K resolution images. In the meantime, PS3 achieves superior results over all previous methods under the same selection ratio. This is due to several possible reasons, for example, the learning-based token selection is more accurate than previous heuristic-based methods given enough data, and PS3 is able to select patches based on the user prompt while several previous methods such as ToMe and VisionZip are prompt-agnostic.

D.7. Generalizability of PS3 Pre-Training to State-Of-The-Art Vision Encoders

We verify the generalizability of PS3 pre-training pipeline by pre-training PS3 on top of several state-of-the-art vision encoders and compare with the same vision encoders without high-res pre-training. Specifically, we take the state-of-the-art language-aligned vision model, SigLIP-SO400M [48], and agglomerative vision model, C-RADIO-v2-L [33], as the baselines. We take their pre-trained model and continue pre-training on high-res images using PS3 pre-training pipeline to get PS3-SigLIP-SO400M and PS3-C-RADIO-v2-L. We compare them with the original models as well as their AnyRes [20] and S^2 [35] variants. Results are shown in Table 10. We can see that both the PS3 models surpass the performances of their corresponding base vision models as well as the AnyRes and S^2 variants, showing that PS3 is a general high-res pre-training pipeline that can be applied to any base vision encoder. We also observe that C-RADIO baseline consistently outperforms SigLIP baseline, and that advantage is also inherited by the PS3-C-RADIO-v2-L model despite C-RADIO-based models have fewer parameters and fewer output tokens than SigLIP-based models.

E. Qualitative Examples of Patch Selection Fine-tuned with MLLMs

We show additional qualitative examples of bottom-up and top-down patch selection of PS3 in Figure 15 and 16.

F. Full Results of Scaling Properties of PS3

Table 11 shows the full results of the experiment of scaling properties (Figure 7) in Section 4.

G. Comparison to State of the Arts

H. Additional Ablation Studies

Trade-off between different image scales. We find the optimal balancing of number of patches between image scales varies for different tasks (Table 12). For example, when selecting in total 729 (20%) high-res patches, ChartQA and V*Bench peaks when selecting 33% patches at 768 scale, while it performs the best on DocVQA when selecting 50% - 67% 756-scale patches. Note that selecting more patches at 756 scale covers more regions of the image because one 756-scale patch takes larger area in the original image than a 1512-scale patch. This indicates ChartQA and V*Bench need smaller coverage of image regions but require more high-res information at 1512 resolution to perform well. DocVQA, on the other hand, needs more coverage of the image, which is probably because one usually needs to read the whole document to answer the questions.

I. 4KPro Data Curation

For each category, we first collect videos with 4K resolution (3840×2160) from YouTube, and for each video, we manually select 5 frames containing rich details that are only recognizable under high resolution and label the bounding boxes around the local details. To obtain QA pairs about the details for each frame, we first use GPT-4o [13] to generate a candidate QA pair based on the context of the global image and the content of the local crop, and then manually review and screen each generated QA pair to ensure each question is answerable under and only under 4K resolution and each answer is correct.

J. Additional Qualitative Results on 4KPro

We show additional qualitative comparison between PS3 and state-of-the-art MLLMs such as GPT-4o and Qwen2-VL in Figure 17.

Table 9. **Comparing PS3 to state-of-the-art token pruning methods.** PS3 has consistently lower ViT latency and achieve better performances than previous methods. PS3 is also the only method that can handle 4K resolution images.

Method	Select (Test)	ViT Latency	LLM Latency	Text VQA	Chart QA	Doc VQA	Info VQA	OCR Bench	V^* Bench	Real World	*Avg*
					1512 Resolution						
Full	100%	0.286s	0.375s	78.6	84.1	92.2	68.1	787	67.9	69.8	*77.1*
ToMe [3]	50%	0.286s	0.260s	74.1	70.2	59.7	47.3	622	66.8	67.2	*63.9*
FastV [6]	50%	0.286s	0.264s	**78.2**	81.2	**90.0**	60.4	769	66.2	69.0	*74.6*
VisionZip [47]	50%	0.286s	0.260s	75.2	77.2	79.8	55.7	722	64.0	67.1	*70.2*
PS3	50%	**0.167s**	0.260s	77.7	**83.4**	89.8	**60.8**	**774**	**67.9**	**69.1**	***75.2***
ToMe [3]	25%	0.286s	0.180s	72.5	65.5	51.7	42.8	61.1	62.2	63.4	*59.9*
FastV [6]	25%	0.286s	0.185s	76.1	66.3	78.1	49.5	651	64.6	65.2	*66.6*
VisionZip [47]	25%	0.286s	0.180s	74.6	76.0	72.8	51.5	694	62.7	64.6	*67.4*
PS3	25%	**0.096s**	0.180s	**76.8**	**80.4**	**84.4**	**54.6**	**738**	**65.7**	**67.8**	***71.9***
					3780 Resolution						
Full	100%	1.812s	OOM	-	-	-	-	-	-	-	-
ToMe [3]	20%	1.812s	OOM	-	-	-	-	-	-	-	-
FastV [6]	20%	1.812s	OOM	-	-	-	-	-	-	-	-
VisionZip [47]	20%	1.812s	OOM	-	-	-	-	-	-	-	-
PS3	20%	**0.417s**	0.383s	77.8	83.9	91.6	65.0	773	72.8	70.1	*76.9*

Table 10. **Generalizability of PS3 pre-training to state-of-the-art vision encoders.** *#Param of ViT* is the number of parameters of the ViT backbone. *Max Res.* is the maximum resolution each model processes in MLLM. *Max #Tok* is the maximum number of vision tokens in MLLM. All the PS3 models select all the high-res patches.

Vision Encoder	#Param of ViT	Max Res.	Max #Tok	Text VQA	Chart QA	Doc VQA	Info VQA	OCR Bench	V^* Bench	Real World	*Avg*
SigLIP-SO400M [48]	400M	378	196	62.3	56.6	51.9	30.7	387	51.8	57.1	*49.9*
+ AnyRes [20]	400M	1512	3332	67.4	58.4	67.9	34.1	468	60.2	59.0	*56.3*
+ S^2 [35]	400M	1512	2916	66.1	71.0	78.3	41.1	526	55.2	61.0	*60.8*
PS3-SigLIP-SO400M	400M	1512	3841	**69.3**	**71.1**	**79.4**	**41.3**	**534**	**64.0**	**63.8**	***63.2***
C-RADIO-v2-L [33]	320M	384	144	65.0	58.8	53.1	30.9	405	51.5	57.5	*51.0*
+ AnyRes [20]	320M	1536	2448	68.1	62.8	70.0	35.8	497	65.9	62.8	*59.3*
+ S^2 [35]	320M	1536	2304	68.1	72.3	82.5	40.4	542	59.7	62.1	*62.8*
PS3-C-RADIO-v2-L	320M	1536	3024	**68.4**	**72.6**	**83.2**	**43.4**	**569**	**68.2**	**61.5**	***64.9***

K. Full Results of Scaling Properties on 4KPro

Table 13 shows the full results of the experiment of scaling properties of PS3 on 4KPro (Figure 9 in Section 5.1).

Table 11. **Full results of scaling properties of PS3.**

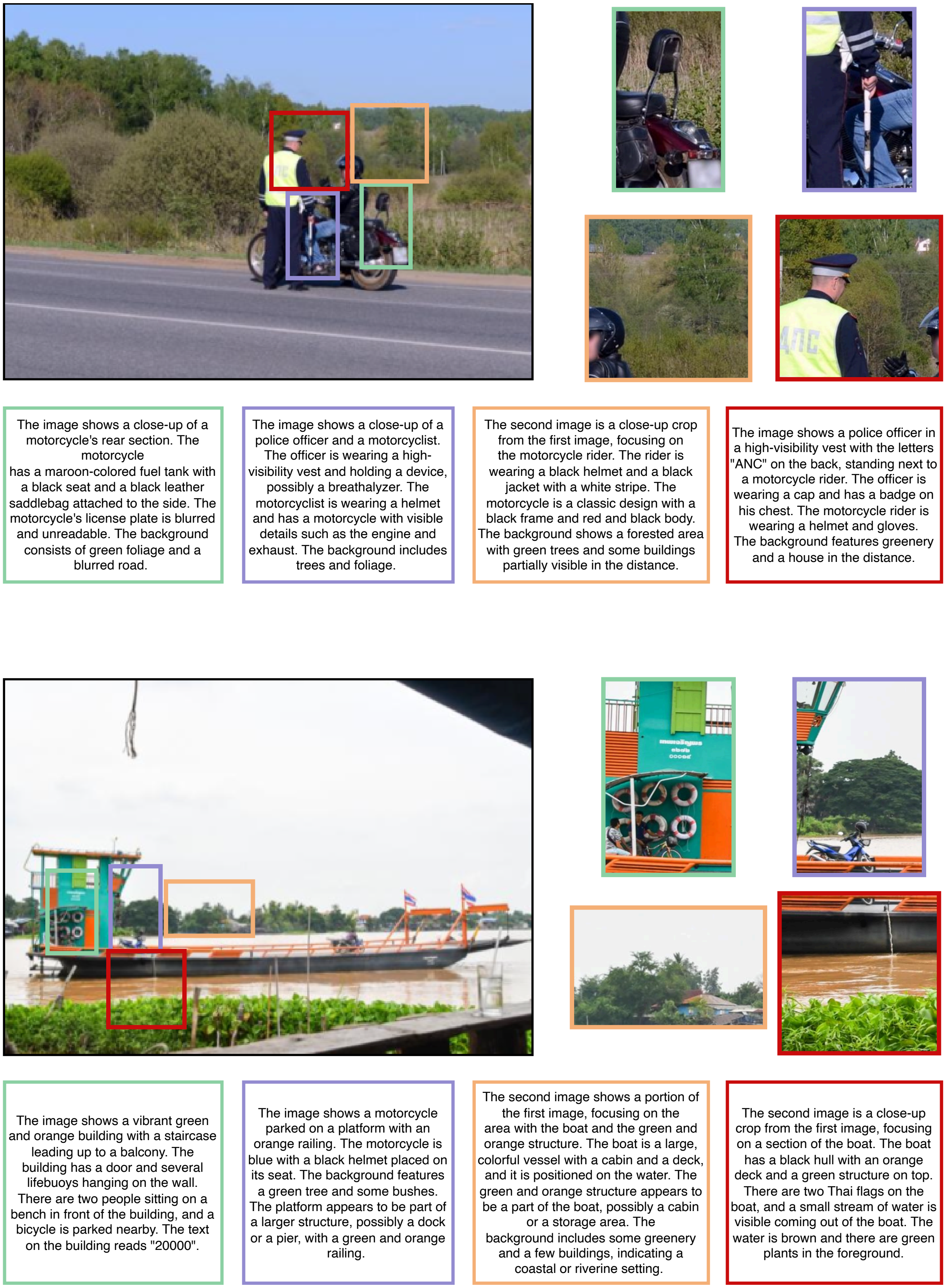
Vision Encoder	Max Res.	#HR Token	Select (Train)	Select (Test)	Text VQA	Chart QA	Doc VQA	Info VQA	OCR Bench	V* Bench	Real World	*Avg*
SigLIP [48]	378	0	-	-	62.3	56.6	51.9	30.7	387	51.8	57.1	*49.9*
SigLIP-DynRes [7]	756†	784†	-	-	65.3	58.0	60.6	32.7	416	59.2	59.1	*53.8*
SigLIP-S^2 [35]	756	784	-	-	65.9	65.5	63.0	32.3	471	53.1	59.6	*55.2*
PS3	756	320	44%	44%	66.7	62.8	62.6	33.1	460	56.3	61.7	*55.6*
PS3	756	729	100%	100%	66.8	63.5	64.6	33.9	462	56.5	61.7	*56.2*
SigLIP-DynRes [7]	1512†	3136†	-	-	67.4	58.4	67.9	34.1	468	60.2	59.0	*56.3*
SigLIP-S^2 [35]	1512	3136	-	-	66.1	71.0	78.3	41.1	526	55.2	61.0	*60.8*
PS3	1512	729	20%	20%	67.3	64.7	66.5	34.8	505	60.7	62.6	*58.2*
PS3	1512	1600	20%	44%	67.7	65.9	70.7	35.7	515	62.0	62.6	*59.4*
PS3	1512	1600	44%	44%	68.4	68.0	74.5	37.3	509	63.1	65.0	*61.0*
PS3	1512	3645	44%	100%	68.4	68.0	76.5	39.4	522	66.7	62.0	*61.9*
PS3	1512	3645	100%	100%	69.3	71.1	79.4	41.3	534	64.0	63.8	*63.2*
PS3	3780	3840	18%	18%	69.8	70.9	76.9	40.5	543	67.8	64.7	*63.6*

Table 12. **Trade-off between image scales.** PS3 can flexibly adjust token selection ratios at different image scales to achieve the best performance for different downstream tasks.

#HR Patch	Select @756	Select @1512	Chart QA	Doc VQA	V* Bench	Avg† Acc
729	20%	20%	64.1	64.7	60.1	57.6
729	33%	17%	**64.7**	66.5	**60.7**	**58.2**
729	50%	13%	64.4	**67.2**	60.2	58.1
729	67%	8%	63.8	**67.2**	58.5	57.6
729	100%	0%	62.0	64.3	53.3	55.6

Table 13. **Full results of scaling properties of PS3 on 4KPro.**

Vision Encoder	Max Res.	#HR Token	Select (Train)	Select (Test)	Acc
SigLIP	378	0	-	-	35.5
SigLIP-DynRes	756†	784†	-	-	37.1
SigLIP-S^2	756	784	-	-	40.7
PS3	756	729	100%	100%	41.9
SigLIP-DynRes	1512†	3136†	-	-	45.2
SigLIP-S^2	1512	3136	-	-	43.6
PS3	1512	3645	100%	100%	46.8
PS3	3780	1280	6%	6%	48.4
PS3	3780	3840	18%	18%	51.6
PS3	3780	7680	18%	35%	59.8



The image shows a close-up of a motorcycle's rear section. The motorcycle has a maroon-colored fuel tank with a black seat and a black leather saddlebag attached to the side. The motorcycle's license plate is blurred and unreadable. The background consists of green foliage and a blurred road.

The image shows a close-up of a police officer and a motorcyclist. The officer is wearing a high-visibility vest and holding a device, possibly a breathalyzer. The motorcyclist is wearing a helmet and has a motorcycle with visible details such as the engine and exhaust. The background includes trees and foliage.

The second image is a close-up crop from the first image, focusing on the motorcycle rider. The rider is wearing a black helmet and a black jacket with a white stripe. The motorcycle is a classic design with a black frame and red and black body. The background shows a forested area with green trees and some buildings partially visible in the distance.

The image shows a police officer in a high-visibility vest with the letters "ANC" on the back, standing next to a motorcycle rider. The officer is wearing a cap and has a badge on his chest. The motorcycle rider is wearing a helmet and gloves. The background features greenery and a house in the distance.

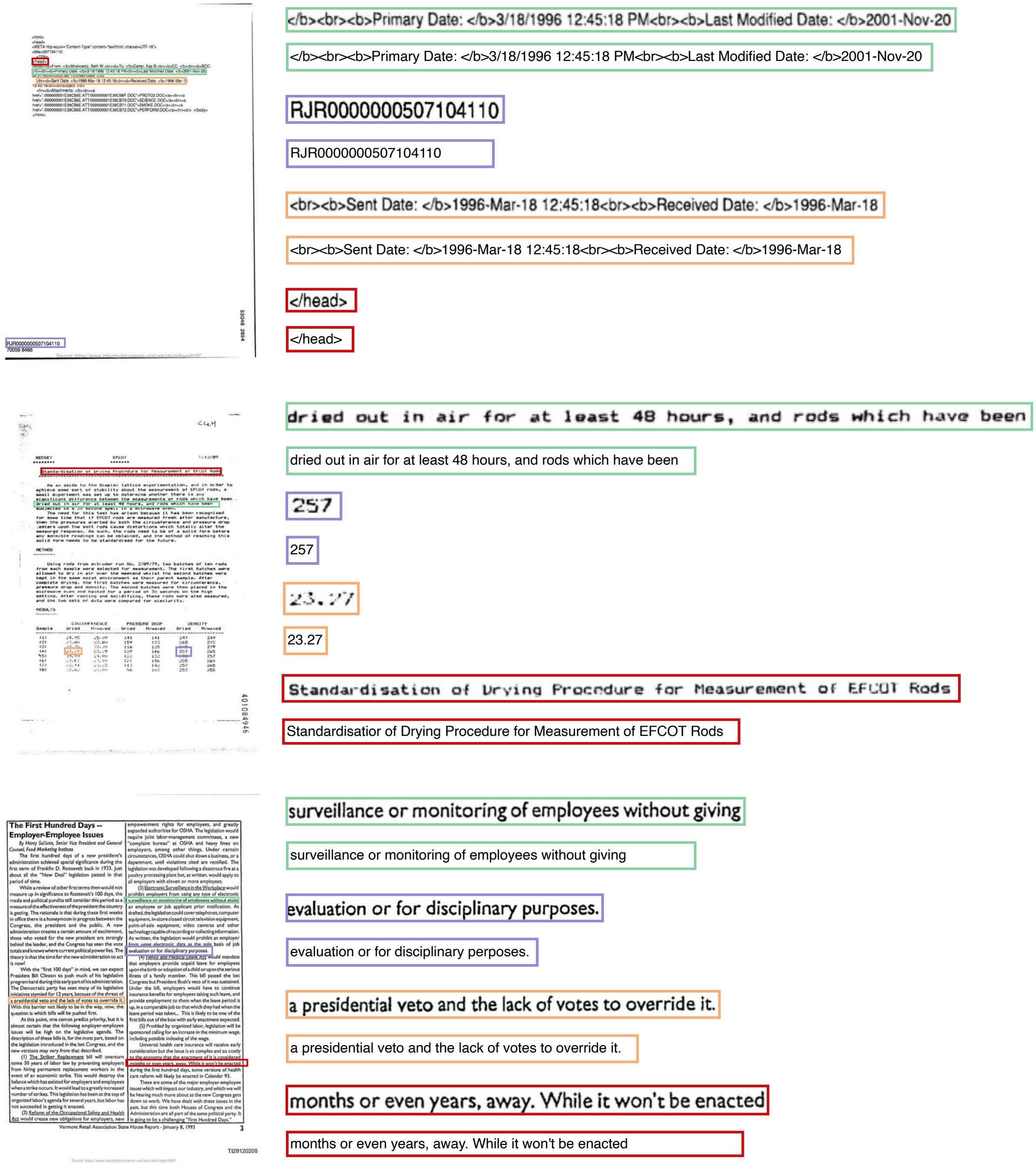
The image shows a vibrant green and orange building with a staircase leading up to a balcony. The building has a door and several lifebuoys hanging on the wall. There are two people sitting on a bench in front of the building, and a bicycle is parked nearby. The text on the building reads "20000".

The image shows a motorcycle parked on a platform with an orange railing. The motorcycle is blue with a black helmet placed on its seat. The background features a green tree and some bushes. The platform appears to be part of a larger structure, possibly a dock or a pier, with a green and orange railing.

The second image shows a portion of the first image, focusing on the area with the boat and the green and orange structure. The boat is a large, colorful vessel with a cabin and a deck, and it is positioned on the water. The green and orange structure appears to be a part of the boat, possibly a cabin or a storage area. The background includes some greenery and a few buildings, indicating a coastal or riverine setting.

The second image is a close-up crop from the first image, focusing on a section of the boat. The boat has a black hull with an orange deck and a green structure on top. There are two Thai flags on the boat, and a small stream of water is visible coming out of the boat. The water is brown and there are green plants in the foreground.

Figure 11. **Examples of pre-training data with natural images.** Here each image is labeled with bounding boxes of four salient regions (highlighted by different colors), together with the local captions of each region. The local captions, generated by Qwen2-VL, contains details in the crops although there are still occasional hallucinations.



</b><br><b>Primary Date: </b>3/18/1996 12:45:18 PM<br><b>Last Modified Date: </b>2001-Nov-20
</b><br><b>Primary Date: </b>3/18/1996 12:45:18 PM<br><b>Last Modified Date: </b>2001-Nov-20
RJR0000000507104110
RJR0000000507104110
<br><b>Sent Date: </b>1996-Mar-18 12:45:18<br><b>Received Date: </b>1996-Mar-18
<br><b>Sent Date: </b>1996-Mar-18 12:45:18<br><b>Received Date: </b>1996-Mar-18
</head>
</head>
dried out in air for at least 48 hours, and rods which have been
dried out in air for at least 48 hours, and rods which have been
257
257
23.27
23.27
Standardisation of Drying Procedure for Measurement of EFCOT Rods
Standardisatior of Drying Procedure for Measurement of EFCOT Rods
surveillance or monitoring of employees without giving
surveillance or monitoring of employees without giving
evaluation or for disciplinary purposes.
evaluation or for disciplinary perposes.
a presidential veto and the lack of votes to override it.
a presidential veto and the lack of votes to override it.
months or even years, away. While it won't be enacted
months or even years, away. While it won't be enacted

Figure 12. **Examples of pre-training data with document images.** Here each image is labeled with four bounding boxes (highlighted by different colors), together with the OCR results as the captions of each region.

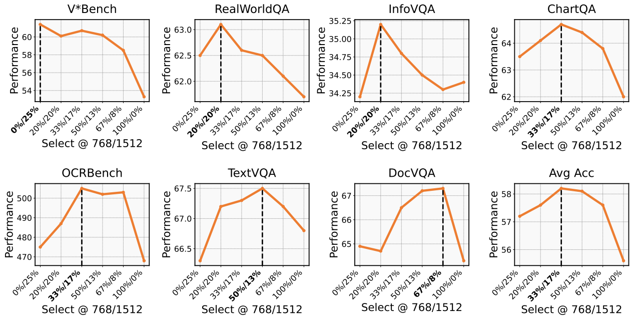
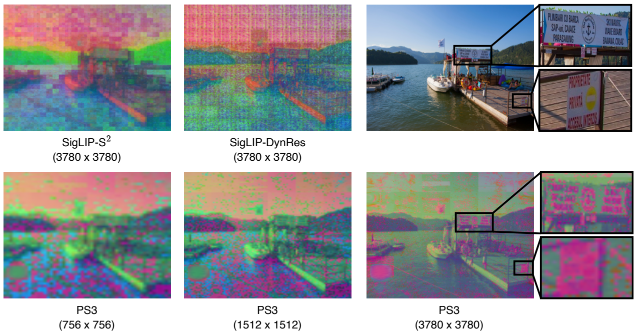


Figure 13. **Trade-off between image scales for different benchmarks.** *Select @ 756/1512* are the percentage of selected patches at 756 and 1512 scales at test time, respectively. PS3 can flexibly adjust token selection ratios at different image scales to achieve the best performance for different downstream tasks.


SigLIP-S^2 (3780 x 3780)

SigLIP-DynRes (3780 x 3780)

PS3 (756 x 756)

PS3 (1512 x 1512)

PS3 (3780 x 3780)

Figure 14. **PCA visualization of visual features.** The baselines, S^2 and AnyRes, have either noisy or blurry features at 4K resolution, while PS3 shows extremely fine-grained features that highlight details such as small texts on the banners.

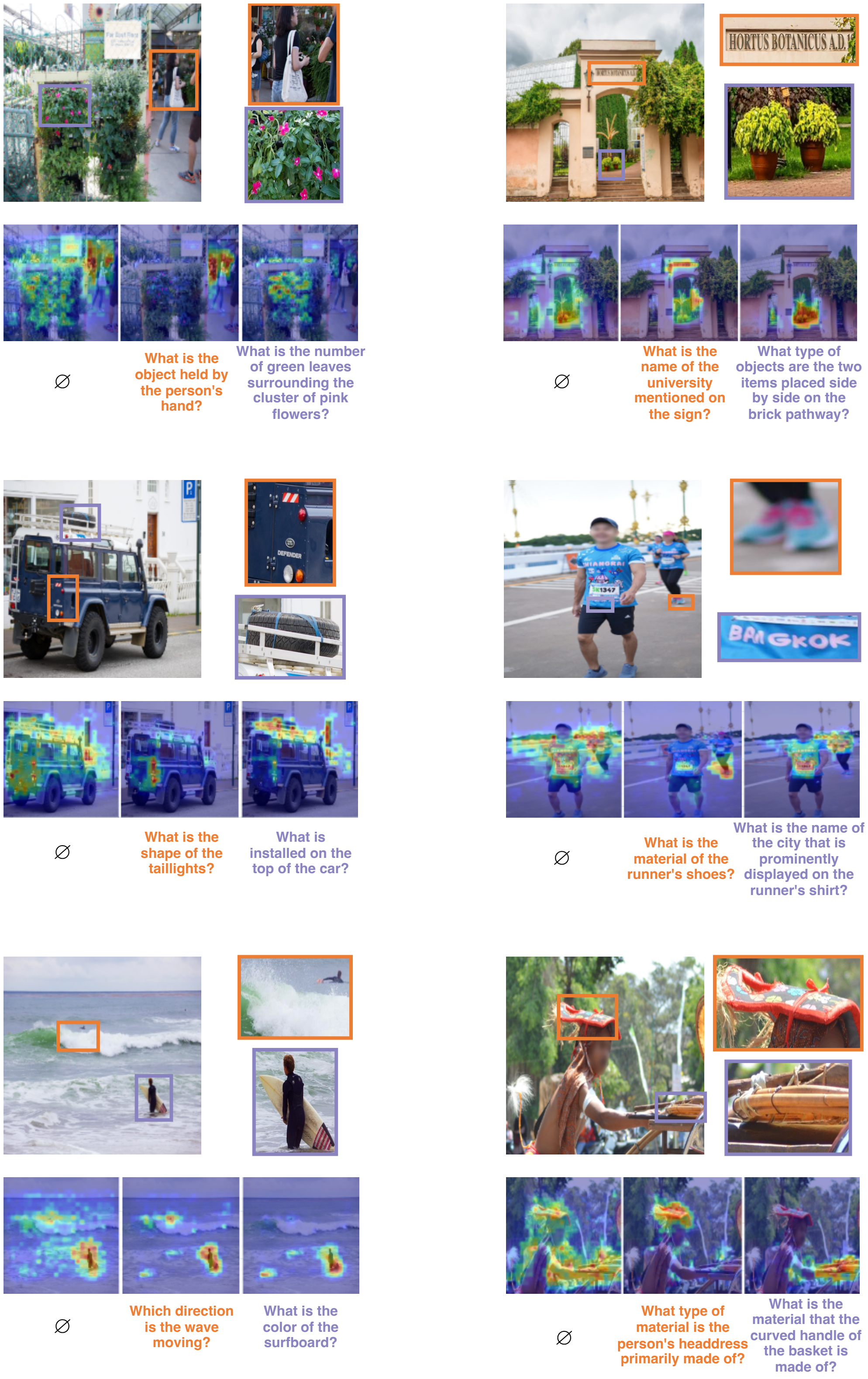
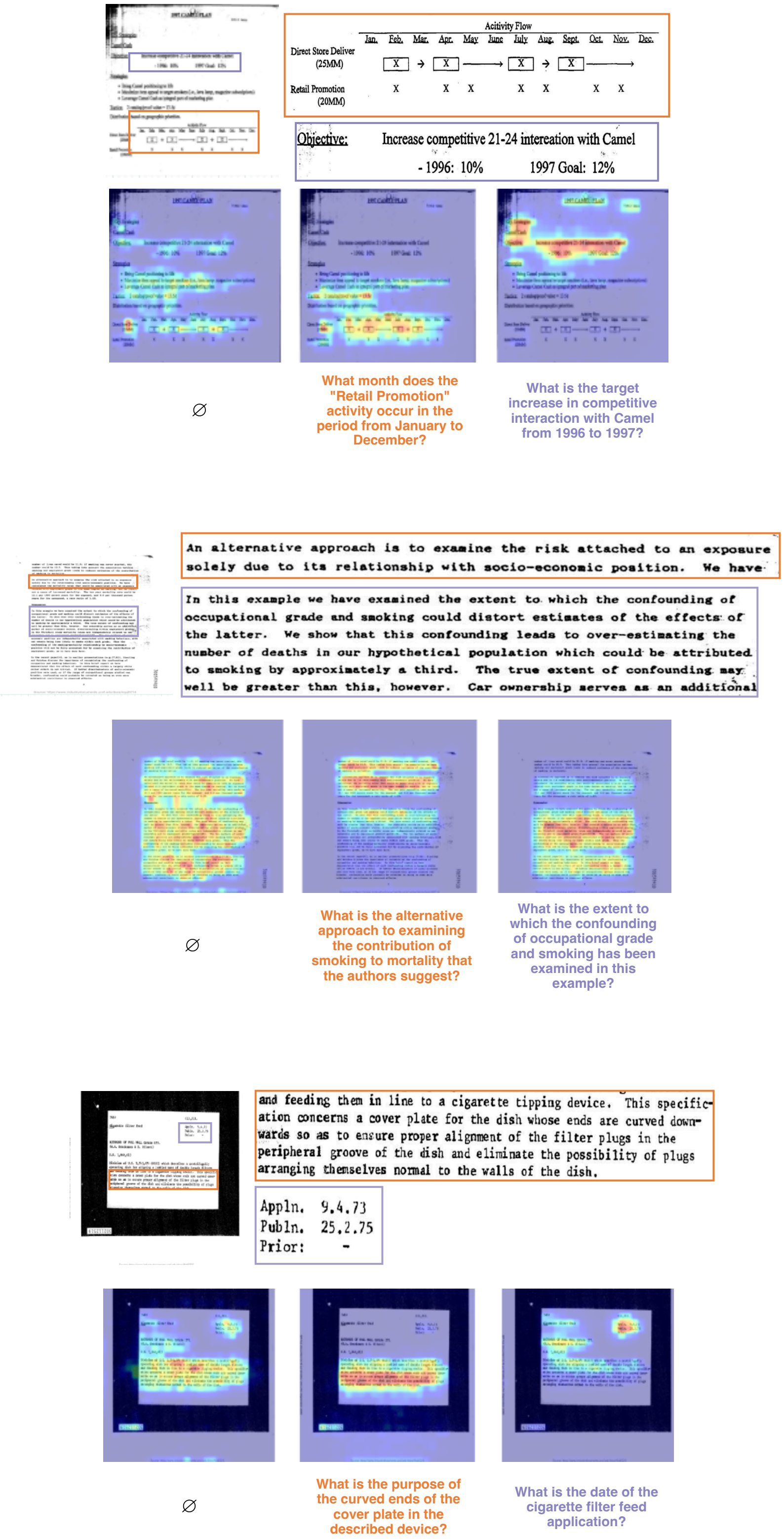

∅
What is the object held by the person's hand?
What is the number of green leaves surrounding the cluster of pink flowers?
HORTUS BOTANICUS A.D.
∅
What is the name of the university mentioned on the sign?
What type of objects are the two items placed side by side on the brick pathway?
∅
What is the shape of the taillights?
What is installed on the top of the car?
BANGKOK
∅
What is the material of the runner's shoes?
What is the name of the city that is prominently displayed on the runner's shirt?
∅
Which direction is the wave moving?
What is the color of the surfboard?
∅
What type of material is the person's headdress primarily made of?
What is the material that the curved handle of the basket is made of?

Figure 15. **Qualitative examples of patch selection on natural images.** PS3 is able to locate different parts of the image that are relevant to the question.



Acitivity Flow

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
Direct Store Deliver (25MM)		X	→	X	⟶		X	→	X	⟶		
Retail Promotion (20MM)	X		X	X		X	X		X	X		

Objective: Increase competitive 21-24 intereation with Camel

- 1996: 10% 1997 Goal: 12%

∅

What month does the "Retail Promotion" activity occur in the period from January to December?

What is the target increase in competitive interaction with Camel from 1996 to 1997?

An alternative approach is to examine the risk attached to an exposure solely due to its relationship with socio-economic position. We have

In this example we have examined the extent to which the confounding of occupational grade and smoking could distort estimates of the effects of the latter. We show that this confounding leads to over-estimating the number of deaths in our hypothetical population which could be attributed to smoking by approximately a third. The true extent of confounding may well be greater than this, however. Car ownership serves as an additional

∅

What is the alternative approach to examining the contribution of smoking to mortality that the authors suggest?

What is the extent to which the confounding of occupational grade and smoking has been examined in this example?

and feeding them in line to a cigarette tipping device. This specification concerns a cover plate for the dish whose ends are curved downwards so as to ensure proper alignment of the filter plugs in the peripheral groove of the dish and eliminate the possibility of plugs arranging themselves normal to the walls of the dish.

Appln. 9.4.73
Publn. 25.2.75
Prior: -

∅

What is the purpose of the curved ends of the cover plate in the described device?

What is the date of the cigarette filter feed application?

Figure 16. **Qualitative examples of patch selection on document images.** PS3 is able to locate different parts of the text in the document based on the question. For example, in the second example, when asked about the alternative approach of examination or the extent to which the confounding is examined, PS3 is able to locate the texts that discuss these topics.

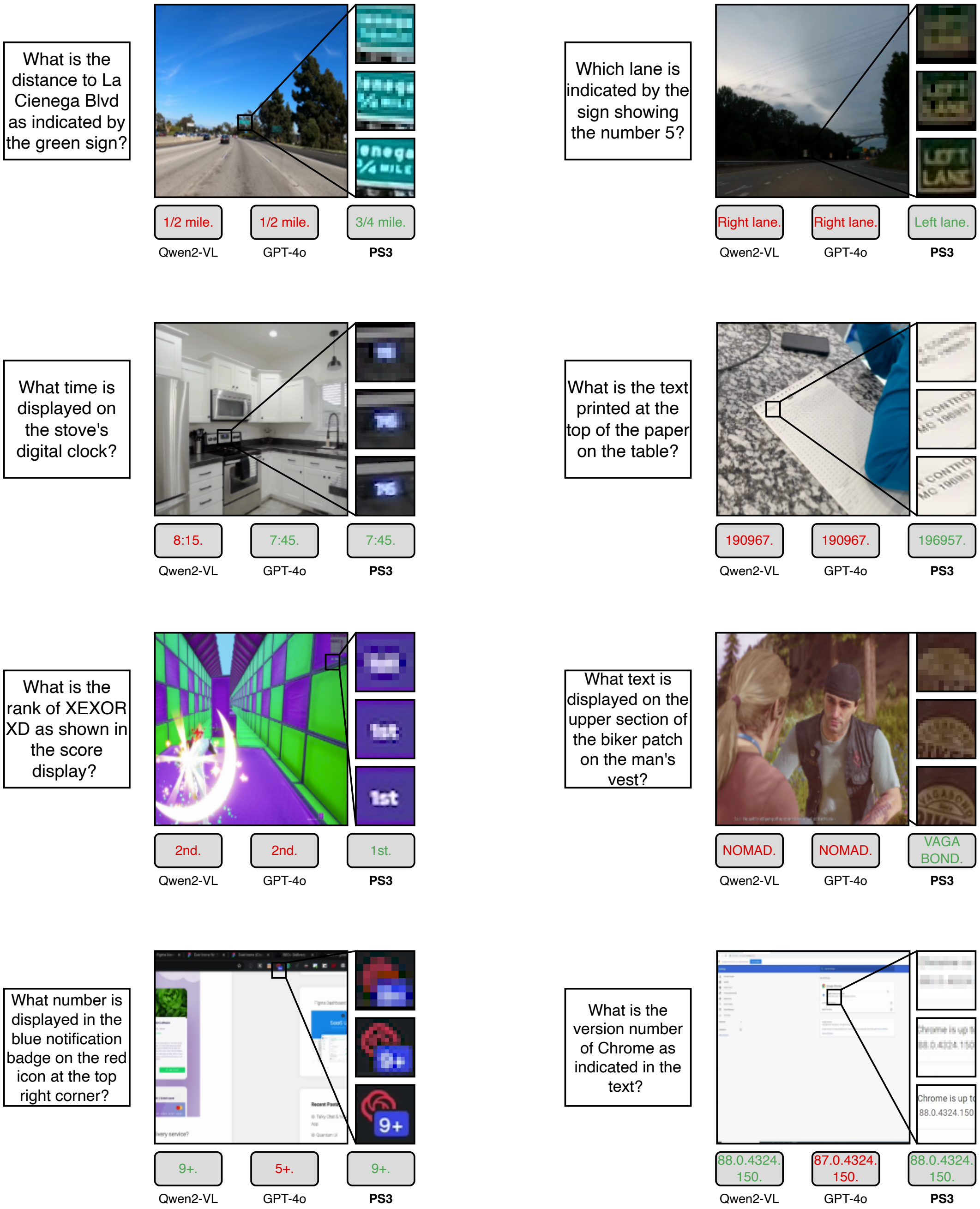

What is the distance to La Cienega Blvd as indicated by the green sign?
1/2 mile.
1/2 mile.
3/4 mile.
Qwen2-VL
GPT-4o
PS3
Which lane is indicated by the sign showing the number 5?
Right lane.
Right lane.
Left lane.
Qwen2-VL
GPT-4o
PS3
What time is displayed on the stove's digital clock?
8:15.
7:45.
7:45.
Qwen2-VL
GPT-4o
PS3
What is the text printed at the top of the paper on the table?
190967.
190967.
196957.
Qwen2-VL
GPT-4o
PS3
What is the rank of XEXOR XD as shown in the score display?
2nd.
2nd.
1st.
Qwen2-VL
GPT-4o
PS3
What text is displayed on the upper section of the biker patch on the man's vest?
NOMAD.
NOMAD.
VAGA BOND.
Qwen2-VL
GPT-4o
PS3
What number is displayed in the blue notification badge on the red icon at the top right corner?
9+.
5+.
9+.
Qwen2-VL
GPT-4o
PS3
What is the version number of Chrome as indicated in the text?
88.0.4324. 150.
87.0.4324. 150.
88.0.4324. 150.
Qwen2-VL
GPT-4o
PS3

Figure 17. **Qualitative examples on 4KPro.** The four rows show examples from categories of autonomous vehicle, household, gaming, and UI understanding, respectively. For each instance, the local crop is shown under the resolutions of 756, 1512, and 3780.