Scaling Vision Pre-Training to 4K Resolution

Baifeng Shi[1,2*] Boyi Li[1,2] Han Cai[2] Yao Lu[2] Sifei Liu[2] Marco Pavone[2]
Jan Kautz[2] Song Han[2] Trevor Darrell[1] Pavlo Molchanov[2] Hongxu Yin[2]
[1]UC Berkeley [2]NVIDIA
https://nvlabs.github.io/PS3

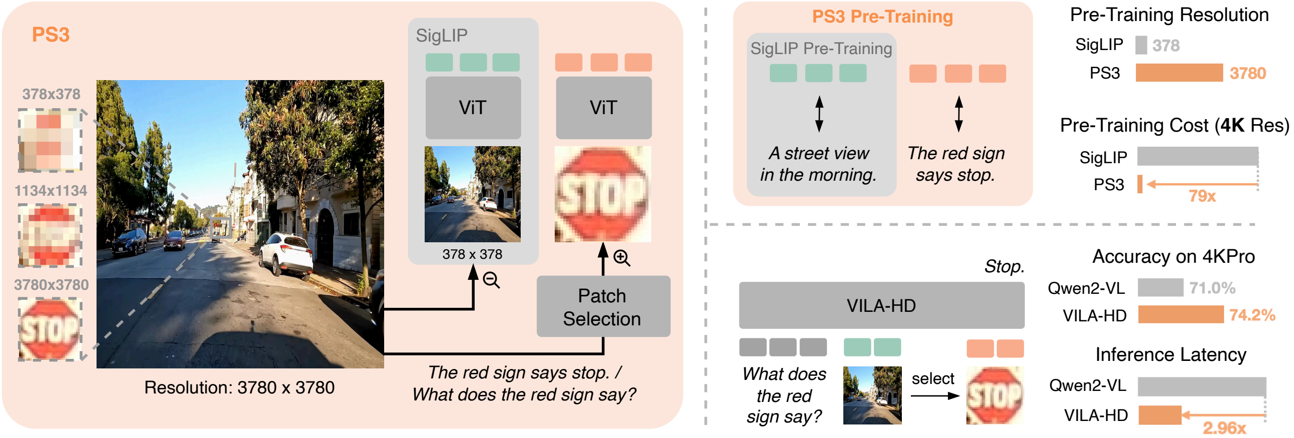


Figure 1. **Left:** Regular vision models such as SigLIP [44] processes images at a low resolution (*e.g.*, 378 × 378 pixels), which is not enough for many daily tasks such as spotting the stop sign while driving. In contrast, PS3 is able to both encode low-res features and efficiently process high-res information of 4K-resolution images via top-down patch selection, *i.e.*, selectively processing relevant patches based on any text prompt. **Top Right:** SigLIP is pre-trained by contrasting global vision features and global captions, which is costly for high-resolution images. PS3 is pre-trained with additional contrast between local high-res features with local captions, enabling pre-training at 4K resolution with 79× less cost than SigLIP. **Bottom Right:** VILA-HD uses PS3 to selectively process high-res regions based on the user prompt, outperforming state-of-the-art MLLMs such as Qwen2-VL [38] on the proposed 4KPro benchmark while achieving 2.96× speedup.

Abstract

High-resolution perception of visual details is crucial for daily tasks. Current vision pre-training, however, is still limited to low resolutions (e.g., 378×378 pixels) due to the quadratic cost of processing larger images. We introduce ***PS3*** *that scales CLIP-style vision pre-training to 4K resolution with a near-constant cost. Instead of contrastive learning on global image representation, PS3 is pre-trained by selectively processing local regions and contrasting them with local detailed captions, enabling high-resolution representation learning with greatly reduced computational overhead. The pre-trained PS3 is able to both encode the global image at low resolution and selectively process local high-resolution regions based on their saliency or relevance to a text prompt. When applying PS3 to multi-modal LLM (MLLM), the resulting model, named* ***VILA-HD****, significantly improves high-resolution visual perception compared to baselines without high-resolution vision pre-training such as AnyRes and S^2 while using up to 4.3× fewer tokens. PS3 also unlocks appealing scaling properties of VILA-HD, including scaling up resolution for free and scaling up test-time compute for better performance. Compared to state of the arts, VILA-HD outperforms previous MLLMs such as NVILA and Qwen2-VL across multiple benchmarks and achieves better efficiency than latest token pruning approaches. Finally, we find current benchmarks do not require 4K-resolution perception, which motivates us to propose* ***4KPro****, a new benchmark of image QA at 4K resolution, on which VILA-HD outperforms all previous MLLMs, including a 14.5% improvement over GPT-4o, and a 3.2% improvement and 2.96× speedup over Qwen2-VL.*

[*] Work done during an internship at NVIDIA.

1. Introduction

Vision models with large-scale pre-training [8, 10, 25, 27] have been the workhorses for both fundamental vision tasks [13, 42] and numerous downstream applications [12, 28, 29]. Notably, CLIP-style vision pre-training (*i.e.*, vision-language contrastive learning) such as CLIP [27] and SigLIP [44] have driven significant advancements in multi-modal large language models (MLLMs) by providing general-purpose language-aligned visual understanding in real-world tasks [15, 17, 21].

However, modern vision models including CLIP and SigLIP have one defect: they are *pre-trained with low resolution only*. Visual perception at high resolution (*e.g.*, 4K resolution) is essential in many real-world scenarios such as spotting the stop sign while driving (Figure 1(Left)). On the other hand, SigLIP, for example, is only pre-trained with a maximum resolution of 378×378 [44], making it incapable of perceiving visual details and thus unsuitable for assisting humans in everyday tasks. Existing methods propose to run pre-trained vision models at higher resolution in a training-free manner for downstream tasks [6, 18, 32]. However, this prevents the model from leveraging large-scale pre-training data to learn high-quality high-resolution perception, resulting in suboptimal performance [32].

What blocks the current vision pre-training from scaling to higher resolution? The computational cost. The compute spent by the vision model grows quadratically for CNNs and quartically for ViTs with increasing resolution, making it even infeasible to pre-train over 1K resolution [25, 44].

In this work, we introduce **P**re-training with **S**cale-**S**elective **S**caling, or **PS3**, that scales CLIP-style pre-training to *4K resolution* with a *near-constant cost*. The key insight is that, instead of contrasting between global images and captions for the whole high-res image, it suffices to contrast between local regions and local captions to learn detailed feature extraction in high-resolution images. For example, in Figure 1(Left, Top Right), to learn to recognize the text on the stop sign, the model only needs to extract the high-resolution feature around the local region of the text and align it with the detailed description about the region. This is analogous to top-down selection mechanism in human vision [3, 48], *i.e.*, one usually focuses on a small portion of the scene that is relevant to the high-level task (*e.g.*, spotting the stop sign). In this way, the model enjoys greatly reduced computational cost by being scale-selective, *i.e.*, selectively processing a small region at fine-grained scale. By disentangling the region size from the image resolution, we are able to scale PS3 pre-training to 4K resolution with a near-constant cost, reducing the pre-training compute by 79× compared to global contrastive learning of SigLIP (Figure 1(Top Right)).

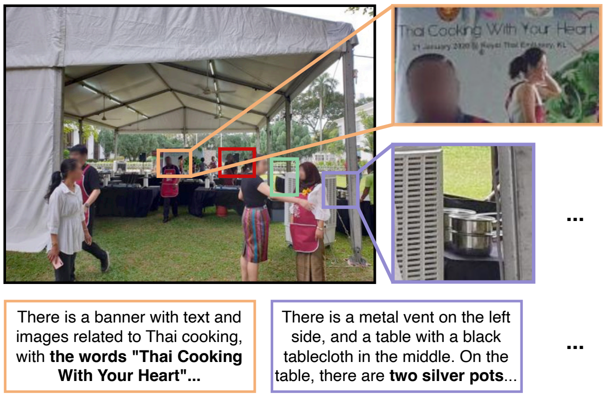
The success of PS3 pre-training hinges on addressing three challenges: *data*, *model*, and *algorithm*. First, since the low-resolution image-text pairs used for CLIP pre-training is not suitable for PS3 pre-training, we collect 75M images with up to 4K resolution and build an automatic pipeline to curate 282M pairs of detailed captions and bounding boxes of salient local regions in the images. Second, we design a vision model that can not only extract low-resolution global features, but also select local patches based on image saliency or text queries and process high-resolution details of the patches. Third, we design an algorithm that pre-trains high-resolution perception through contrastive loss between local regions and local captions and pre-trains patch selection with supervision from the curated bounding boxes.

We show PS3 enables high-quality and efficient high-resolution perception in multi-modal LLMs (MLLMs). Specifically, we train a modern MLLM [21] using pre-trained PS3 as the vision encoder. The resulting MLLM, named **VILA-HD**, is capable of capturing the global image at low resolution and extracting high-resolution details in the local regions selected based on the user prompt. Evaluated on seven benchmarks that require high-res perception, VILA-HD significantly improves the performance over baseline MLLMs that use either the original low-res SigLIP or approaches such as S^2 [32] and AnyRes [6, 18] that scale up the resolution of SigLIP without high-resolution vision pre-training, while using 4.3× fewer tokens compared to the AnyRes baseline. PS3 also unlocks several intriguing scaling properties of VILA-HD, for example, scaling up the resolution without extra cost by selecting a constant number of high-res patches and trading more compute for higher performance at test time by selecting larger high-res regions. We further show in the Appendix that, with a more advanced training recipe, VILA-HD is able to surpass state-of-the-art MLLMs such as NVILA [21] and Qwen2-VL [38] on various benchmarks and achieve superior efficiency and performance over latest token pruning approaches [2, 5, 43].

Despite the superior performance of PS3, we find most existing benchmarks do not actually require 4K resolution. Therefore, we introduce **4KPro**, a benchmark that evaluates visual perception at 4K resolution in four professional use cases including autonomous vehicle, household, gaming, and UI understanding. For each category, 4KPro contains image QA pairs where each question can only be answered under 4K resolution. On 4KPro, PS3 shows a significant improvement of 15% over S^2 and AnyRes baselines and achieves state-of-the-art results compared to both proprietary and open-sourced MLLMs including GPT-4o [11] and Qwen2-VL [38] while being up to 2.96× faster than Qwen2-VL (Figure 1(Bottom Right)).

2. PS3: Vision Pre-Training at 4K Resolution

Based on the paradigm of contrastive language-image pre-training (CLIP) [27] which optimizes a contrastive loss between global images and global captions, we propose



There is a banner with text and images related to Thai cooking, with **the words "Thai Cooking With Your Heart"...**

There is a metal vent on the left side, and a table with a black tablecloth in the middle. On the table, there are **two silver pots**...

Figure 2. **Pre-training data example.** Each instance contains an image with resolution up to 4K, bounding boxes of the salient regions in the image, and captions about details in the regions such as text or small objects.

PS3 which instead optimizes the contrast loss between local regions and detailed captions about the regions (Figure 1(Upper Right)). In this way, the model efficiently learns language-aligned detailed representation by being scale-selective, *i.e.*, only processing the selected regions at a fine-grained scale. This detaches the computational cost from the global image size, allowing us to scale up to ultra-high image resolution during pre-training by controlling the size of the local regions.

The scale-selective pre-training requires a redesign of data, model, and algorithm. We first collect 75M high-resolution images with 282M pairs of bounding boxes and captions of salient local regions (Section 2.1). We then design the architecture of PS3 that can both encode low-resolution global images and select local high-resolution patches to process based on image saliency or their relevance to a text prompt (Section 2.2). We finally pre-train PS3 jointly with localized contrastive loss for high-res perception and box supervision for patch selection (Section 2.3).

2.1. Pre-Training Data of PS3

To learn fine-grained perception in high-res images through contrastive loss between local regions and captions, we need to collect high-res images together with bounding boxes and captions of local regions in each image. We need to make sure the local regions contain rich details in order for the model to learn fine-grained representation. In this work, we collect 75M high-res images with 282M pairs of bounding boxes and local captions for both natural images and document images. Specifically, we propose a pipeline of first detecting the salient regions containing fine-grained details using SAM [13, 47] and then captioning the salient regions using off-the-shelf MLLMs [38]. The detail of data curation approach is explained in the Appendix. An example of the pre-training data is shown in Figure 2, and more examples can be found in the Appendix.

2.2. Model Design of PS3

We design the model such that given a high-res image, it can 1) extract low-res global features, 2) select local regions based on saliency or their relevance to a input text prompt, and 3) extract high-res features of the selected regions. The whole model can be divided into three stages corresponding to these three capabilities respectively, as illustrated in Figure 3. The design of each stage is detailed below.

Stage 1: Low-res feature extraction. We use the same vision transformer (ViT) architecture as SigLIP-SO400M [44] to extract low-res features. The image is resized to 378×378 which corresponds to 27×27 output tokens.

Stage 2: Top-down or bottom-up patch selection. In this stage, the model selects important regions either based on their relevance to a text prompt (*i.e.*, top-down selection) or based on the saliency of the region itself (*i.e.*, bottom-up selection) [3, 48]. See Figure 4(Left) for examples of such selection. To achieve this, the model predicts a selection score for each spatial position of the image by calculating the cosine similarity between the low-res visual features (from Stage 1) and the embedding of the prompt. The prompt embedding is either the text embedding for top-down selection or a constant learnable vector for bottom-up selection, following [31]. The text embedding comes from the text encoder in our contrastive pre-training.

The selection score is calculated with low-res features only, making it infeasible to locate fine-grained details. To alleviate this issue, we predict additional high-res selection score following the same process but with auxiliary high-res features extracted by a light-weight encoder. The light-weight encoder is a ConvNeXt [20] model with only 3 blocks and extracts features at 1512 resolution. The high-res and low-res selection scores are then interpolated to the same size and averaged as the final score.

Stage 3: High-res multi-scale feature extraction. Stage 3 consists of a few key steps. **1) Selecting top-k multi-scale high-res patches.** The model first resizes the high-res image to a set of pre-defined scales and patchifies each. For example, we use three scales of 756×756, 1512×1512, and 3780×3780 for a maximum resolution of 4K. Each is then patchified to 54×54, 108×108, and 270×270 patches, respectively. The selection score from Stage 2 is also interpolated into the same each size. Then for each scale, top-k patches with the highest score are selected. Note that k can vary for different scales. During pre-training, we set k for each scale to be proportional to the total number of patches at that scale. **2) Scale-aware positional embedding.** We add positional embedding for each token by interpolating the original low-res positional embedding and selecting the embeddings that correspond to the selected patches. On top of that, we add a new learnable scale-specific positional embedding to tokens from each scale. **3) High-res feature**

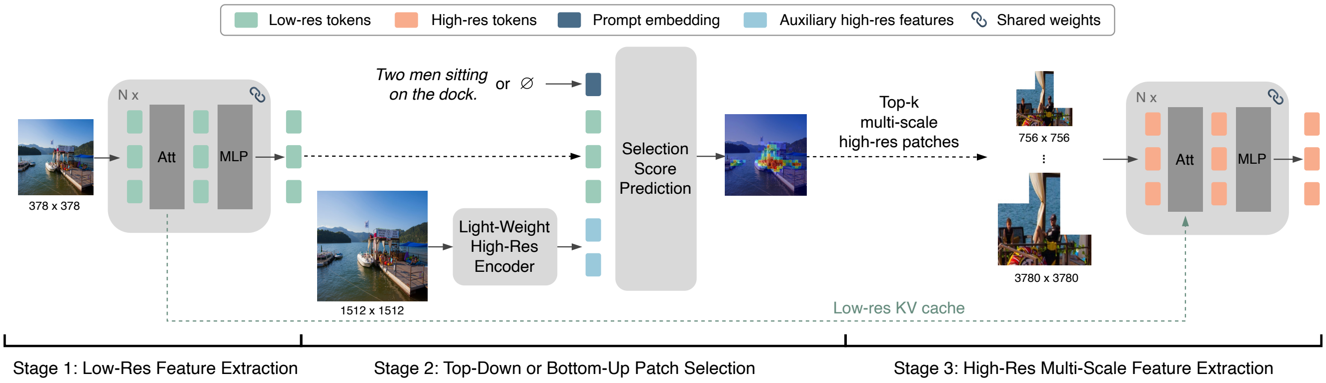

Low-res tokens
High-res tokens
Prompt embedding
Auxiliary high-res features
Shared weights
N x
Att
MLP
378 x 378
Two men sitting on the dock.
or ∅
Selection Score Prediction
Light-Weight High-Res Encoder
1512 x 1512
Top-k multi-scale high-res patches
756 x 756
⋮
3780 x 3780
N x
Att
MLP
Low-res KV cache
Stage 1: Low-Res Feature Extraction
Stage 2: Top-Down or Bottom-Up Patch Selection
Stage 3: High-Res Multi-Scale Feature Extraction

Figure 3. **Model architecture of PS3.** The model consists of 3 stages. In Stage 1, the model encodes global low-resolution features. In Stage 2, based on the low-resolution features as well as auxiliary high-resolution features extracted by a light-weight encoder, the model selects local regions that are either relevant to a text prompt (top-down selection) or salient by themselves (bottom-up selection). In Stage 3, the model processes multi-scale high-res patches from the selected regions with the same encoder from Stage 1. KV cache from the low-res tokens in Stage 1 is added to the self-attention layers to provide a global context for local high-res encoding.

extraction with low-res KV cache. The selected patches from different scales are gathered and simultaneously processed by the same ViT as in Stage 1. To make the local high-res features aware of the global visual context, we augment the K and V in the self-attention layers with the K and V from the corresponding layer in the low-res feature extraction, similar to the KV cache in modern LLMs [26].

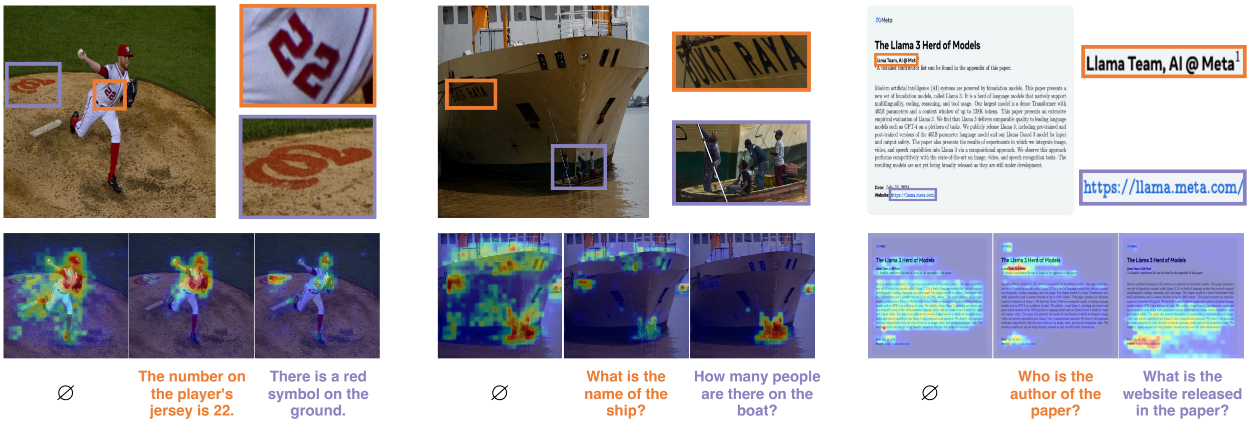
2.3. Pre-Training Algorithm of PS3

PS3 is pre-trained to jointly learn 1) detailed visual representation through localized contrastive loss and 2) top-down and bottom-up patch selection from box supervision.

Learning high-res visual representation. Given the paired data of high-res images and detailed local captions, we use PS3 to extract the high-res features of the local regions as described in Section 2.2, extract text embedding of the local captions using a text encoder, and optimize a contrastive loss between the high-res visual features and the text embeddings. The total number of selected high-res patches for each image is limited to 2560 during pre-training for efficiency, while one can choose to select more tokens for downstream applications (Section 3.1). We use the same sigmoid contrastive loss as in SigLIP [44]. Both the ViT backbone in PS3 and the text encoder are initialized with the pre-trained SigLIP.

There are several key designs in the contrastive pre-training. The effect of each design is studied in Appendix. **1) Using ground-truth selection score for patch selection.** Normally PS3 selects patches based on the local caption. However, in pre-training, to avoid inaccuracy in the selection score predicted by the model which may lead to selecting irrelevant regions, we use selection score generated from the ground-truth bounding box. This is similar to Teacher Forcing [39] in training recurrent neural networks. The ground truth selection score is generated by setting the score inside the box to 1 and others to 0. **2) 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. Pooling both tokens inside and outside the box results in aligning irrelevant visual features to the text embedding in contrastive loss. To avoid this, we constrain the attention pooling to only tokens inside the box. **3) Mixing global and local contrast.** We empirically find that optimizing contrastive loss only between local regions and captions can degrade the quality of global low-res representations. To this end, we mix global and local contrast, *i.e.*, each batch contains both pairs of global low-resolution features and global caption embeddings and pairs of local high-resolution features and local caption embeddings. **4) Avoiding intra-image contrast.** Since we have multiple local boxes 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 visually similar regions of the same image [4]. We make sure each image only appears once in a batch to avoid intra-image contrast.

Learning top-down and bottom-up patch selection. For top-down patch selection, we supervise the selection score predicted from local captions by treating it as a binary semantic segmentation problem. Specifically, the ground truth segmentation map has value 1 inside the ground-truth box and 0 outside. A position-wise cross entropy loss as well as a DICE loss [34] is then optimized between the selection score map and the ground-truth map. For bottom-up selection, since the boxes for each image are already labeled around salient regions, we directly use these boxes to generate ground truth segmentation map and supervise the predicted bottom-up selection score in the same way as above. Figure 4(Left) shows examples of the learned top-down and bottom-up patch selection.



Llama Team, AI @ Meta
https://llama.meta.com/
∅
The number on the player's jersey is 22.
There is a red symbol on the ground.
∅
What is the name of the ship?
How many people are there on the boat?
∅
Who is the author of the paper?
What is the website released in the paper?

Figure 4. **Qualitative examples of patch selection.** *Left*: PS3 can select patches based on image saliency (denoted by ∅) or local captions. *Middle & Right*: VILA-HD with PS3 is fine-tuned to select patches based on questions about local regions.

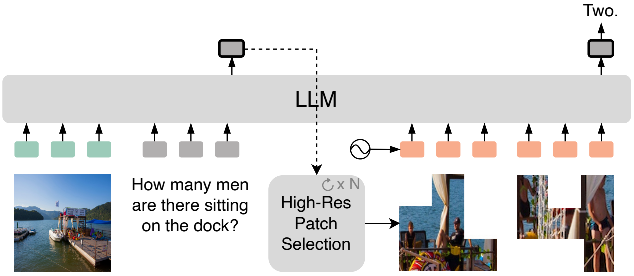


Figure 5. **Model design of VILA-HD.** VILA-HD first extracts the low-res image features using PS3 and sends them along with the text tokens to the LLM. The last-layer embedding of the last token is used to select high-res patches to encode in PS3. VILA-HD can select and encode up to 2560×N high-res patches by running patch selection for N times.

3. VILA-HD: Enabling High-Resolution MLLM with PS3

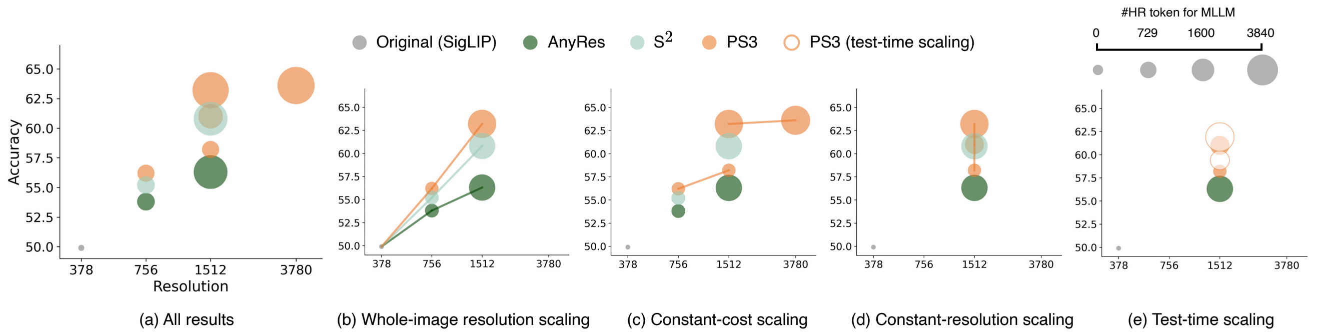
We apply PS3 to MLLMs to enhance their high-resolution perception capability. Specifically, we propose VILA-HD, an MLLM with PS3 as the vision encoder that shows suprior performance and efficiency in processing images of up to 4K resolution. In the following we introduce the model design (Section 3.1) and training recipe (Section 3.2) of VILA-HD.

3.1. Building VILA-HD with PS3

VILA-HD is built based on NVILA [21] but with the vision encoder replaced by PS3. The model design of VILA-HD is illustrated in Figure 5. We first extract the global low-res features following Stage 1 of PS3 and send them along with the text tokens to LLM. We then select high-res patches in either a bottom-up or a top-down way. Bottom-up selection is exactly the same as in pre-training. For top-down selection, since we need to select regions that can help answer the user's question, instead of using the text embedding from the SigLIP text encoder as the prompt, we use the latent embedding of the last token in the user's text input from the last layer of LLM as the prompt embedding. This is inspired by LISA [14] which uses the same embedding for reasoning segmentation. Finally, the selected high-res patches are encoded by Stage 3 of PS3 and sent to LLM after the text tokens, from which the following text generation resumes. We also add an additional positional embedding to the high-res features such that LLM is aware of the spatial positions of the selected patches. Note that while the number of selected high-res patches is limited to 2560 during pre-training, one can select an arbitrary number of patches when applying to MLLMs by running patch selection and high-res feature extraction for multiple times. For example, to select 3840 high-res patches, one can first select the top 2560 patches to process in Stage 3, and then select another top 1280 patches among the rest of the unselected patches and process them in Stage 3.

3.2. Training VILA-HD

The whole model is trained with the normal next-token-prediction loss. We also jointly fine-tune the top-down patch selection since it uses different prompt embeddings from pre-training. To achieve this, we collect data of high-res images paired with *questions* about local regions as well as their bounding boxes. Images and bounding boxes are directly sampled from the pre-training data (Section 2.1) and questions are automatically generated from the local captions using LLaMA-3.1 [9]. During training, low-res image and question are input to LLM, the output last-layer embedding is used for patch selection, and the predicted selection map is supervised by the ground-truth bounding box following the same objective as in Section 2.3. See Figure 4(Middle & Right) for visualization of the fine-tuned patch selection. To better align the high-res features from PS3 to the text space of VILA-HD, we collect high-res QA data and mix it into the training data mixture, which is detailed in Appendix.



(a) All results (b) Whole-image resolution scaling (c) Constant-cost scaling (d) Constant-resolution scaling (e) Test-time scaling

Figure 6. **Scaling properties of PS3.** **(a)** Overall results. We report average performance of VILA-HD with PS3 on seven benchmarks. **(b)** PS3 scales better than baselines without high-res pre-training when all high-res patches are selected. **(c)** PS3 can benefit from processing higher resolution even while selecting a fixed number of high-res patches. **(d)** At a fixed resolution, PS3 trades compute for performance by selecting more high-res patches. **(e)** PS3 can select more high-res patches at test time for better performance.

4. Scaling Properties of PS3

We first evaluate how the performance of PS3 scales with the pre-training resolution. Specifically, we pre-train PS3 at three resolutions of 756, 1512, and 3780. We then train VILA-HD with each PS3 model and evaluate. We show four types of scaling: **1) Whole-image resolution scaling.** PS3 selects all the high-res patches for VILA-HD. This is to compare the high-res feature quality of PS3 with the baselines that also process all high-res patches. **2) Constant-cost scaling.** For PS3 pre-trained at different resolutions, we select a constant number of high-res patches for VILA-HD. This evaluates if performance scales "for free", *i.e.*, by maintaining a constant downstream training and inference cost (note that the pre-training cost is already near-constant). **3) Constant-resolution scaling.** Pre-trained at the same resolution, we select increasingly more high-res patches when training VILA-HD. This evaluates if we gain benefits from selecting more patches in downstream without touching pre-training. **4) Test-time scaling.** Similar to 3), but we increase the number of high-res patches at *test* time. Figure 6 shows the scaling curves and the full results are in Appendix.

Experiment settings. PS3 is initialized with SigLIP-SO400M [44] before pre-training. For evaluation, we report average accuracy on seven resolution-sensitive benchmarks: TextVQA [33], ChartQA [22], DocVQA [23], InfoVQA [24], OCRBench [19], V*Bench [40], and RealWorldQA [41]. We compare PS3 to the original SigLIP as well as two baselines, AnyRes [6, 18] and S^2 [32], that run SigLIP at larger resolution in a training-free way by splitting large images into smaller tiles. See Appendix for the detailed training and evaluation setting.

4.1. Whole-Image Resolution Scaling

Results are shown in Figure 6(b). We can see that the effect of scaling up resolution is significant, where PS3 at 1512 resolution improves 14.2% over SigLIP baseline. Compared to AnyRes and S^2, PS3 shows consistent improvements across different resolution while using a similar number of high-res tokens. For example, at 1512 resolution, PS3 improves by 2.4% over S^2 and 6.9% over AnyRes which is commonly used by modern MLLMs [6, 15]. Since all the methods are processing the whole high-res image, the advantage of PS3 mainly comes from the improved high-res feature quality which is brought by our high-res pre-training.

4.2. Constant-Cost Scaling

Scaling up resolution comes at a cost of quadratically increasing number of tokens in Section 4.1. However, for PS3, higher resolution is still beneficial even when selecting a constant number of patches (Figure 6(c)). For example, selecting 729 (20%) patches with 1512 resolution improves the accuracy by 2% over selecting 729 (100%) patches with 756 resolution. This is because at 756 resolution, not all patches are relevant to the user's questions, and by scaling up the resolution, there are fewer irrelevant 756-resolution patches but more relevant 1512-resolution patches being selected, thanks to our top-down selection mechanism. Comparing to AnyRes, PS3 achieves 58.2% accuracy at resolution of 1512 through constant-cost scaling, improving over AnyRes by 2% accuracy while using 5× fewer tokens. This also enables scaling up 4K resolution for VILA-HD with a constant cost and achieving 63.9% accuracy, improving over S^2 by 3.1% with fewer tokens.

4.3. Constant-Resolution Scaling

Pre-trained at a fixed resolution, PS3 can flexibly select different number of patches for VILA-HD to trade compute for performance. As shown by Figure 6(d), selecting more patches at 1512 resolution consistently improves performance. By increasing the number of patches from 729 (20%) to 1600 (44%), PS3 is able to outperform S^2 with only half the number of high-res tokens. Further increasing to 3645 (100%) patches gives extra 2.2% performance boost.

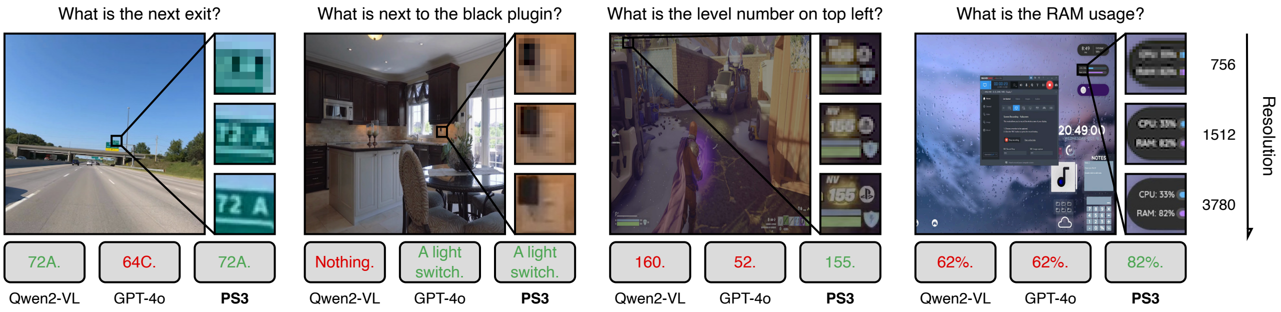


Figure 7. **Examples from 4KPro and comparison of different models**. Each example corresponds to one out of four categories and each question can only be answered without ambiguity under 4K resolution. PS3 improves over the state-of-the-art MLLMs.

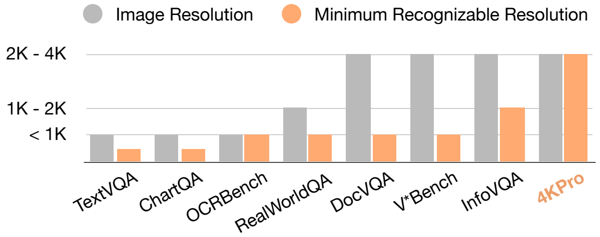


Figure 8. **Image resolution and MRR of different benchmarks.** Existing benchmarks contain high-res images but the resolution required to answer the questions (MRR) is mostly under 1K. In contrast, 4KPro contains questions only solvable at 4K resolution.

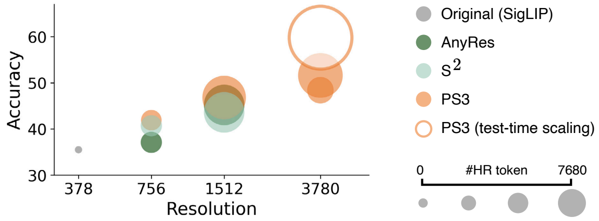


Figure 9. **Scaling properties of PS3 on 4KPro.** PS3 shows consistently improved performance by scaling to 4K resolution and greatly outperforms the baselines.

4.4. Test-Time Scaling

Constant-resolution scaling is still valid at test time, *i.e.*, we select a fixed number of high-res patches during training but select more at test time. As shown in Figure 6, at 1512 resolution, we can train with 20% high-res patches and test with 44% patches, which improves the accuracy by 1.2%. Similarly, training with 40% patches but testing with 100% patches gives an improvement of 0.9%. Note that scaling at test time still performs worse than training time since the MLLM learns better by seeing more patches at training time.

5. 4KPro: Benchmarking PS3 at 4K Resolution

Despite the suprior performance of PS3 on existing benchmarks (Section 4), we find these benchmarks do not actually require high resolution visual perception, especially 4K-resolution perception, even though some of them contain high-resolution images. Specifically, we examine the minimum recognizable resolution (MRR) of the existing benchmarks, *i.e.*, the minimum resolution required to answer the questions. We calculate the MRR by randomly sampling examples from a benchmark, manually checking the minimum resolution (4K, 2K, or 1K) under which the visual details are clear enough to answer each question, and averaging the minimum resolutions of different samples. As shown in Figure 8, even though benchmarks like DocVQA and V*Bench already contain images at 4K resolution, the MRR is mostly around 1K. InfoVQA has the highest MRR of 2K, although it is solely focused on infographic understanding. To effectively evaluate 4K-resolution perception in real-world tasks, we introduce 4KPro which collects image QA pairs with MRR of 4K from four professional use cases including autonomous vehicle, household, gaming, and UI understanding. Each QA pair is in the form of multi-choice problem with four options. Examples of 4KPro are shown in Figure 7. We detail the data curation process in Appendix.

5.1. Main Results

Scaling properties of PS3. We evaluate the performance of VILA-HD when scaling up the resolution of PS3. Results are shown in Figure 9. We can see PS3 outperforms other baselines at resolution of 756 and 1512. While it is infeasible to scale to 4K resolution for the baselines, we are able to train PS3 at 4K resolution by selecting the same number of high-res patches as 1512 resolution. This constant-cost scaling improves the performance by 4.8%. Taking a step further, we can double the number of high-res patches at test time to boost the performance by another 8.2%. On the other direction, we can also shrink the number of patches by 3×, achieving 48.4% accuracy which is 3.2% higher than AnyRes at 1512 resolution while using 2.5× fewer tokens.

Comparison to state of the arts. We compare the performance of VILA-HD with PS3 at 3780 resolution with other proprietary or open-source MLLMs (Table 2). Please see training details in Appendix. The best proprietary MLLMs achieves accuracy of 59.7%. For open-source mod-

Table 1. **Comparing VILA-HD to state-of-the-art MLLMs.** *Res.* is the maximum resolution each model supports. Some models (*e.g.*, Qwen2-VL, InternVL2) can accept input images of different aspect ratios, for which the resolution is calculated as square root of the maximum number of pixels the model can take in. *Select* is the high-res patch selection ratio of PS3 at test time. *#Token* is the total number of visual tokens fed into LLM under the maximum input resolution.

	Res.	Select	#Token	Chart	Doc	Info	Math	MMB	MMMU$_P$	OCR	V*	RWQA	Text	4KPro
VILA-1.5-8B [16]	336	-	576	52.7	40.6	25.9	36.7	68.9	-	-	-	52.7	68.5	33.9
Cambrian-1-8B [37]	1024	-	-	73.3	77.8	-	49.0	75.9	-	624	59.2	64.2	71.7	50.0
NVILA-8B [21]	1552	-	3072	**86.1**	93.7	70.7	65.4	87.6	33.6	794	67.2	66.4	80.1	58.1
MM1.5-7B [45]	2016	-	5184	78.6	88.1	59.5	47.6	-	-	635	-	62.5	76.5	-
LLaVA-OV-7B [15]	2304	-	7252	80.0	87.5	68.8	63.2	80.8	29.5	-	69.2	66.3	-	67.7
IXC2-4KHD [7]	2479	-	7920	81.0	90.0	68.6	57.8	80.2	-	675	-	-	77.2	42.8
IXC-2.5-7B [46]	2743	-	10000	82.2	90.9	70.0	59.6	82.2	-	690	45.6	67.8	78.2	32.3
InternVL2-8B [36]	2833	-	10496	83.3	91.6	74.8	58.3	81.7	32.5	794	65.8	64.4	77.4	58.1
Qwen2-VL-7B [38]	3584	-	16384	83.0	**94.5**	**76.5**	58.2	-	-	**866**	71.0	70.1	**84.3**	71.0
VILA-HD-1.5K-8B	1512	33%	1411	81.3	88.4	58.2	65.3	91.8	35.0	768	67.3	68.4	77.3	50.0
	1512	67%	2626	84.2	91.9	65.3	**66.0**	91.8	**35.1**	776	67.5	68.6	78.0	53.2
	1512	100%	3841	84.3	92.0	67.4	64.6	**92.6**	35.0	782	68.1	68.9	78.4	59.7
VILA-HD-4K-8B	3780	6%	1476	82.2	87.1	57.9	63.9	90.8	34.6	753	68.2	66.5	72.2	62.9
	3780	12%	2756	83.8	91.5	64.5	64.6	91.8	34.7	773	68.8	66.9	77.9	68.8
	3780	18%	4036	84.3	91.7	65.3	64.5	91.8	33.5	774	**71.2**	**70.3**	77.9	**72.6**

Table 2. **Comparing VILA-HD to state-of-the-art MLLMs on 4KPro.** VILA-HD outperforms Qwen2-VL which has the best performance among existing MLLMs while having a lower latency.

Model	Select	Latency	Acc.
GPT-4o [11]	-	-	59.7
Claude 3.5 Sonnet [1]	-	-	29.0
Gemini-1.5-Pro [35]	-	-	59.7
NVILA-8B [21]	-	0.82s	58.1
Cambrian-1-8B [37]	-	2.78s	50.0
InternVL2-8B [36]	-	1.65s	58.1
IXC-2.5-7B [46]	-	2.11s	32.3
LLaVA-OneVision-7B [15]	-	1.75s	67.7
Qwen2-VL-7B-Instruct [38]	-	3.61s	71.0
VILA-HD-4K	18%	1.22s	72.6
VILA-HD-4K	35%	1.91s	**74.2**

els, Qwen2-VL-7B-Instruct performs the best at 71.0% accuracy, although at a cost of larger latency than other models due to processing the full high-resolution images in its vision encoder. On the other hand, VILA-HD-4K (*i.e.*, VILA-HD model with 4K-resolution PS3) achieves 74.2% accuracy at 3780 resolution when selecting 35% patches, which improves over both the proprietary models and the state-of-the-art open-source model (Qwen2-VL) while having a lower latency. By selecting fewer patches (*e.g.*, 18%), VILA-HD-4K still maintains superior performance of 72.6% while enjoying only 1/3 of the latency comparing to Qwen2-VL. See Figure 7 for qualitative examples.

6. Comparison to State of the Arts

In this section, we compare VILA-HD with other state-of-the-art MLLMs. See the experiment settings in Appendix. We also compare the efficiency of PS3 with other token pruning methods [2, 5, 43] and verify the generalizability of PS3 pre-training to different state-of-the-art vision encoders [30, 44] in Appendix.

As shown in Table 1, VILA-HD shows competitive performance compared to state-of-the-art MLLMs such as NVILA and Qwen2-VL and achieves the best results in 6 out of 11 benchmarks. Specifically, VILA-HD-1.5K achieves the best results on benchmarks that have the MRR around 512-1K including MathVista, MMBench, and MMMU-Pro, and VILA-HD-4K obtains state-of-the-art performance on benchmarks that require more detailed understanding (MRR between 1K and 4K) such as V*Bench, RealWorldQA, and 4KPro, surpassing NVILA despite using the same recipe and less data, showing the effect of PS3 pre-training. Note that this is achieved by selecting only 18% of the high-res patches, showing both the efficacy and efficiency of our model. We further show that PS3 can achieve even higher efficiency with only minor performance degradation on most benchmarks. Specifically, by selecting only 6% patches for PS3, it maintains competitive performance of VILA-HD-4K while only using 1476 tokens for a maximum resolution of 4K which is less than 1/10 of #token of Qwen2-VL. We can see the performance stays similar compared to 18% patch selection for benchmarks such as TextVQA and RealWorldQA, and only has minor drops for CharQA and V*Bench.

7. Conclusion

We propose PS3, a scalable CLIP-style vision pre-training method for 4K resolution with near-constant cost. It learns high-res perception via localized contrast by encoding a low-res global image and selectively processing key high-res regions based on image saliency or text prompts. PS3 powers VILA-HD, a high-res MLLM that scales with pre-training resolution and outperforms state-of-the-art MLLMs efficiently. We also introduce 4KPro, a 4K visual perception benchmark, where VILA-HD sets a new state of the art.

References

[1] Anthropic. Claude 3.5 sonnet, 2024. 8

[2] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022. 2, 8

[3] Marisa Carrasco. Visual attention: The past 25 years. *Vision research*, 51(13):1484–1525, 2011. 2, 3

[4] Hong-You Chen, Zhengfeng Lai, Haotian Zhang, Xinze Wang, Marcin Eichner, Keen You, Meng Cao, Bowen Zhang, Yinfei Yang, and Zhe Gan. Contrastive localized language-image pre-training. *arXiv preprint arXiv:2410.02746*, 2024. 4

[5] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pages 19–35. Springer, 2025. 2, 8

[6] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 2, 6

[7] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, et al. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *arXiv preprint arXiv:2404.06512*, 2024. 8

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[9] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 5

[10] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2

[11] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 2, 8

[12] Shubham Juneja, Povilas Daniušis, and Virginijus Marcinkevičius. Dino pre-training for vision-based end-to-end autonomous driving. *arXiv preprint arXiv:2407.10803*, 2024. 2

[13] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 3

[14] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 5

[15] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2, 6, 8

[16] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. *CVPR*, 2024. 8

[17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 2

[18] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 2, 6

[19] Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023. 6

[20] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 3

[21] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. *arXiv preprint arXiv:2412.04468*, 2024. 2, 5, 8

[22] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 6

[23] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 6

[24] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 6

[25] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[26] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 4

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning

transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[28] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. In *Conference on Robot Learning*, pages 416–426. PMLR, 2023. 2

[29] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3, 2022. 2

[30] Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12490–12500, 2024. 8

[31] Baifeng Shi, Trevor Darrell, and Xin Wang. Top-down visual attention from analysis by synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2102–2112, 2023. 3

[32] Baifeng Shi, Ziyang Wu, Maolin Mao, Xin Wang, and Trevor Darrell. When do we not need larger vision models? In *European Conference on Computer Vision*, pages 444–462. Springer, 2025. 2, 6

[33] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 6

[34] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248. Springer, 2017. 4

[35] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 8

[36] OpenGVLab Team. Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy. https://internvl.github.io/blog/2024-07-02-InternVL-2.0/. 8

[37] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 8

[38] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 2, 3, 8

[39] Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989. 4

[40] Penghao Wu and Saining Xie. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094, 2024. 6

[41] x.ai. https://x.ai/blog/grok-1.5v. 6

[42] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 2

[43] Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. *arXiv preprint arXiv:2412.04467*, 2024. 2, 8

[44] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 1, 2, 3, 4, 6, 8

[45] Haotian Zhang, Mingfei Gao, Zhe Gan, Philipp Dufter, Nina Wenzel, Forrest Huang, Dhruti Shah, Xianzhi Du, Bowen Zhang, Yanghao Li, et al. Mm1. 5: Methods, analysis & insights from multimodal llm fine-tuning. *arXiv preprint arXiv:2409.20566*, 2024. 8

[46] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024. 8

[47] Zhuoyang Zhang, Han Cai, and Song Han. Efficientvit-sam: Accelerated segment anything model without performance loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7859–7863, 2024. 3

[48] Li Zhaoping. *Understanding vision: theory, models, and data*. Oxford University Press (UK), 2014. 2, 3