VersaViT: Enhancing MLLM Vision Backbones via Task-Guided Optimization

Yikun Liu[1,2*], Yuan Liu[3†], Shangzhe Di[1,2], Haicheng Wang[1,2], Zhongyin Zhao[3], Le Tian[3], Xiao Zhou[3], Jie Zhou[3], Jiangchao Yao[2], Weidi Xie[1], and Yanfeng Wang[1†]

[1] School of Artificial Intelligence, Shanghai Jiao Tong University
[2] CMIC, Shanghai Jiao Tong University, China
[3] WeChat AI, Tencent Inc., China
yikunliu@sjtu.edu.cn

Abstract. Multimodal Large Language Models (MLLMs) have recently achieved remarkable success in visual-language understanding, demonstrating superior high-level semantic alignment within their vision encoders. An important question thus arises: *Can these encoders serve as versatile vision backbones, capable of reliably performing classic vision-centric tasks as well?* To address the question, we make the following contributions: (i) We identify that the vision encoders within MLLMs exhibit deficiencies in their dense feature representations, as evidenced by their suboptimal performance on dense prediction tasks (*e.g.*, semantic segmentation, depth estimation); (ii) We propose VersaViT, a well-rounded vision transformer that instantiates a novel multi-task framework for collaborative post-training. This framework facilitates the optimization of the vision backbone via lightweight task heads with multi-granularity supervision; (iii) Extensive experiments across various downstream tasks demonstrate the effectiveness of our method, yielding a versatile vision backbone suited for both language-mediated reasoning and pixel-level understanding. The project page is available here.

Keywords: Vision Encoder · MLLM · Multi-task Learning

1 Introduction

Multimodal large language models (MLLMs) have emerged as the prevailing paradigm for vision-language understanding [1, 13, 20, 57]. Architecturally, contemporary MLLMs consist of a vision encoder, a vision–language projection module, and a large language model (LLM). Among these, the vision encoder plays a pivotal role in linking visual perception with language reasoning. In practice, it is commonly initialized from CLIP-style encoders [47, 59] trained with contrastive learning [44], and then jointly optimized with the LLM via instruction tuning on large-scale visual question answering (VQA) and captioning

*Work was done during internship in WeChat.
†Corresponding author.

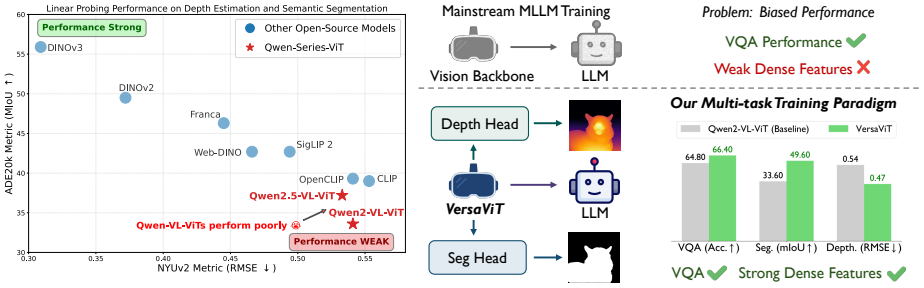

Fig. 1: Overcoming the dense feature limitation of vision backbone within MLLMs. We observe the vision encoder within MLLMs (Left & Top), which yields strong VQA but suboptimal dense features. Conversely, our multi-task collaborative post-training (Bottom) is designed to overcome this limitation by comprehensively enhancing the vision encoder's capabilities.

corpora. This pipeline has yielded state-of-the-art results on a broad spectrum of visual–language understanding tasks.

Despite the excellent performance on instruction following and VQA [1, 20, 39, 71], it remains unclear whether MLLM vision encoders retain the pixel-level perception and spatial precision, that are often required for classic vision-centric problems, for example, segmentation and depth estimation. This question is central if we aim to evolve these encoders into true vision foundation models (VFMs) that are equally competent at language reasoning and pixel-level perception. To probe this, we conduct linear probing on monocular depth estimation and semantic segmentation using the vision backbones of the Qwen-VL series [1, 66]. Our initial results (Section 3) reveal notable deficiencies on dense prediction benchmarks, suggesting a representational misalignment: pretraining and instruction tuning emphasize global semantic alignment and language grounding, but do not sufficiently regularize the fine-grained visual features critical for dense tasks.

This observation motivates our central question: *Can we adapt MLLM vision encoders so that they simultaneously excel at vision–language reasoning and vision-centric dense prediction, yielding a genuinely well-rounded vision foundation model ?* We posit that the gap arises less from architectural limitations and more from the training curriculum and supervision granularity. Specifically, contrastive pretraining followed by text-centric instruction tuning provides limited multi-scale spatial supervision and sparse pixel-level signals, which can ultimately erode fine-grained dense representations.

To this end, we introduce **VersaViT**, by post-training the vision encoder of existing MLLMs with multi-granularity supervision and lightweight task-specific heads. Our framework couples the backbone with auxiliary heads to (i) expose the encoder to complementary signals spanning high-level semantics, mid-level spatial reasoning, and pixel-level perception, while (ii) isolating task-specific conflicts in the heads rather than in the shared representation. Concretely, we instantiate three representative supervision families: (1) VQA and captioning for semantic grounding and instruction-following, (2) monocular depth estima-

tion for 3D-aware spatial structure, and (3) referring segmentation for localized, language-conditioned pixel precision. This framework jointly optimizes the backbone and all auxiliary heads with a simple multi-task objective, biasing the shared representation toward transferable, dense-friendly features without sacrificing language grounding.

We validate **VersaViT** across a wide range of downstream tasks, including VQA, semantic segmentation, referring segmentation, monocular depth estimation, and image–text retrieval. Empirically, VersaViT improves VQA performance while substantially improving dense prediction accuracy, indicating that multi-granularity signals can be synergistic rather than competitive under our framework. Moreover, our framework is modular and aligns with MLLM training pipelines, enabling drop-in integration with minimal engineering overhead.

In summary, we make the following contribution: (i) diagnosis. We identify a consistent shortfall in dense representations of mainstream MLLM vision encoders, manifested by underperformance on depth and segmentation under linear probing; (ii) method. We propose VersaViT, a well-rounded vision transformer that instantiates a multi-task collaborative post-training framework. This framework augments an MLLM vision encoder with lightweight auxiliary heads for VQA/captioning, depth estimation, and referring segmentation, delivering multi-granularity supervision to the shared backbone; (iii) evidence. Extensive experiments show that VersaViT simultaneously advances VQA and dense prediction, yielding a more balanced, transferable vision backbone suited for both language-mediated reasoning and pixel-level understanding.

2 Related Work

Vision Foundation Models. Developing a robust foundation model for computer vision remains a fundamental challenge within the field. Various approaches have been adopted to advance the development of vision foundation models, which can be broadly categorized into two scalable learning paradigms: self-supervised and weakly-supervised learning. Recent self-supervised methods, including MAE [21], iBOT [88], and the DINO series [7, 45, 54], mainly use large-scale curated image datasets to construct suitable proxy tasks for learning generalizable visual representations. In contrast, weakly-supervised learning leverages web-scale image-text pairs for training. This paradigm includes models such as CLIP [47], ALIGN [24], and SigLIP [84], which utilize contrastive learning [44], as well as more recent models like CapPa [60], LocCa [62] and OpenVision2 [36], together with the Qwen-VL series [1, 66] and other MLLMs [20, 57, 76, 81], which rely on autoregressive loss for training from scratch or for fine-tuning the vision encoder to achieve improved visual representations.

Recent studies have also explored vision encoder enhancement from complementary perspectives. DUNE [50] and UNIC [51] rely on distillation from frozen vision teachers for vision-only tasks; Perception Encoder [5] employs decoupled backbones to separately address language and spatial alignment; and [55] investigates reinforcement learning as an alternative to supervised fine-tuning

for MLLM vision encoders. In contrast, our work focuses on task-guided post-training of MLLM vision encoders while retaining a single unified backbone. This makes our approach complementary to teacher distillation, decoupled encoder designs, and RL-based optimization.

Multimodal Large Language Models. With recent advancements in LLMs [6, 25, 58], increasing attention has been directed toward MLLMs, which aim to align visual and textual modalities via visual instruction tuning. Recent studies [1, 20, 57, 63, 66] either seek to enhance the performance of MLLMs across various VQA benchmarks by improving data scaling and selection, architectural design, and training strategies, or leverage the powerful capabilities of MLLMs to repurpose them for other visual-language tasks [31,32,37,64], such as multimodal retrieval, temporal grounding, and image referring segmentation. However, the intrinsic quality and characteristics of the visual features learned by the MLLM's vision encoder remain largely unexplored and unaddressed. Therefore, we aim to investigate this critical issue and further leverage these insights to enhance the vision encoder's representation.

Multi-task Learning. Multi-task learning [8, 12, 14, 61, 73, 74, 77, 89] enables models to generalize across diverse tasks, fostering the development of versatile systems that no longer rely on task-specific architectures. In natural language processing (NLP), this paradigm has been successfully realized by framing text-related tasks within a sequence-to-sequence formulation, which has facilitated the emergence of LLMs [6,58]. Inspired by this idea, the multimodal domain has also witnessed a surge of research on unified multimodal models [22, 49, 65, 69, 75]. These approaches typically reformulate multimodal tasks, which span from image captioning to visual question answering, as sequence-to-sequence problems or use LLMs to connect different task heads. Motivated by these advances, we explore joint training of multiple vision tasks to enhance the vision backbone.

3 Motivation

Recent vision encoders developed under the MLLM paradigm have demonstrated strong performance on visual question answering (VQA) and other language-centered vision tasks. However, a practical vision foundation model (VFM) should be broadly useful beyond dialogue, for example, to support perception systems that demand spatial precision, geometric consistency. This motivates a central question: *Do current MLLM vision encoders serve as effective VFMs, or do they exhibit systematic deficiencies when transferred to classic vision-centric tasks?* To investigate this question, we conduct a linear probing evaluation of two representative encoders, Qwen2-VL-ViT [66] and Qwen2.5-VL-ViT [1], on monocular depth estimation and semantic segmentation. As shown in Figure 1, both encoders underperform substantially under this setting. These results reveal a representational skew induced by prevailing MLLM training recipes, which emphasize global semantic alignment while insufficiently preserving intermediate geometric cues and fine-grained visual structures.

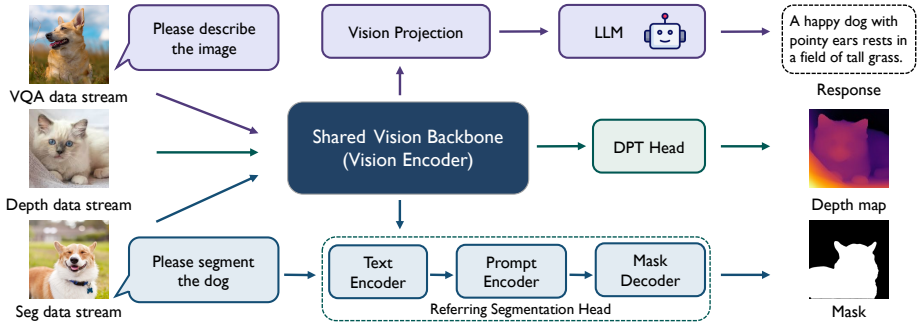

Fig. 2: Overview of the proposed multi-task collaborative training framework. The proposed framework jointly trains three distinct tasks: VQA and Image Captioning, Monocular Depth Estimation, and Image Referring Segmentation. By incorporating lightweight task heads, this collaborative training strategy is designed to enhance the representational capabilities of the underlying vision backbone.

Guided by this observation, we further ask: *Can MLLM vision encoders be rebalanced toward dense-friendly representations through low-cost post-training, without compromising their language grounding?* We hypothesize that multi-granularity supervision can mitigate the observed skew by jointly reinforcing high-level semantic grounding, mid-level spatial reasoning, and pixel-level localization. Concretely, we combine supervision from VQA/captioning, monocular depth estimation, and referring segmentation to cover these complementary granularities. We instantiate this idea in VersaViT, a multi-task post-training framework that attaches lightweight task-specific heads to a shared vision backbone. These heads provide task-targeted gradients and help absorb task conflicts, while the shared backbone is gradually nudged toward representations that remain instruction-competent yet become more effective for dense prediction. The resulting design prioritizes practicality: it is modular, compatible with standard MLLM pipelines, and avoids full-scale retraining.

In summary, our motivation is empirical and application-driven. Although current MLLM vision encoders excel at VQA, they do not reliably transfer to dense prediction tasks. We show that low-cost multi-granularity post-training can close this gap, producing a more balanced and broadly useful VFM.

4 Method

This section starts by formulating our considered problem in Section 4.1, then elaborates on the overall multi-task collaborative post-training framework in Section 4.2, and finally details the training strategy of VersaViT in Section 4.3.

4.1 Problem Formulation

Our goal is to strengthen the MLLM vision backbone so that its representations support both language-mediated reasoning and dense, pixel-level perception. To

this end, we propose VersaViT, a versatile vision backbone that is optimized by a multi-task post-training framework. This framework augments a shared vision encoder with lightweight, task-specific heads, enabling multi-granularity supervision with minimal architectural change. As shown in Figure 2, we select three representative tasks: (i) VQA and image captioning for semantic alignment with language, (ii) monocular depth estimation for 3D-aware spatial reasoning, and (iii) referring segmentation for localized, language-conditioned perception.

Our framework uses separate data streams for each task. To simplify, we illustrate it with a single image for all tasks. Given an input image $\boldsymbol{I} \in \mathbb{R}^{H\times W\times 3}$, the shared vision encoder (Φ_V, parameterized by $\boldsymbol{\theta}_V$) first extracts visual features ($\{\mathcal{F}_i^V\}_{i=1}^N$), where N denotes the number of the transformer layers in the encoder. The extracted features are subsequently routed to task-specific heads with parameters $\boldsymbol{\theta}_{\text{heads}}$. The resulting respective task losses ($\mathcal{L}_{\text{cap}}$, $\mathcal{L}_{\text{depth}}$, $\mathcal{L}_{\text{seg}}$) are aggregated to form a combined objective $\mathcal{L}_{\text{all}}$, which is backpropagated to collaboratively optimize and strengthen the shared vision backbone (Φ_V) across all objectives. The overall optimization goal is to find the optimal parameters for both the backbone and the heads by minimizing the combined loss:

$$\min_{\boldsymbol{\theta}_V,\boldsymbol{\theta}_{\text{heads}}} \mathcal{L}_{\text{all}} = \sum_{\tau\in\mathcal{T}} \lambda_\tau \cdot \mathcal{L}_\tau(\boldsymbol{\theta}_V, \boldsymbol{\theta}_\tau),$$

where $\mathcal{T} = \{\texttt{cap}, \texttt{depth}, \texttt{seg}\}$ is the set of all tasks, λ_τ is the task weight hyperparameter, and $\boldsymbol{\theta}_\tau$ represents the parameters for the head of task τ.

4.2 Overall Framework

In this section, we provide a detailed breakdown of our proposed multi-task collaborative training framework by successively introducing the three auxiliary training tasks: VQA and image captioning, monocular depth estimation, and image referring segmentation.

Generative Vision-Language Fine-Tuning. To enhance the high-level semantic understanding of the vision backbone, the combined VQA and image captioning tasks are integrated. This section details the components and training for the VQA&image captioning tasks.

(1) Core Components. To accomplish the tasks of VQA and image captioning, we adopt the mainstream MLLM architecture [35]. Specifically, our main components consist of a vision encoder (`Qwen2-VL-ViT`), a vision projector (an MLP Layer), and a large language model (`Qwen3-1.7B`).

(2) Training Pipeline. For the VQA and image captioning task, we utilize the final layer feature ($\mathcal{F}_N^V$) from the vision encoder. Given an input image $\boldsymbol{I}$ and an associated text $\boldsymbol{t} = (t_0, t_1, \cdots, t_L)$ with the token length of L, the visual feature ($\mathcal{F}_N^V$) is first passed through a vision projector to obtain the projected image embeddings. These projected image embeddings are then concatenated with the input text embedding derived from $\boldsymbol{t}$ via a text embedding layer. The resulting sequence forms the full input for the LLMs. The training objective are as follows: $\mathcal{L}_{\text{cap}} = \frac{1}{L}\sum_{i=1}^{L} \mathcal{L}_{\text{ce}}(t_i|\boldsymbol{I}, t_0, \cdots, t_{i-1})$.

(3) Captioning Alignment Warm-Up. Prior to the multi-task collaborative training phase, we employ the captioning task to warm up the vision projection. Crucially, during this stage, both the vision backbone and the LLM are frozen; we exclusively train the vision projection using autoregressive loss ($\mathcal{L}_{\text{cap}}$). For training data, we incorporate a variety of sources, including natural image-text pairs (*e.g.*, pixmo-cap [13] and SA-1B-InternVL [29]) and optical character recognition (OCR) data (*e.g.*, idl-wds [4]).

Monocular Depth Estimation. To address the feature deficiency in spatial awareness and improve geometric representation, the monocular depth estimation task is incorporated. This section introduces the methodology for obtaining the depth prediction and corresponding pseudo ground truth label for model training, as well as the training objective employed.

(1) Depth Prediction. For monocular depth estimation, we select a uniform subset of features from the extracted visual features ($\{\mathcal{F}_i^V\}_{i=1}^N$). Specifically, we uniformly sample three of these features and feed them into a DPT head [48] to generate the final depth prediction ($d^* \in \mathbb{R}^{H\times W}$).

(2) Pseudo Depth Labeling. We do not rely on manually annotated ground truth data. Instead, we generate depth labels via pseudo-labeling. Specifically, we adopt the Depth Anything V2 model [78] to annotate the images to get the depth labels ($d \in \mathbb{R}^{H\times W}$). Data source details are provided in Table 1.

(3) Training Objective. We use a scale- and shift-invariant loss $\mathcal{L}_{\text{ssi}}$ and a gradient matching loss $\mathcal{L}_{\text{gm}}$ to optimize depth estimation. Specifically, $\mathcal{L}_{\text{ssi}} = \frac{1}{HW}\sum_{i=1}^{HW}\rho(d_i^*, d_i)$, where ρ is the affine-invariant mean absolute error loss. $\mathcal{L}_{\text{gm}}$ evaluate the gradient difference between the ground truth and the rescaled estimates on multiple scales k:

$$\mathcal{L}_{\text{gm}} = \sum_{k=1}^{K}\sum_{i=1}^{HW}|s\nabla_x^k d_i - \nabla_x^k d_i^*| + |s\nabla_y^k d_i - \nabla_y^k d_i^*|.$$

The final loss is: $\mathcal{L}_{\text{depth}} = \lambda_{\text{ssi}}\cdot\mathcal{L}_{\text{ssi}} + \lambda_{\text{gm}}\cdot\mathcal{L}_{\text{gm}}$, where λ_{ssi} and λ_{gm} are weight hyperparameters. Please refer to the supplementary materials for more details.

Image Referring Segmentation. To instill robust localization and pixel-level fine-grained perception into the visual representation, the image referring segmentation task is employed. This section outlines the way to get the mask prediction, the data scaling strategy, and the corresponding training objective.

(1) Mask Prediction. Given an input image $\boldsymbol{I}$ and an associated text prompt $\boldsymbol{t}$, we use the final layer feature ($\mathcal{F}_N^V$) from the vision encoder. Subsequently, we utilize a pretrained text embedding model (`Qwen3-Embedding-8B [85]`) as the text encoder to offline extract the corresponding text embedding. Furthermore, this text embedding is then processed by the prompt encoder to yield the prompt embedding ($\mathcal{F}^P$). Ultimately, the mask decoder efficiently mapping $\mathcal{F}_N^V$, $\mathcal{F}^P$, and an output token to generate the final mask ($\hat{M}$). For further details about the architecture of the prompt encoder and mask decoder, please refer to [29].

(2) Segmentation Data Construction. To efficiently scale the segmentation data, which is structured as an image $\boldsymbol{I}$ associated with a set of P instance-level

Table 1: Statistics of the training data.

Training Stage	Task	Data Source	Data Size
Captioning Alignment	Image Captioning	Pixmo-cap [13]	0.6M
		IDL-WDS [4]	4.7M
		SA-1B-InternVL [29]	4.1M
Multi-task Collaborative Training	Depth Estimation	Open-images [30]	1.7M
		Objects365 [52]	1.7M
		Google-landmark [67]	2.7M
		Places365 [87]	8.0M
		SA-1B [29]	11.0M
		CC3M [9]	2.8M
	Referring Segmentation	GRIT [46]	13.9M
		Objects365 [52]	1.6M
		SA-1B-DAM [33]	0.6M
	VQA&Image Captioning	FineVision [68]	14.0M

pairs $\{(\boldsymbol{t}_i, M_i)\}_{i=1}^{P}$ (where $\boldsymbol{t}_i$ is the text description and M_i is the corresponding mask), we employ the following three strategies: (i) We directly integrate existing referring segmentation datasets. (ii) For datasets already possessing mask annotations (*e.g.*, SA-1B), we leverage a dense captioning model (*e.g.*, Describe Anything [33]) to generate a text annotation for each mask. (iii) For datasets annotated only with bounding boxes and class labels, we first use SAM [29] to generate precise masks based on the bounding boxes, and subsequently apply the Describe Anything or the existing class label for text description annotation. For detailed data regarding segmentation, please refer to Table 1.

(3) Training Objective. We utilize a combination of per-pixel binary cross-entropy (BCE) loss and DICE loss, with corresponding loss weights λ_{bce} and λ_{dice}. Given the ground-truth target masks (M), the overall loss is formulated as $\mathcal{L}_{\text{seg}} = \lambda_{\text{bce}} \cdot \mathcal{L}_{\text{bce}}(\hat{M}, M) + \lambda_{\text{dice}} \cdot \mathcal{L}_{\text{dice}}(\hat{M}, M)$.

4.3 Model Training

In this section, we describe the details of our multi-task collaborative training, which uses diverse objectives to advance the capabilities of the vision backbone.

Multi-task Collaborative Training. After alignment, we adopt the multi-task collaborative training schema involving the three objectives detailed in Section 4.2: VQA&captioning ($\mathcal{L}_{\text{cap}}$), monocular depth estimation ($\mathcal{L}_{\text{depth}}$), and image referring segmentation ($\mathcal{L}_{\text{seg}}$). During this phase of training, the weights of the vision encoder, the task-specific heads, and the LLM are all unfrozen. We perform joint optimization via alternating batch sampling combined with gradient accumulation. Specifically, we alternately sample a batch of data from each of the three tasks for a forward pass. The loss of each task is used to compute and accumulate gradients, with the collective parameter update executed after iterating through all three tasks. The final loss is a weighted sum:

$$\mathcal{L}_{\text{all}} = \lambda_{\text{cap}} \cdot \mathcal{L}_{\text{cap}} + \lambda_{\text{depth}} \cdot \mathcal{L}_{\text{depth}} + \lambda_{\text{seg}} \cdot \mathcal{L}_{\text{seg}},$$

Table 2: Performance on OpenCompass benchmarks, a collection of eight general VQA benchmarks. We assess our method on the following benchmarks, MMBench (MMB) [38], MMstar [10], MMMU [83], MathVista [41], Hallusion-Bench (Hallu.) [18], AI2D [28], OCRBench (OCR.) [40], MMVet [82]. For details on our VQA training and data, please refer to the supplementary materials.

Method	MMB	MMStar	MMMU	MathVista	Hallu.	AI2D	OCR.	MMVet	Avg.
Open-Source Models (without access to the training data)									
MiniCPM-V-2.6 [80]	78.0	57.5	49.8	60.8	48.1	82.1	85.2	60.0	65.2
DeepSeek-VL2 [70]	81.2	61.9	54.0	63.9	45.3	83.8	80.9	60.0	66.4
Qwen2-VL-7B [66]	81.0	60.7	53.7	61.6	50.4	83.0	84.3	61.8	67.1
Qwen2.5-VL-7B [1]	82.2	64.1	58.0	68.1	51.9	84.3	88.8	69.7	70.9
Ours (LLM = Qwen3-8B)									
Qwen2-VL-ViT [66]	78.2	59.3	51.0	62.1	49.0	79.4	80.6	58.7	64.8
VersaViT	78.0	60.9	53.1	63.7	51.3	79.1	82.2	62.6	66.4
Absolute Gain △	(-0.2)	(+1.6)	(+2.1)	(+1.6)	(+2.3)	(-0.3)	(+1.6)	(+3.9)	(+1.6)

where λ_{cap}, λ_{depth} and λ_{seg} are weight hyperparameters. Jointly optimizing these objectives enables the vision backbone to capture high-level semantics and strengthen pixel-level and spatial-aware representations.

Discussions. The core novelty of VersaViT lies in **rethinking the training paradigm** rather than seeking **complex architectural design**. VersaViT diverges from standard multi-task learning by using task-specific gradients to **shape representation** rather than merely maximize metrics. We argue that the "blindness" of MLLM encoders to dense tasks stems from representational skew, not capacity limits. By employing complementary objectives as **regularizers** within a decoupled head-backbone framework, we force the backbone to reclaim fine-grained visual evidence while buffering task conflicts. This yields a versatile VFM that balances high-level semantics with pixel-level precision.

5 Experiments

5.1 Experimental Setup

Training Datasets. Our training is supported by a diverse dataset from multiple sources. Specifically, the captioning alignment stage mainly uses high-quality image-caption and OCR data. Subsequently, during the multi-task collaborative training phase, we either collect or synthesize new data tailored to each respective task. The comprehensive details of the training data are presented in Table 1.

Evaluation Settings. To comprehensively assess the performance of VersaViT, we conduct evaluations across multiple downstream tasks. (i) VQA performance: To evaluate VersaViT's performance on VQA, we adopt an MLLM evaluation strategy. Specifically, we integrate VersaViT with a vision projection layer and an LLM (`Qwen3-8B`). The training is performed in two stages: in the first stage, only the vision projection layer is trained; in the second stage, both the vision projection layer and the LLM are trained. Throughout this entire training process, the parameters of the vision encoder are kept frozen. For details regarding

Table 3: Linear probing results on semantic segmentation and monocular depth estimation with frozen backbones. We report the mean Intersection-over-Union (mIoU) metric for ADE20k, Cityscapes, and VOC. We report the Root Mean Squared Error (RMSE) metric for the depth benchmarks NYUv2 and KITTI. For segmentation, we use an image resolution of 560×560.

		Segmentation			Depth	
Method	**Arch.**	**ADE20k**	**Citysc.**	**VOC**	**NYUv2 ↓**	**KITTI ↓**
Self-supervised Backbone						
MAE [21]	H/14	33.3	58.4	67.6	0.517	3.660
Web-DINO [16]	7B/14	42.7	68.3	76.1	0.466	3.158
DINOv3 [54]	7B/16	55.9	81.1	86.6	0.309	2.346
Weakly-supervised Backbone						
OpenCLIP [11]	G/14	39.3	60.3	71.4	0.541	3.570
SigLIP 2 [59]	G/16	42.7	64.8	72.7	0.494	3.273
PEcore [5]	G/14	38.9	61.1	69.2	0.590	4.119
Ours						
Qwen2-VL-ViT [66]	H/14	33.6	57.6	67.5	0.541	3.735
VersaViT	H/14	49.6	74.5	86.6	0.473	3.136
Absolute Gain △		(+16.0)	(+16.9)	(+19.1)	(-0.068)	(-0.599)

the training data utilized for VQA evaluation, please refer to the supplementary materials. (ii) Dense Feature Probing: To demonstrate VersaViT's capability for dense feature representation, we follow the methodology of [54] by employing a linear probing approach with the vision backbone kept frozen. (iii) In-Task and Out-of-Task Performance: To assess our model's performance on the tasks included in the multi-task collaborative training stage, we conduct evaluations using established benchmarks specific to those tasks and compare the results against specialized models. Furthermore, for out-of-task generalization, we evaluate VersaViT's retrieval performance on corresponding benchmarks.

Implementation Details. Our framework is built on PyTorch and is initialized by Qwen2-VL-7B-ViT [66]. Our model is capable of accepting dynamic resolutions. In the captioning alignment stage, we conduct experiments with a batch size of 1024 and a learning rate of 1e-3, training for one epoch. During the multi-task collaborative training phase, the learning rate for the vision encoder, the vision projection, and the text decoder is set to 1e-5, while the learning rate for the remaining parameters is set to 1e-4, again for one epoch. For more implementation details, please refer to the supplementary materials.

5.2 Experimental Results

In this section, we compare VersaViT with the Qwen2-VL-ViT baseline on both VQA and classic vision tasks to demonstrate the effectiveness of our method.

Improved VQA Performance. As outlined in Section 5.1, we adopt a two-stage training paradigm based on MLLMs to evaluate our vision backbone on various VQA tasks. Table 2 presents the performance of our model on the OpenCompass benchmark. The results indicate that (i) Crucially, under a fair comparison setting, our model demonstrates a significant improvement over the Qwen2-

Table 4: Ablation study on multi-task training. Here, we report the average score on the OpenCompass, the RMSE on the NYUv2, and the MIoU on the ADE20k.

VQA&Cap.	Depth	Referring Seg	OpenCompass↑	NYUv2↓	ADE20k↑
✘	✘	✘	64.8	0.541	33.6
✔	✘	✘	66.1	0.557	32.0
✔	✔	✘	66.3	0.495	35.1
✔	✔	✔	66.4	0.473	49.6

Table 5: Ablation study on loss weights.

λ_{cap}	λ_{depth}	λ_{seg}	OpenCompass↑	NYUv2↓	ADE20k↑
Baseline (Qwen2-VL-ViT)			64.8	0.541	33.6
0.50	0.25	0.25	66.5	0.470	48.3
0.25	0.50	0.25	65.7	0.455	48.9
0.25	0.25	0.50	65.7	0.476	50.0
0.33	0.33	0.33	66.4	0.473	49.6

Table 6: Ablation study on different vision backbones.

Model	OpenCompass↑	NYUv2↓	ADE20k↑
Qwen2-VL-ViT	64.8	0.541	33.6
VersaViT	66.4 (+1.6)	0.473 (-0.068)	49.6 (+16.0)
Qwen2.5-VL-ViT	64.0	0.533	37.2
VersaViT-Qwen-2.5	65.2 (+1.2)	0.464 (-0.069)	47.2 (+10.0)
Siglip2-so400M-512	59.3	0.497	40.9
VersaViT-Siglip2	60.4 (+1.1)	0.481 (-0.016)	49.2 (+8.3)

VL-ViT [66] baseline, achieving an average gain of 1.6 points across eight benchmarks. This consistent gain strongly suggests that our backbone's high-level semantic representations are superior to the baseline. (ii) When VersaViT is coupled with an LLM through instruction tuning with small-scale data, it generally outperforms several well-established open-source models, such as MiniCPM-V-2.6 [80] and DeepSeek-VL2 [70]. However, a performance gap remains when compared to top-tier MLLMs, such as Qwen2.5-VL-7B [1], primarily due to differences in data scale. For further details regarding the data and training for these two stages, please refer to the supplementary materials.

Improved Dense Features. To evaluate the dense features of our method, we utilize the frozen patch features from the final layer and assess their performance through linear probing on both semantic segmentation and monocular depth estimation tasks. The experimental results, as shown in Table 3, reveal that (i) VersaViT demonstrates superior performance to SigLIP 2 on both tasks, although there remains a slight gap compared to DINOv3. (ii) In particular, our method delivers substantial improvements over the Qwen2-VL-ViT baseline. For semantic segmentation, the performance increases from 33.6 to 49.6 on ADE20k and from 67.5 to 86.6 on Pascal VOC. For monocular depth estimation, the RMSE on the KITTI dataset is dramatically reduced from 3.735 to 3.136, and on the NYUv2 dataset, it drops from 0.541 to 0.473. These results demonstrate that our method significantly enhances the representational quality of dense features within the vision backbone.

Table 7: Effectiveness of multi-tasking vs. data scaling.

Model	OpenCompass↑	NYUv2↓	ADE20k↑
Baseline (35M Data, Single-task)	66.2	0.547	34.7
VersaViT	66.4	0.473	49.6

Table 8: Evaluation of 3D consistency of dense representations.

Method	Geometric	Semantic
	NAVI (Recall)	SPair (Recall)
Qwen2-VL-ViT [66]	39.27	17.05
VersaViT	41.64 (+2.37)	26.99 (+9.94)

5.3 Ablation Study & Analysis

In this section, we further investigate the effect of our method, for example, the effectiveness of multi-task training, the generalization of our method, *etc.*

Effectiveness of Multi-task Training. As shown in Table 4, it can be observed that as the number of tasks during the multi-task collaborative training phase increases, the model's overall performance improves. The following key observations can be made: (i) Training only the VQA and Image Captioning tasks effectively enhances the model's performance on the VQA benchmark, but it degrades the vision backbone's dense feature representation. This finding aligns with the observation made in Section 3. (ii) Jointly training the VQA and depth tasks significantly improves the model's performance on both VQA and depth-related tasks. (iii) Adding the segmentation task further enhances the model's capabilities, not only improving segmentation performance but also boosting performance in both VQA and depth tasks. Collectively, integrating these three tasks results in a more comprehensive visual representation.

Ablation Study on Loss Weights. As shown in Table 5, we conduct ablation experiments under different loss-weight settings. The experimental results indicate that increasing the loss weight of a specific task generally leads to slightly improved performance on that task. Moreover, balanced weights achieve the optimal trade-off among all tasks. Importantly, our method is robust to weight variations, as all configurations significantly outperform the baseline.

Generalization on Different Vision Backbones. In this section, we investigate the application of our method across different vision backbones. As shown in Table 6, our approach is compatible with various ViT architectures and consistently enhances the performance of the backbones across multiple aspects, thereby highlighting its flexibility and practical applicability.

Data-driven vs. Method-driven Gains. To isolate the effect of multi-tasking from data scaling, we introduce a baseline model trained on a 35M-scale dataset using only captioning and VQA tasks, excluding the multi-tasking framework. As demonstrated in Table 7, this comparison confirms that the performance gains stem from our multi-tasking methodology rather than mere data scaling.

Enhanced 3D Correspondence. We follow the evaluation setting established by Probe3D [15] to assess geometric correspondence on the NAVI [23] and se-

Table 9: Comparison of zero-shot relative depth estimation. Ours is trained via multi-tasking without task-specific fine-tuning. $*$ denotes depth fine-tuning.

Method	**KITTI** [17]		**NYUv2** [53]		**DA-2K** [78]
	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	Acc(%)
Specialist Models					
DPT [48]	10.0	90.1	9.8	90.3	-
MiDaS V3.1 [3]	12.7	85.0	4.8	98.0	-
DepthFM [19]	8.9	91.3	5.5	96.3	85.8
Marigold [27]	9.9	91.6	5.5	96.4	86.8
Depth Anything V2 [78]	7.5	94.8	4.4	97.9	97.4
Ours	9.4	90.5	7.7	94.4	91.3
Ours*	7.1	94.5	5.3	96.9	94.2

Table 10: Comparison of image referring segmentation performance. Results are shown for the fine-tuning setting (post-training on RefCOCO datasets).

Method	**refCOCO** [26]			**refCOCO+** [26]			**refCOCOg** [42]	
	val	testA	testB	val	testA	testB	val(U)	test(U)
Specialist Models								
LAVT [79]	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1
ReLA [34]	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0
LISA-7B [31]	74.1	76.5	71.1	62.4	67.4	56.5	66.4	68.5
VISA-7B [72]	72.4	75.5	68.1	59.8	64.8	53.1	65.5	66.4
VideoLISA-3.8B [2]	73.8	76.6	68.8	63.4	68.8	56.2	68.3	68.8
Ours	78.8	81.2	76.2	69.4	75.0	63.9	72.0	74.3

mantic correspondence on the SPair [43]. As demonstrated in Table 8, VersaViT significantly outperforms the baseline, achieving a 2.3-point increase on NAVI and a significant 10-point increase on SPair. This compelling result strongly indicates that our method effectively enhances the 3D-aware representations of the vision backbone. Further details are provided in the supplementary materials.

Transfer to Tasks Learned During Multi-task Training. Through the multi-task training stage, our model naturally acquires the ability to perform monocular depth estimation and image referring segmentation. For monocular depth estimation, we evaluate our model on three benchmarks, with the results detailed in Table 9. Despite being trained solely via multi-task learning, our method achieves performance on par with that of specialized depth estimation models. Moreover, after fine-tuning specifically for the depth estimation task, our model attains performance comparable to state-of-the-art methods, demonstrating that VersaViT is highly adaptable to the depth estimation task.

Similarly, for image referring segmentation, the experimental results are shown in Table 10. After fine-tuning on the RefCOCO, our model substantially outperforms a wide range of specialized methods, demonstrating strong transferability.

Transfer to Retrieval Tasks. To assess the performance of our backbone on retrieval tasks, we select `siglip2-so400m-patch14-384` as the text encoder. We consider two settings: a frozen configuration, where only the vision attention pooling layer and the text projection are trained; and a more extensive setting where the vision encoder training is enabled to fully explore the potential of our model. Following [86], we train on CC3M, CC12M, and YFCC15M.

Table 11: Comparison of zero-shot image-text retrieval. Results are evaluated under two settings: (i) fine-tuning only the projection layers, and (ii) unfreezing the vision backbone to fully exploit the retrieval potential. $*$ denotes the second setting.

Method	COCO		Flickr	
	Text → Image	Image → Text	Text → Image	Image → Text
CLIP-L [47]	36.5	56.3	65.2	85.2
EVA-CLIP-8B [56]	53.0	70.3	80.8	95.6
SigLIP [84]	52.0	70.2	80.5	93.5
SigLIP 2 [59]	55.8	71.7	85.7	94.9
DINOv3 [54]	45.6	63.7	-	-
Aligned on 30M image-text pairs				
Ours	46.9	60.5	77.0	88.3
Ours*	54.8	69.5	82.2	92.5

The experimental results, presented in Table 11, show that our model exhibits exceptional retrieval performance across various settings. Notably, with only 30M training samples and training solely the lightweight projection layer, our model surpasses CLIP and achieves comparable performance to DINOv3. Furthermore, when the vision encoder is enabled, our model performs slightly behind SigLIP 2 by about 2 points. This outcome clearly indicates that VersaViT still demonstrates strong potential and competitiveness in retrieval tasks.

Discussion on Seamless Integration. While our multi-task collaborative training is validated via post-training on the vision encoder, its design enables seamless integration into any stage of MLLM training (*e.g.*, pre-training, decay, and post-training). Typically, MLLM training relies exclusively on autoregressive loss; in contrast, our method augments this with two pixel-level objectives. This integration is facilitated by batch-wise task alternation, wherein only one task-specific loss is computed per batch. Such a design ensures that our method can be seamlessly adapted to existing MLLM training frameworks. Moreover, we anticipate that applying our approach during the earlier phases of MLLM training will allow the vision backbone to acquire more precise and semantically rich feature representations, which we consider an avenue for future work.

5.4 Qualitative Results

In Figure 3, we provide a qualitative comparison of VersaViT with the Qwen2-VL-ViT baseline across three tasks: VQA, semantic segmentation, and depth estimation. As demonstrated, after undergoing multi-task collaborative training, VersaViT significantly outperforms the baseline in each of these tasks. For instance, the example in the first row illustrates that our model exhibits a superior comprehension of spatial relationships. Furthermore, the examples in the second and

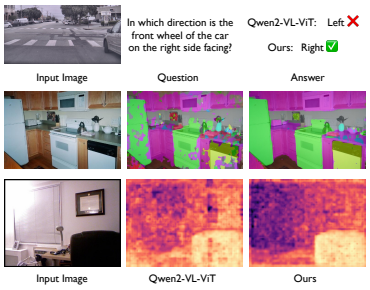

Fig. 3: Qualitative examples.

third rows highlight our model's enhanced ability to capture fine-grained features, resulting in more precise object segmentation and depth estimation.

6 Conclusion

In this paper, we have identified that the vision encoders in current MLLMs exhibit deficiencies in their dense feature representations, as manifested in their suboptimal performance on dense prediction tasks. To address this issue, we proposed VersaViT, a well-rounded vision backbone that employs a cost-effective multi-task collaborative post-training paradigm. Specifically, our approach simultaneously leverages three distinct tasks: VQA and image captioning (semantic), monocular depth estimation (geometric), and image referring segmentation (localized). The collaborative training aims to adapt MLLM vision encoders to achieve superior performance on both VQA and vision-centric tasks, ultimately enabling a versatile VFM. Furthermore, extensive experiments on a wide range of downstream tasks have demonstrated the significant improvement achieved by our method. We anticipate that our framework will provide valuable insight into visual foundation models and stimulate further research in this area.

Acknowledgements

This work is supported by the Shanghai Special Fund for Promoting High-Quality Industrial Development: Pilot Industry Innovation and Development Program (Artificial Intelligence Track, ID: 2025-GZL-RGZN-02020).

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025) 1, 2, 3, 4, 9, 11
2. Bai, Z., He, T., Mei, H., Wang, P., Gao, Z., Chen, J., Zhang, Z., Shou, M.Z.: One token to seg them all: Language instructed reasoning segmentation in videos. In: NIPS (2024) 13
3. Birkl, R., Wofk, D., Müller, M.: Midas v3. 1–a model zoo for robust monocular relative depth estimation. arXiv preprint arXiv:2307.14460 (2023) 13
4. Biten, A.F., Tito, R., Gomez, L., Valveny, E., Karatzas, D.: Ocr-idl: Ocr annotations for industry document library dataset. In: ECCV (2022) 7, 8
5. Bolya, D., Huang, P.Y., Sun, P., Cho, J.H., Madotto, A., Wei, C., Ma, T., Zhi, J., Rajasegaran, J., Rasheed, H., et al.: Perception encoder: The best visual embeddings are not at the output of the network. arXiv preprint arXiv:2504.13181 (2025) 3, 10
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. In: NIPS (2020) 4
7. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021) 3

8. Caruana, R.: Multitask learning. Machine learning (1997) 4
9. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: CVPR (2021) 8
10. Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., et al.: Are we on the right way for evaluating large vision-language models? In: NIPS (2024) 9
11. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: CVPR (2023) 10
12. Crawshaw, M.: Multi-task learning with deep neural networks: A survey. arXiv preprint arXiv:2009.09796 (2020) 4
13. Deitke, M., Clark, C., Lee, S., Tripathi, R., Yang, Y., Park, J.S., Salehi, M., Muennighoff, N., Lo, K., Soldaini, L., et al.: Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In: CVPR (2025) 1, 7, 8
14. Di, S., Zhai, Z., Xie, W.: Revisiting multi-task visual representation learning. arXiv preprint arXiv:2601.13886 (2026) 4
15. El Banani, M., Raj, A., Maninis, K.K., Kar, A., Li, Y., Rubinstein, M., Sun, D., Guibas, L., Johnson, J., Jampani, V.: Probing the 3d awareness of visual foundation models. In: CVPR (2024) 12
16. Fan, D., Tong, S., Zhu, J., Sinha, K., Liu, Z., Chen, X., Rabbat, M., Ballas, N., LeCun, Y., Bar, A., et al.: Scaling language-free visual representation learning. arXiv preprint arXiv:2504.01017 (2025) 10
17. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The international journal of robotics research (2013) 13
18. Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., et al.: Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In: CVPR (2024) 9
19. Gui, M., Schusterbauer, J., Prestel, U., Ma, P., Kotovenko, D., Grebenkova, O., Baumann, S.A., Hu, V.T., Ommer, B.: Depthfm: Fast generative monocular depth estimation with flow matching. In: AAAI (2025) 13
20. Guo, D., Wu, F., Zhu, F., Leng, F., Shi, G., Chen, H., Fan, H., Wang, J., Jiang, J., Wang, J., et al.: Seed1. 5-vl technical report. arXiv preprint arXiv:2505.07062 (2025) 1, 2, 3, 4
21. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR (2022) 3, 10
22. Heinrich, G., Ranzinger, M., Yin, H., Lu, Y., Kautz, J., Tao, A., Catanzaro, B., Molchanov, P.: Radiov2. 5: Improved baselines for agglomerative vision foundation models. In: CVPR (2025) 4
23. Jampani, V., Maninis, K.K., Engelhardt, A., Karpur, A., Truong, K., Sargent, K., Popov, S., Araujo, A., Martin Brualla, R., Patel, K., et al.: Navi: Category-agnostic image collections with high-quality 3d shape and pose annotations. In: NIPS (2023) 12
24. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML (2021) 3
25. Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., Casas, D.d.l., Hanna, E.B., Bressand, F., et al.: Mixtral of experts. arXiv preprint arXiv:2401.04088 (2024) 4

26. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: EMNLP (2014) 13
27. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: CVPR (2024) 13
28. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images. In: ECCV (2016) 9
29. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: ICCV (2023) 7, 8
30. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. IJCV (2020) 8
31. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: CVPR (2024) 4, 13
32. Li, Z., Di, S., Zhai, Z., Huang, W., Wang, Y., Xie, W.: Universal video temporal grounding with generative multi-modal large language models. In: NIPS (2025) 4
33. Lian, L., Ding, Y., Ge, Y., Liu, S., Mao, H., Li, B., Pavone, M., Liu, M.Y., Darrell, T., Yala, A., Cui, Y.: Describe anything: Detailed localized image and video captioning. In: ICCV (2025) 8
34. Liu, C., Ding, H., Jiang, X.: Gres: Generalized referring expression segmentation. In: CVPR (2023) 13
35. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NIPS (2023) 6
36. Liu, Y., Li, X., Zhang, L., Wang, Z., Zheng, Z., Zhou, Y., Xie, C.: Openvision 2: A family of generative pretrained visual encoders for multimodal learning. arXiv preprint arXiv:2509.01644 (2025) 3
37. Liu, Y., Zhang, Y., Cai, J., Jiang, X., Hu, Y., Yao, J., Wang, Y., Xie, W.: Lamra: Large multimodal model as your advanced retrieval assistant. In: CVPR (2025) 4
38. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: ECCV (2024) 9
39. Liu, Y., Tian, L., Zhou, X., Gao, X., Yu, K., Yu, Y., Zhou, J.: Points1. 5: Building a vision-language model towards real world applications. arXiv preprint arXiv:2412.08443 (2024) 2
40. Liu, Y., Li, Z., Li, H., Yu, W., Huang, M., Peng, D., Liu, M., Chen, M., Li, C., Jin, L., et al.: On the hidden mystery of ocr in large multimodal models. arXiv preprint arXiv:2305.07895 (2023) 9
41. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255 (2023) 9
42. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: CVPR (2016) 13
43. Min, J., Lee, J., Ponce, J., Cho, M.: Spair-71k: A large-scale benchmark for semantic correspondence. arXiv preprint arXiv:1908.10543 (2019) 13
44. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018) 1, 3
45. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) 3

46. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824 (2023) 8
47. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021) 1, 3, 14
48. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV (2021) 7, 13
49. Ranzinger, M., Heinrich, G., Kautz, J., Molchanov, P.: Am-radio: Agglomerative vision foundation model reduce all domains into one. In: CVPR (2024) 4
50. Sarıyıldız, M.B., Weinzaepfel, P., Lucas, T., De Jorge, P., Larlus, D., Kalantidis, Y.: Dune: Distilling a universal encoder from heterogeneous 2d and 3d teachers. In: CVPR (2025) 3
51. Sarıyıldız, M.B., Weinzaepfel, P., Lucas, T., Larlus, D., Kalantidis, Y.: Unic: Universal classification models via multi-teacher distillation. In: ECCV (2024) 3
52. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: ICCV (2019) 8
53. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012) 13
54. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025) 3, 10, 14
55. Song, J., Yun, S., Han, D., Choo, J., Heo, B.: Rl makes mllms see better than sft. arXiv preprint arXiv:2510.16333 (2025) 3
56. Sun, Q., Wang, J., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, X.: Eva-clip-18b: Scaling clip to 18 billion parameters. arXiv preprint arXiv:2402.04252 (2024) 14
57. Team, K., Du, A., Yin, B., Xing, B., Qu, B., Wang, B., Chen, C., Zhang, C., Du, C., Wei, C., et al.: Kimi-vl technical report. arXiv preprint arXiv:2504.07491 (2025) 1, 3, 4
58. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023) 4
59. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025) 1, 10, 14
60. Tschannen, M., Kumar, M., Steiner, A., Zhai, X., Houlsby, N., Beyer, L.: Image captioners are scalable vision learners too. In: NIPS (2023) 3
61. Vandenhende, S., Georgoulis, S., Van Gansbeke, W., Proesmans, M., Dai, D., Van Gool, L.: Multi-task learning for dense prediction tasks: A survey. TPAMI (2021) 4
62. Wan, B., Tschannen, M., Xian, Y., Pavetic, F., Alabdulmohsin, I.M., Wang, X., Susano Pinto, A., Steiner, A., Beyer, L., Zhai, X.: Locca: Visual pretraining with location-aware captioners. In: NIPS (2024) 3
63. Wang, H., Ju, C., Lin, W., Xiao, S., Chen, M., Huang, Y., Liu, C., Yao, M., Lan, J., Chen, Y., et al.: Advancing myopia to holism: Fully contrastive language-image pre-training. In: CVPR (2025) 4
64. Wang, H., Chen, Q., Yan, C., Cai, J., Jiang, X., Hu, Y., Xie, W., Gavves, S.: Object-centric video question answering with visual grounding and referring. In: ICCV (2025) 4

65. Wang, J., Hu, X., Gan, Z., Yang, Z., Dai, X., Liu, Z., Lu, Y., Wang, L.: Ufo: A unified transformer for vision-language representation learning. arXiv preprint arXiv:2111.10023 (2021) 4
66. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024) 2, 3, 4, 9, 10, 11, 12
67. Weyand, T., Araujo, A., Cao, B., Sim, J.: Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In: CVPR (2020) 8
68. Wiedmann, L., Zohar, O., Mahla, A., Wang, X., Li, R., Frere, T., von Werra, L., Gosthipaty, A.R., Marafioti, A.: Finevision - open data is all you need (September 2025), https://huggingface.co/datasets/HuggingFaceM4/FineVision/ 8
69. Wu, J., Zhong, M., Xing, S., Lai, Z., Liu, Z., Chen, Z., Wang, W., Zhu, X., Lu, L., Lu, T., et al.: Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. In: NIPS (2024) 4
70. Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al.: Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. arXiv preprint arXiv:2412.10302 (2024) 9, 11
71. Xiaomi, L.C.T.: Mimo-vl technical report (2025), https://arxiv.org/abs/2506.03569 2
72. Yan, C., Wang, H., Yan, S., Jiang, X., Hu, Y., Kang, G., Xie, W., Gavves, E.: Visa: Reasoning video object segmentation via large language models. In: ECCV (2024) 13
73. Yan, Y., Xu, J., Di, S., Liu, Y., Shi, Y., Chen, Q., Li, Z., Huang, Y., Xie, W.: Learning streaming video representation via multitask training. In: ICCV (2025) 4
74. Yan, Y., Xu, J., Di, S., Wu, H., Xie, W.: Omnistream: Mastering perception, reconstruction and action in continuous streams. arXiv preprint arXiv:2603.12265 (2026) 4
75. Yan, Z., Li, Z., He, Y., Wang, C., Li, K., Li, X., Zeng, X., Wang, Z., Wang, Y., Qiao, Y., et al.: Task preference optimization: Improving multimodal large language models with vision task alignment. In: CVPR (2025) 4
76. Yang, B., Wen, B., Ding, B., Liu, C., Chu, C., Song, C., Rao, C., Yi, C., Li, D., Zang, D., et al.: Kwai keye-vl 1.5 technical report. arXiv preprint arXiv:2509.01563 (2025) 3
77. Yang, H., Rao, J., Wu, H., Xie, W.: Soccermaster: A vision foundation model for soccer understanding. arXiv preprint arXiv:2512.11016 (2025) 4
78. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. In: NIPS (2024) 7, 13
79. Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: Lavt: Language-aware vision transformer for referring image segmentation. In: CVPR (2022) 13
80. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., et al.: Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800 (2024) 9, 11
81. Yin, W., Yang, D., Dong, H., Kang, Z., Wang, J., Liang, X., Feng, C., Ran, J.: Sailvit: Towards robust and generalizable visual backbones for mllms via gradual feature refinement. arXiv preprint arXiv:2507.01643 (2025) 3
82. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:2308.02490 (2023) 9

83. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: CVPR (2024) 9
84. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: ICCV (2023) 3, 14
85. Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., et al.: Qwen3 embedding: Advancing text embedding and reranking through foundation models. arXiv preprint arXiv:2506.05176 (2025) 7
86. Zheng, K., Zhang, Y., Wu, W., Lu, F., Ma, S., Jin, X., Chen, W., Shen, Y.: Dreamlip: Language-image pre-training with long captions. In: ECCV (2024) 13
87. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. TPAMI (2017) 8
88. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832 (2021) 3
89. Zhuang, W., Chen, C., Li, Z., Sajadmanesh, S., Li, J., Huang, J., Sehwag, V., Sharma, V., Shinozaki, H., Garcia, F.C., et al.: Argus: A compact and versatile foundation model for vision. In: CVPR (2025) 4