

Florence-VL: Enhancing Vision-Language Models with Generative Vision Encoder and Depth-Breadth Fusion

Jiuhai Chen^{1*}, Jianwei Yang², Haiping Wu², Dianqi Li, Jianfeng Gao², Tianyi Zhou¹, Bin Xiao²

¹University of Maryland

²Microsoft Research

Abstract

We present *Florence-VL*, a new family of multimodal large language models (MLLMs) with enriched visual representations produced by *Florence-2* [45], a generative vision foundation model. Unlike the widely used CLIP-style vision transformer [35] trained by contrastive learning, *Florence-2* can capture different levels and aspects of visual features, which are more versatile to be adapted to diverse downstream tasks. We propose a novel feature-fusion architecture and an innovative training recipe that effectively integrates *Florence-2*'s visual features into pre-trained LLMs, such as *Phi 3.5* and *LLama 3*. In particular, we propose “depth-breath fusion (DBFusion)” to fuse the visual features extracted from different depths and under multiple prompts. Our model training is composed of end-to-end pretraining of the whole model followed by finetuning of the projection layer and the LLM, on a carefully designed recipe of diverse open-source datasets that include high-quality image captions and instruction-tuning pairs. Our quantitative analysis and visualization of *Florence-VL*'s visual features show its advantages over popular vision encoders on vision-language alignment, where the enriched depth and breath play important roles. *Florence-VL* achieves significant improvements over existing state-of-the-art MLLMs across various multi-modal and vision-centric benchmarks covering general VQA, perception, hallucination, OCR, Chart, knowledge-intensive understanding, etc. To facilitate future research, our models and the complete training recipe are open-sourced. <https://github.com/JiuhaiChen/Florence-VL>

1. Introduction

Recent progress in multimodal large language models (MLLMs) are largely driven by progress in large language

*The work is done during Jiuhai Chen's internship at Microsoft Research.

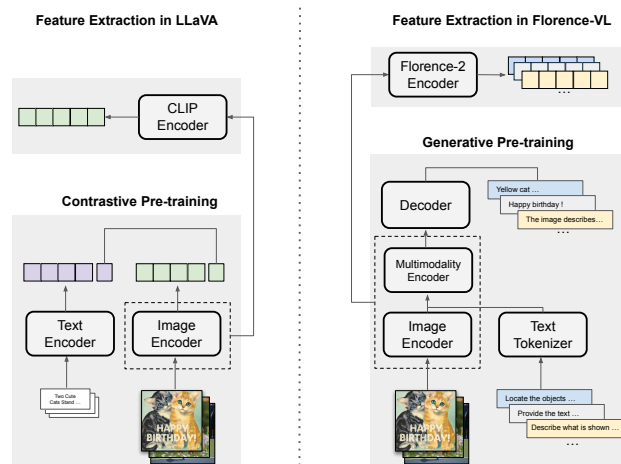


Figure 1. Comparison of LLaVA-style MLLMs with our Florence-VL. LLaVA-style models use CLIP, pretrained with contrastive learning, to generate a **single high-level image feature**. In contrast, Florence-VL leverages Florence-2, pretrained with generative modeling across various vision tasks such as image captioning, OCR, and grounding. This enables Florence-VL to flexibly extract **multiple task-specific image features** using Florence-2 as the image encoder.

models [26, 49]. However, when it comes to visual encoders, transformer-based models like CLIP or SigLIP remain the most commonly used choices. Despite CLIP and SigLIP's effectiveness, they come with limitations; for instance, their last-layer features usually provide an image-level semantic representation that captures the overall scene and context, but often overlook pixel or region-level details and low-level features that are critical to various downstream tasks. There is a much broader range of visual representation, such as the self-supervised DINOv2 model [34], diffusion model [37] and segmentation [20], [41] shows these different visual encoders can benefit well in some specific tasks.

In order to leverage distinctive representations of multiple vision encoders, some recent works such as [38, 41]

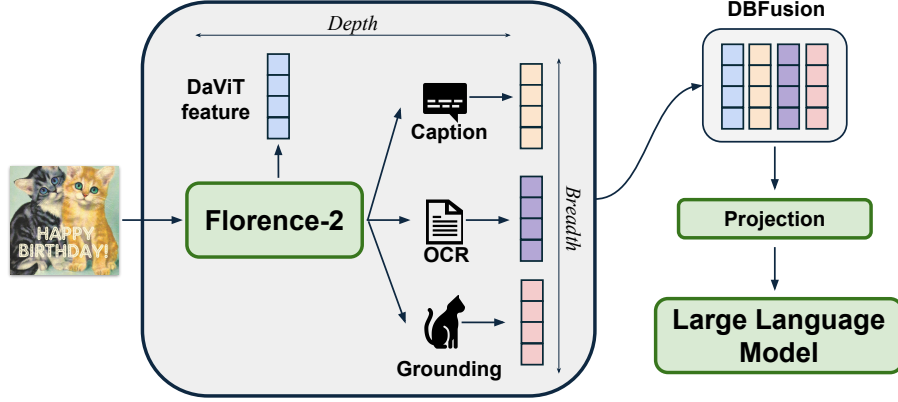


Figure 2. An overview of Florence-VL, which extracts visual features of different depths (levels of feature concepts) and breaths (prompts) from Florence-2, combines them using DBFusion, and project the fused features to an LLM’s input space. Florence-VL is fully pretrained on image captioning data and then partially finetuned on instruction-tuning data.

adopt a mixture of vision encoders that specialize in different feature aspects or skills. However, integrating multiple vision encoders increases the computational expense for both model training and deployment. *Could a single vision model be designed to generate distinct visual features, each emphasizing different perceptual information in the input image?* In this paper, we propose Florence-VL, which leverages the generative vision foundation model Florence-2 [45] as the vision encoder. Florence-2 offers a prompt-based representation for various computer vision tasks, including captioning, object detection, grounding, and OCR. Its versatile visual representations can benefit different types of downstream tasks. For instance, OCR-based representations are advantageous for tasks that require extracting textual information from images, and grounding-based representation can benefit for tasks that require the relationships between objects and their spatial contexts. However, to build a better MLLM, how to extract these diverse features and align them with a pretrained LLM remains unexplored.

To address this, we propose *Depth-Breadth Fusion (DBFusion)* to effectively selecting and utilizing diverse visual features. Visual features from different layers capture various levels of concepts, with the final layers typically representing higher-level concepts. Integrating lower-level features can therefore complement these high-level representations, which we refer to as the “Depth” of visual features. Additionally, since different downstream tasks need different perceptual information within images, a single image feature often falls short in capturing all relevant information. Thus, we leverage multiple image features, with each feature capturing different visual representations. We refer to this as the “Breadth” of visual features. For utilizing these diverse visual features, we find that a straightforward channel concatenation serves as a simple yet effective fu-

sion strategy. Specifically, we concatenate multiple features along the channel dimension, and these combined features, spanning various depths and breadths, are then projected as input embedding to LLMs.

We train Florence-VL on a novel recipe of open-sourced training data, which is composed of a large-scale detailed captioning dataset and a mix of instruction tuning datasets for whole-model pretraining and partial-model finetuning, respectively. The resulted Florence-VL achieves significant advantages on 25 benchmarks covering vision-centric, knowledge-based, and OCR & Chart tasks, outperforming other advanced MLLMs like Cambrian [41]. Moreover, we provide quantitative analysis and visualization demonstrating that Florence-VL’s visual representation achieves better alignment to LLMs than the widely adopted vision encoders such as CLIP and SigLIP [26].

2. Preliminary: Florence-2

Florence-2 [45] is a vision foundation model that utilizes a unified, prompt-based approach to handle various vision tasks with simple instructions, such as captioning, object detection, grounding, and segmentation. The architecture consists of a vision encoder DaViT [9] and a standard encoder-decoder model. It processes an input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ (where H and W indicate height and width, respectively) into flattened visual token embeddings. The model then applies a standard encoder-decoder transformer architecture to process both visual and language token embeddings. It first generates prompt text embeddings $\mathbf{T} \in \mathbb{R}^{N_t \times D}$ using the language tokenizer and word embedding layer, with N_t and D representing the number and dimensionality of prompt tokens, respectively. The vision token embeddings are then concatenated with the prompt embeddings to create the input for the multi-modality encoder module, $\mathbf{X} = [\mathbf{V}, \mathbf{T}]$, where $\mathbf{V} \in \mathbb{R}^{N_v \times D}$ is pro-

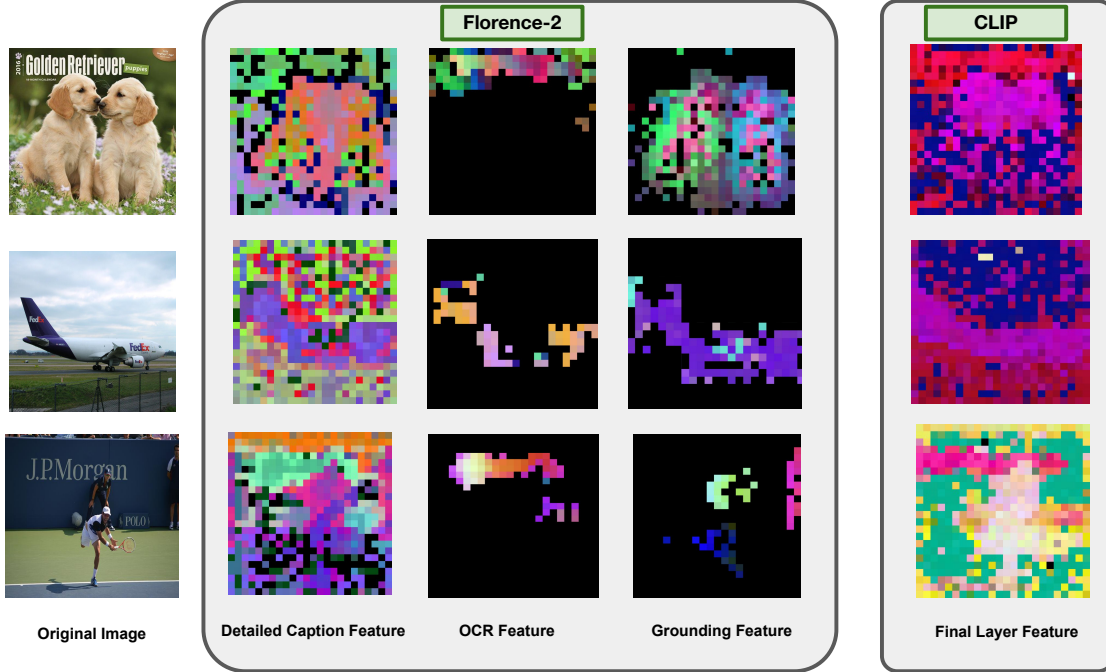


Figure 3. Visualization of the first three PCA components: we apply PCA to image features generated from Detailed Caption, OCR, and Grounding prompts, excluding the background by setting a threshold on the first PCA component. The image features derived from the Detailed Caption prompt (second column) capture the general context of the image, those from the OCR prompt (third column) focus primarily on text information, and those from the Grounding prompt (fourth column) highlight spatial relationships between objects. Additionally, we visualize the final layer features from OpenAI CLIP (ViT-L/14@336) in the last column, showing that CLIP features often miss certain region-level details, such as text information in many cases.

duced by applying a linear projection and LayerNorm layer to visual embedding from DaViT, with N_v and D representing the number and dimensionality of vision tokens, respectively. The linear projection and LayerNorm layer are used to ensure dimensionality alignment with $\mathbf{T}$. Encoder-decoder model will process the $\mathbf{X}$ and generate the desirable results, such as captions, object detections, grounding in textual form.

3. Method

3.1. Using Florence-2 as Vision Backbone

To address the limitations of existing vision backbones in MLLMs, specifically, last layer features typically yield an image-level representation that captures overall scene and context but often misses pixel- or region-level details, we utilize the vision foundation model Florence-2 as our visual encoder for extracting visual features. Unlike the CLIP pre-trained vision transformers that provide a single, universal image feature, Florence-2 can identify spatial details at different scales, by using different tasks prompts.

In MLLMs, effective image understanding requires capturing multiple levels of granularity, from global semantics to local details, and understanding spatial relationships between objects and entities within their semantic context.

Florence-2, with its capability to manage diverse granularity levels, is an ideal vision encoder to address these core aspects of image comprehension. In the following section, we explore how to leverage Florence-2’s strengths in integrating it into MLLMs.

3.2. Visual Features spanning Depth and Breadth

Breadth. Since different downstream tasks require varying perceptual information from images, we consider expanding the breadth of visual representation. Given an input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ and a task-specific prompt, such as “provide the text shown in the image”, Florence-2 will process the image feature and prompt feature into $\mathbf{X} = [\mathbf{V}, \mathbf{T}]$ and then feed into the encoder-decoder transformer architecture. The encoder employs an attention mechanism to process $\mathbf{X}$, producing an output $\mathbf{X}' = [\mathbf{V}', \mathbf{T}']$. Due to the cross-attention between $\mathbf{V}$ and $\mathbf{T}$, the updated image feature $\mathbf{V}'$ becomes more focused on the prompt “provide the text shown in the image”, specifically extracting more text information from the image.

We focus on three distinct tasks that contribute to image understanding, resulting in three different image embeddings $[\mathbf{V}'_{t_1}, \mathbf{V}'_{t_2}, \mathbf{V}'_{t_3}]$, each tailored to a specific task:

- **Detailed Image Caption:** describe what is shown in

	# Vis tok	MMBench (EN)	POPE	MM-Vet	MME-P	Seed-image	HallusionBench	LLaVA-bench	AI2D	MathVista	MMMU	OCRBench	ChartQA	DocVQA	InfoVQA	Average
Token Integration	1728	66.6	88.7	34.1	1536.3	70.9	45.0	63.3	56.9	28.1	36.4	40.8	23.0	44.6	29.5	50.3
Average Pooling	576	65.7	88.8	32.3	1551.3	70.3	45.7	64.6	56.6	27.4	36.0	41.2	24.6	44.8	29.3	50.4
Channel Integration	576	66.1	89.4	35.2	1543.5	70.3	46.8	65.0	57.2	28.0	35.6	41.4	24.3	44.5	29.4	50.8

Table 1. Experiments for different fusion strategies. The vision token count is 1728 for token integration, which leads to longer training and inference times. The channel integration strategy shows better performance and training efficiency compared to the other two fusion methods.

the image with a paragraph. It enables the model to give a overall context of an image.

- **OCR:** provide the text shown in the image. It extracts more text information from the image.
- **Dense Region Caption:** locate the objects in the image, with their descriptions. It captures the spatial relationships between objects.

We visualize the image features with different task prompts, applying PCA to the visual embeddings and setting a threshold for the visualization. As illustrated in Figure 3, different image embeddings emphasize distinct conceptual information within the images. Additionally, we also visualize the final layer image features from OpenAI CLIP in Figure 3, which often lacks certain region-level details in most cases.

Depth. We also integrate lower-level features using $\mathbf{V}$ from DaViT, combined with higher-level features $[\mathbf{V}'_{t_1}, \mathbf{V}'_{t_2}, \mathbf{V}'_{t_3}]$ derived from the three prompts, allows us to capture multiple levels of conceptual detail.

3.3. Depth-Breadth Fusion

Since we have image feature with different level of granularity, feature fusion is commonly used. When dealing with multiple feature embeddings, such as $[\mathbf{V}, \mathbf{V}'_{t_1}, \mathbf{V}'_{t_2}, \mathbf{V}'_{t_3}]$, the next question becomes how to fuse these features and align them with the language model space. To take advantage of all these four features, several approaches can be considered for this fusion process:

- **Token Integration:** This approach involves concatenating all features along the token dimension. However, this can make the visual token excessively long and complicate model training.
- **Average Pooling:** Alternatively, average pooling over all features can be used, but this method may result in information loss.

- **Channel Integration:** A more effective method is to concatenate features along the channel dimension, which does not increase the sequence length.

To quickly assess which feature fusion method provides the best overall performance, we use datasets from LLaVA-1.5 [26], which include 558K image captions for pre-training and 665K entries for instruction tuning. In the Table 1, the channel integration strategy shows better performance and training efficiency compared to the other two fusion methods. Thus we choose channel integration simple yet effective fusion strategy.

3.4. Florence-VL

As shown in Figure 2, Florence-VL is composed of the vision foundation model Florence-2 and the large language model. After extracting multiple image features, we use MLP to project these features into the language model space. During the pretraining stage, we align Florence-2 with the language model using image detailed caption data. In the instruction tuning stage, we use diverse and high-quality instruction-tuning dataset to effectively adapt the model to downstream tasks.

4. Analysis on Different Vision Encoders

To demonstrate that Florence-2 is a superior vision encoder compared to others, we quantify the cross-modal alignment quality between various vision encoders and language models, allowing us to assess the impact of different vision encoders without requiring subsequent supervised fine-tuning and evaluations on benchmarks [15, 43]. Specifically, consider a pretrained MLLM $\mathcal{M} = (\mathcal{V}, \mathcal{L})$ where $\mathcal{V}$ is the vision encoder and $\mathcal{L}$ represents the language model, we input a set of image-text pairs, $(V, T) = (\{v_n\}_{n=1}^N, \{t_n\}_{n=1}^N)$, into the model. For the n^{th} image-text pair, the vision encoder produces vision representations $f^{v_n} \in \mathbb{R}^{r_n \times d'}$, and the text representations $f^{t_n} \in \mathbb{R}^{s_n \times d}$ from last layer of the language decoder, where r_n and s_n are the number of tokens in the vision and text representations, and d' and d are the hidden state dimensions for the vision and text tokens.

We apply the trainable projection $\mathcal{P}$ to f^{v_n} to ensure dimensionality alignment with f^{t_n} , that is $\mathcal{P}(f^{v_n}) \in \mathbb{R}^{r_n \times d}$. We also apply average pooling along token dimension and normalize along the hidden dimension for both $\mathcal{P}(f^{v_n})$ and f^{t_n} . For all image-text pairs, we concatenate all vision features along the first dimension to form a matrix $F^{v_n} \in \mathbb{R}^{N \times d}$, and similarly concatenate all text features into a matrix $F^{t_n} \in \mathbb{R}^{N \times d}$. Since we need to measure the modality gap between vision tokens and text tokens, we compute the divergence between these two token representations. Specifically, we optimize the trainable projection $\mathcal{P}$, which is used to bring these two representations closer together by minimizing a cross-entropy loss function:

$$\mathcal{L} = - \sum_{i,j} \mathcal{I}_n^{(i,j)} \log (\text{softmax}(F^{v_n} \times (F^{t_n})^T)_{i,j})$$

, where $\mathcal{I}_n$ is the target (indicator) matrix. The multiplication of F^{v_n} with the transpose of F^{t_n} calculates the correlation between vision and text token representations. In short, the loss function is designed to minimize the distance between vision tokens and their corresponding text tokens by maximizing the likelihood that each vision token aligns correctly with its associated text token.

We use a set of image-text pairs $(V, T) = (\{v_n\}_{n=1}^N, \{t_n\}_{n=1}^N)$ from the LLaVA 1.5 pretraining image captioning datasets and select various vision encoders to assess how well we can optimize the alignment between the vision encoder and the language model. The vision encoders we evaluate include: Stable Diffusion [36], Dinov2 [34] (ViT-G/14, ViT-L/14, ViT-B/14), SigLIP, OpenAI CLIP, and our Florence-2 model. The chosen language model is Llama 3 8B Instruct. We plot the alignment loss in Figure 4, which clearly shows that Florence-2 vision encoder achieves the lowest alignment loss compared to the other vision encoders, demonstrating the best alignment with text embeddings. Additionally, SigLIP demonstrates competitive results, as noted in [41], which highlights SigLIP’s strong benchmark performance relative to other vision encoders, aligning with the findings of our study.

5. Experiments

Implementation Details. In order to build a state-of-the-art MLLM, we use images from CC12M [4], Redcaps [8], and Commonpool [12] during the pretraining stage, with detailed captions sourced from PixelProse [40]. For the instruction tuning stage, we also curate our high quality instruction tuning datasets, sourcing from Cambrian-7M [41], Vision Flan [46], ShareGPT4V [5], along with additional data from Docmatix [17] to improve chart and diagram comprehension [3]. The detail of training datasets and experiment details can be found in the appendix.

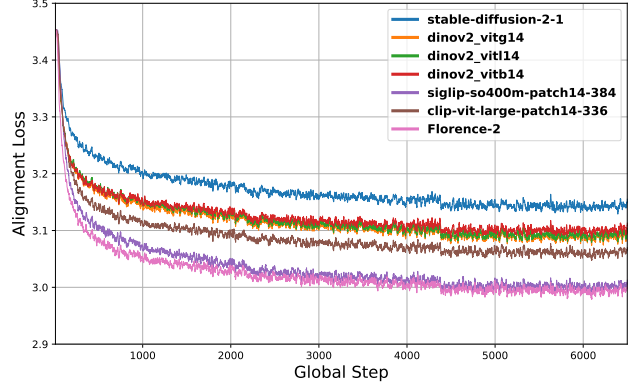


Figure 4. We plot the alignment loss for different vision encoders, which clearly shows that Florence-2 vision encoder achieves the lowest alignment loss compared to the other vision encoders, demonstrating the best alignment with text embeddings.

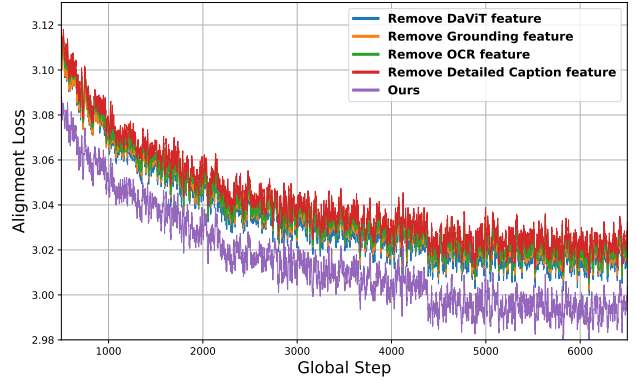


Figure 5. We plot the alignment loss for various feature combinations, removing one feature at a time from different depths and breadths. The results clearly show that our method achieves the lowest alignment loss compared to others, highlighting the importance of all features from different depths and breadths for optimal alignment.

Evaluation. We evaluate the performance of different MLLM models on 25 benchmarks with four different categories:

- General multimodal benchmarks: VQAv2 [13], GQA [16], MMBench (EN and CN) [27], VisWiz [14], POPE [22], MM-Vet [47], MME Perception [11], MME Cognition [11], SeedBench [21], Hallusion-Bench, LLaVA in the Wild [26] and MMStar [6].
- OCR & Chart benchmark: TextVQA [39], OCR-Bench [28], ChartQA [31], DocVQA [32] and InforVQA [33].
- Knowledge based benchmark: AI2D [19], Math-Vista [30], MMMU [48] and ScienceQA [29].

		General Benchmarks												
	# Vis tok.	VQAv2	GQA	MMBench (EN)	MMBench (CN)	VizWiz	POPE	MM-Vet	MME-P	MME-C	Seed-image	HallusionBench	LLaVA-bench	MMStar
Vila 3B	-	80.4	61.5	63.4	52.7	53.5	86.9	35.4	1442.4	-	67.9	-	-	40.3
Phi 3.5-Vision	-	-	63.5	75.5	64.2	58.2	82.2	46.5	1473.4	412.1	69.9	53.3	68.8	49.0
Florence-VL 3B (ours)	576	82.1	61.8	71.6	60.8	59.1	88.3	51.0	1498.7	403.9	70.6	58.1	71.1	44.9
LLaVA next 8B	2880	-	65.4	72.2	-	57.7	86.6	41.7	1595.1	379.3	72.7	47.7	76.8	-
Vila 8B	-	80.9	61.7	72.3	66.2	58.7	84.4	38.3	1577.0	-	71.4	-	-	-
Mini-Gemini-HD 8B	2880	-	64.5	72.7	-	-	-	-	1606.0	-	73.2	-	-	-
Cambrain 8B	576	-	64.6	75.9	67.9	-	87.4	48.0	1547.1	-	74.7	48.7	71.0	50.0
Florence-VL 8B (ours)	576	84.7	64.4	76.2	69.5	59.1	89.9	56.3	1560.0	381.1	74.9	57.3	74.2	50.0

(a) Results on general multimodal benchmarks.

# Vis tok.	Vision centric			Knowledge based				OCR & Chart				
	Realworldqa	CV-Bench*	MMVP	AI2D	MathVista	MMMU	SciQA-IMG	TextVQA	OCRBench	ChartQA	DocVQA	InfoVQA
Vila 3B	-	53.3	55.2	-	30.6	34.1	67.9	58.1	-	-	-	-
Phi 3.5 Vision	-	53.5	69.3	77.4	-	43.3	89.0	61.1	59.8	72.0	75.9	40.7
Florence-VL 3B (ours)	576	60.4	70.2	64.7	73.8	52.2	41.8	69.1	63.0	70.7	82.1	51.3
LLaVA next 8B	2880	59.6	63.8	38.7	71.6	37.4	40.1	65.4	55.2	69.3	78.2	-
Vila 8B	-	-	-	-	-	-	36.9	-	-	-	-	-
Mini-Gemini-HD 8B	2880	62.1	62.6	18.7	73.5	37.0	37.3	70.2	47.7	59.1	74.6	-
Cambrian 8B	576	64.2	72.2	51.3	73.0	49.0	42.7	71.7	62.4	73.3	77.8	-
Florence-VL 8B (ours)	576	64.2	73.4	73.3	74.2	55.5	43.7	74.2	63.4	74.7	84.9	51.7

(b) Results on Vision centric, Knowledge based, and OCR & Chart benchmarks.

Table 2. Results on general multimodal benchmarks, Vision centric, Knowledge based, and OCR & Chart benchmarks.

- Vision Centric benchmark: MMVP [42], Real-worldQA [44] and CV-Bench [41].

Baselines. We select two language backbones: Phi-3.5-mini-Instruct and LLaMA-3-8B-Instruct. For baseline comparisons among small models, we chose Vila 1.5 3B [24] and Phi 3.5-Vision-Instruct [1]. For the larger models, we select the baselines: LLaVA Next 8B [25], Vila 8B [24], Mini-Gemini-HD 8B [23] and Cambrian 8B [41], using LLaMA 3 8B Instruct as the language backbone.

Results. In the Table 2, we present the results of Florence-VL compared to various baselines across a range of benchmarks, along with the number of visual tokens used. For the smaller-sized model, our model outperforms Vila 3B, and surpasses Phi 3.5 Vision on 12 out of 24 tasks. Notably, Phi 3.5 Vision utilizes 500 billion vision and text tokens [1], with its training data being proprietary and significantly

larger than ours. Nonetheless, our Florence-VL 3B remains competitive with this model. For the larger-sized model, our model shows a significant improvement over other baselines on most benchmarks. Notably, our model significantly outperforms Cambrian-8B, which utilizes multiple vision encoders and combines their image features, whereas we achieve superior results using only a single vision encoder.

6. Discussion

Results using LLaVA 1.5 Data. Since we curate our training data when building our MLLMs, we disentangle the effects of training data and model architecture to clearly demonstrate our method’s effectiveness. Specifically, to highlight the advantages of our model architecture, we use the **exact same** pretraining and instruction dataset as LLaVA 1.5 [26]. We test different language backbones, including Phi-3.5-mini-Instruct, Vicuna 1.5 7B, and LLaMA-3-8B-Instruct. As shown in Tables 3, our model

LLM		GQA	MMBench (EN)	MMBench (CN)	VizWiz	POPE	MM-Vet	MME-P	MME-C	HallusionBench	LLaVA-bench	MMStar
LLaVA 1.5 3B	Phi 3.5	61.4	69.4	60.6	38.4	86.2	35.4	1399.5	284.6	44.5	68.0	40.6
Florence-VL 3B	Phi 3.5	62.7	68.7	61.7	42.6	89.9	35.4	1448.5	299.6	45.5	64.9	40.8
LLaVA 1.5 7B	Vicuna 1.5	62.0	64.8	57.6	50.0	85.9	30.6	1510.7	294.0	44.8	64.2	30.3
Florence-VL 7B	Vicuna 1.5	62.7	66.1	55.8	54.5	89.4	35.2	1543.5	316.4	46.8	65.0	36.8
LLaVA 1.5 8B	Llama 3	62.8	71.4	65.5	49.3	84.8	34.2	1539.4	292.5	45.7	71.0	38.5
Florence-VL 8B	Llama 3	63.8	71.1	65.8	54.0	88.4	36.4	1584.1	346.8	46.8	66.2	39.1

(a) Results on general multimodal benchmarks.

LLM		Realworldqa	MMVP	AI2D	MathVista	MMMU	SciQA-IMG	TextVQA	OCRBench	ChartQA	DocVQA	InfoVQA
LLaVA 1.5 3B	Phi 3.5	54.4	2.0	63.3	30.6	40.7	72.0	43.7	30.4	16.4	28.1	26.4
Florence-VL 3B	Phi 3.5	58.4	6.0	64.9	30.6	39.6	68.7	61.6	40.3	21.8	46.1	29.6
LLaVA 1.5 7B	Vicuna 1.5	54.8	6.0	54.8	26.7	35.3	66.8	58.2	31.4	18.2	28.1	25.8
Florence-VL 7B	Vicuna 1.5	60.4	12.3	57.2	28.0	35.6	66.5	62.8	41.4	24.3	44.5	29.4
LLaVA 1.5 8B	Llama 3	55.7	7.3	60.2	29.3	39.4	76.5	45.4	34.6	15.4	28.6	26.4
Florence-VL 8B	Llama 3	59.9	8.3	62.4	31.8	39.9	73.6	68.0	41.1	23.4	44.4	29.0

(b) Results on Vision centric, Knowledge based, and OCR & Chart benchmarks.

Table 3. We compare LLaVA 1.5 with our model (Florence-VL 3B/7B/8B) across multiple multimodal benchmarks. The key difference between them lies in the vision encoders used (CLIP for LLaVA vs. Florence-2 for our model), while we maintain the same training data and backbone LLMs for both. The results show that our models significantly outperform LLaVA 1.5 with the same training data.

design significantly outperforms the LLaVA architectures when trained on the same dataset. Notably, for OCR & Chart tasks, Florence-VL significantly outperforms LLaVA 1.5, demonstrating that OCR image features are essential for effective text-based image understanding.

Study on Depth Features Impacts. We aim to examine the impact of image features from different depths. For the feature set $[\mathbf{V}, \mathbf{V}'_{t_1}, \mathbf{V}'_{t_2}, \mathbf{V}'_{t_3}]$, we first remove all higher-level features $[\mathbf{V}'_{t_1}, \mathbf{V}'_{t_2}, \mathbf{V}'_{t_3}]$ and retain only the lower-level feature $[\mathbf{V}]$. We then evaluate the performance across different benchmarks, and as shown in Table 4, using only the lower-level feature $[\mathbf{V}]$ performs worse than our complete method. Next, we remove the lower-level feature $[\mathbf{V}]$ and keep only the higher-level features $[\mathbf{V}'_{t_1}, \mathbf{V}'_{t_2}, \mathbf{V}'_{t_3}]$. The alignment loss, displayed in Figure 5, clearly indicates that excluding the lower-level features (i.e., removing DaViT features) results in a higher alignment loss compared to our method. Therefore, both ablation studies confirm that features from different depths are essential for achieving optimal performance.

Study on Breadth Features Impacts. In Table 5, we analyze the impact of each feature from different breadths by individually removing one feature at a time from $[\mathbf{V}'_{t_1}, \mathbf{V}'_{t_2}, \mathbf{V}'_{t_3}]$. For instance, to assess the effect of the caption feature, we retain only the OCR and grounding features. The results in Table 5 show that combining all three features yields the best average benchmark performance. Additionally, we plot the alignment loss when each feature is removed individually, as shown in Figure 5. This further demonstrates incorporating all three features from different breadths is essential for effectively extracting visual information.

7. Related Work

LLMs have significantly advanced the development of MLLMs, including models like LLaVA [26], MiniGPT-4 [49], Qwen-VL [2], and Vila [24]. Most of these models integrate a language-supervised vision encoder, such as CLIP or SigLIP, with a language model backbone. Beyond these, there is a wider range of visual models available, including self-supervised models [34], segmentation models [20], and diffusion models [37]. Departing from conventional

Features used	MMBench (EN)	POPE	MM-Vet	MME-P	Seed-image	HallusionBench	LLaVA-bench	AI2D	Math Vista	MMMU	OCRBench	ChartQA	DocVQA	InfoVQA
[V]	64.3	86.1	31.1	1510.7	66.0	44.8	64.2	54.7	26.7	35.2	31.2	18.3	27.9	25.7
[V, V' _{t1} , V' _{t2} , V' _{t3}]	66.1	89.4	35.2	1543.5	70.3	46.8	65.0	57.2	28.0	35.6	41.4	24.3	44.5	29.4

Table 4. The comparison between keeping only the lower-level feature [V] and our method, which includes both lower- and higher-level features, clearly demonstrates that maintaining both types of features achieves better performance.

	GQA	MMBench (EN)	MMBench (CN)	VizWiz	POPE	MM-Vet	MME-P	MME-C	Seed-image	HallusionBench	LLaVA-bench	MMStar	Average
Florence-VL 7B	62.7	66.1	55.8	54.5	89.4	35.2	1543.5	316.4	70.3	46.8	65.0	36.8	58.3
Remove Caption Feature V' _{t1}	62.2	64.9	56.1	53.5	89.3	31.8	1477.8	354.3	69.0	44.9	65.2	36.0	57.6
Remove OCR Feature V' _{t2}	62.0	65.6	55.4	56.0	88.8	30.2	1506.3	345.4	67.6	45.4	62.6	35.2	57.3
Remove Grounding Feature V' _{t3}	63.0	66.6	56.8	56.5	88.8	32.9	1494.8	338.9	70.8	44.7	65.1	36.2	58.2

Table 5. Ablation study was conducted by removing one high level image feature at a time, demonstrating that all high-level features are essential for maintaining optimal performance.

vision encoder designs, our work introduces an innovative approach by using the generative vision foundation model Florence-2 as the vision encoder.

While other studies, such as Cambrian [41], Brave [18] and MouSi [10] have explored the advantages of combining multiple visual signals, our approach avoids the added complexity and cost of using multiple vision encoders. Instead, we use a single vision model to generate multiple visual features, which each one emphasizing different perceptual information in the input image. This approach allows us to achieve superior performance with a single vision encoder, surpassing models that rely on multiple vision encoders, such as Cambrian [41].

High-resolution adaptation is commonly applied to increase the input resolution for MLLMs [25]. Besides, models like LLaVA-NeXT [25] and InternVL [7] achieve this by using tiling or adaptive tiling, dividing high-resolution inputs into smaller patches for separate processing. Although our method does not incorporate these techniques, both approaches are compatible and could be combined with our method.

8. Conclusion

In conclusion, Florence-VL uses Florence-2 as a versatile vision encoder, which provides diverse, task-specific visual representations across multiple computer vision tasks like captioning, OCR, and Grounding. By leveraging Depth-Breadth Fusion (DBFusion), we incorporate a range of vi-

sual features from different layers ("Depth") and prompts ("Breadth") to create enriched representations that meet varied perceptual demands of downstream tasks. Our fusion strategy, based on channel concatenation, effectively combines these diverse features, which are then projected as input to the language model.

Through training on a novel data recipe that includes detailed captions for pretraining and diverse instruction tuning data, Florence-VL demonstrates superior alignment between the vision encoder and the LLM, outperforming other models across 25 benchmarks covering vision-centric, knowledge-based, and OCR & Chart tasks. Our analysis underscores the effectiveness of Florence-2's generative capabilities in enhancing MLLM alignment and versatility for a wide range of applications.

For future work, several avenues could further enhance the capabilities and efficiency of Florence-VL. One direction involves improving the DBFusion strategy by exploring more sophisticated fusion techniques that could dynamically adapt the Depth-Breadth balance based on specific downstream task requirements. Additionally, while Florence-2 provides diverse visual representations, future research could explore adaptive vision encoders that select features on-the-fly, optimizing computational efficiency without compromising performance.

References

- [1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 6
- [2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 7
- [3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. 5
- [4] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568, 2021. 5
- [5] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions, 2023. 5
- [6] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 5
- [7] 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. 8
- [8] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431*, 2021. 5
- [9] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers. In *European conference on computer vision*, pages 74–92. Springer, 2022. 2
- [10] Xiaoran Fan, Tao Ji, Changhao Jiang, Shuo Li, Senjie Jin, Sirui Song, Junke Wang, Boyang Hong, Lu Chen, Guodong Zheng, et al. Mousi: Poly-visual-expert vision-language models. *arXiv preprint arXiv:2401.17221*, 2024. 8
- [11] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. 5
- [12] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruva Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024. 5
- [13] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 5
- [14] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 5
- [15] Qidong Huang, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Deciphering cross-modal alignment in large vision-language models with modality integration rate. *arXiv preprint arXiv:2410.07167*, 2024. 4
- [16] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 5
- [17] HuggingFaceM4/Docmatix. <https://huggingface.co/datasets/huggingfacem4/docmatix>. <https://huggingface.co/datasets/HuggingFaceM4/Docmatix>, 2024. 5
- [18] Oğuzhan Fatih Kar, Alessio Tonioni, Petra Poklukar, Achin Kulshrestha, Amir Zamir, and Federico Tombari. Brave: Broadening the visual encoding of vision-language models. *arXiv preprint arXiv:2404.07204*, 2024. 8
- [19] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016. 5
- [20] 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. 1, 7
- [21] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 5
- [22] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 5
- [23] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024. 6
- [24] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeibi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF*

- Conference on Computer Vision and Pattern Recognition, pages 26689–26699, 2024. 6, 7
- [25] 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. 6, 8
- [26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 2, 4, 5, 6, 7
- [27] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 5
- [28] 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, 2024. 5
- [29] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 5
- [30] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 5
- [31] 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. 5
- [32] 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. 5
- [33] 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. 5
- [34] 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. 1, 5, 7
- [35] 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. 1
- [36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 5
- [37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 7
- [38] Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, et al. Eagle: Exploring the design space for multimodal llms with mixture of encoders. *arXiv preprint arXiv:2408.15998*, 2024. 1
- [39] 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. 5
- [40] Vasu Singla, Kaiyu Yue, Sukriti Paul, Reza Shirkavand, Mayuka Jayawardhana, Alireza Ganjdanesh, Heng Huang, Abhinav Bhatele, Gowthami Somepalli, and Tom Goldstein. From pixels to prose: A large dataset of dense image captions, 2024. 5
- [41] 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. 1, 2, 5, 6, 8
- [42] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024. 6
- [43] Lai Wei, Zhiquan Tan, Chenghai Li, Jindong Wang, and Weiran Huang. Large language model evaluation via matrix entropy. *arXiv preprint arXiv:2401.17139*, 2024. 4
- [44] x.ai. Grok 1.5v: The next generation of ai. <https://x.ai/blog/grok-1.5v>, 2023. Accessed: 2024-07-26. 6
- [45] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829, 2024. 1, 2
- [46] Zhiyang Xu, Chao Feng, Rulin Shao, Trevor Ashby, Ying Shen, Di Jin, Yu Cheng, Qifan Wang, and Lifu Huang. Vision-flan: Scaling human-labeled tasks in visual instruction tuning, 2024. 5
- [47] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 5
- [48] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 5
- [49] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language

understanding with advanced large language models. arXiv preprint arXiv:2304.10592, 2023. [1](#), [7](#)