This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

FINECAPTION: Compositional Image Captioning Focusing on Wherever You Want at Any Granularity

Hang Hua[1], Qing Liu[2], Lingzhi Zhang[2], Jing Shi[2], Soo Ye Kim[2], Zhifei Zhang[2], Yilin Wang[2], Jianming Zhang[2], Zhe Lin[2], Jiebo Luo[1]
[1]University of Rochester, [2]Adobe Research
{hhua2,jluo}@cs.rochester.edu, {qingl,lingzzha,jingshi,sooyek,zzhang,yilwang,zlin}@adobe.com

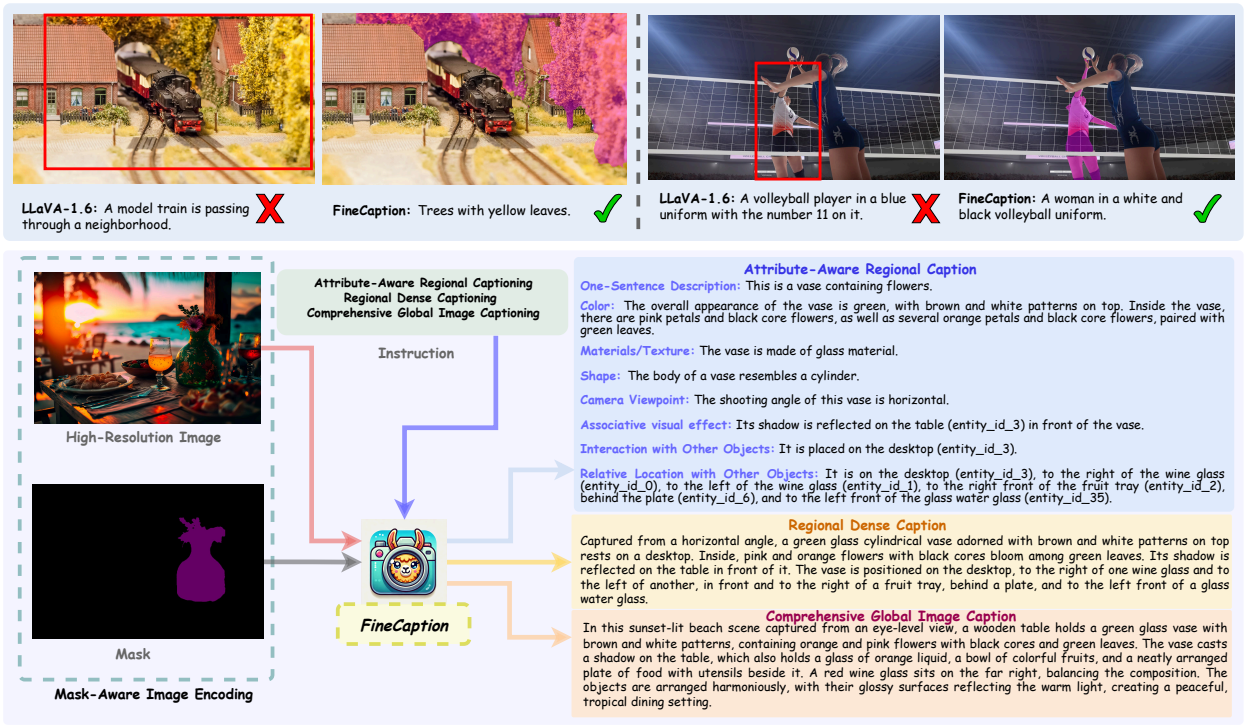


Figure 1. We propose **FINECAPTION**, a novel Vision-Language model with the improved capabilities of **Attribute-Aware Regional Captioning**, **Regional Dense Captioning**, and **Comprehensive Global Image Captioning**. FINECAPTION can recognize arbitrary masks as referential inputs and process high-resolution images. Moreover, models trained using the traditional bounding boxes as region reference are inadequate to precisely describe the region of interest.

Abstract

The advent of large Vision-Language Models (VLMs) has significantly advanced multimodal tasks, enabling more sophisticated and accurate reasoning across various applications, including image and video captioning, visual question answering, and cross-modal retrieval. Despite their superior capabilities, VLMs struggle with fine-grained image regional composition information perception. Specifically, they have difficulty accurately aligning the segmentation masks with the corresponding semantics and precisely describing the compositional aspects of the referred regions. However, compositionality – the ability to understand and generate novel combinations of known visual and textual components – is critical for facilitating coherent reasoning and understanding across modalities by VLMs. To address this issue, we propose FINECAPTION*, a novel VLM that can recognize arbitrary masks as referential inputs and process high-resolution images for compositional image captioning at different granularity levels. To support this endeavor, we introduce* COMPOSITIONCAP*, a new dataset for multi-grained region compositional image captioning, which introduces the task of compositional attribute-aware regional image captioning. Empirical results demonstrate*

the effectiveness of our proposed model compared to other state-of-the-art VLMs. Additionally, we analyze the capabilities of current VLMs in recognizing various visual prompts for compositional region image captioning, highlighting areas for improvement in VLM design and training. https://hanghuacs.github.io/FineCaption/

1. Introduction

Pre-trained vision-language models, such as GPT-4o [1], LLaVA [27], InternVL [5], BLIP [21], and VILA [24], have demonstrated impressive capabilities in complex reasoning, and have achieved remarkable results in various vision-language (VL) tasks. Among these tasks, one fundamental challenge is detailed image content perception and captioning, which involves understanding visual inputs and generating descriptive textual outputs. This capability is essential for applications such as assistive technologies, content accessibility, and enhanced human-computer interaction. While extensive research has focused on improving the captioning abilities of VLMs in areas like generic image captioning [22, 40, 44, 46, 52], dense image captioning [16, 31, 42], and referring image captioning [32, 35, 45, 51, 56], less attention has been devoted to region compositional captioning and detailed region-level captioning tasks. These tasks require models to recognize both compositional attribute prompts and region prompts, generating captions specific to the attributes and regions of interest.

Previous works have aimed at enabling region-level understanding in VLMs. Methods such as Kosmos-2 [32], Shikra [4], and GPT4RoI [54] have attempted to process regions specified by bounding boxes, leveraging visual instruction tuning with object-level spatial features. Other approaches aim to enable VLMs to recognize arbitrary visual prompts and focus on the referred regions by incorporating overlaid images and visual prompts as input [3, 45]. However, these methods have limitations. As illustrated in previous work [34, 35, 51, 55] and the examples in Figure 1, bounding boxes are inadequate for providing precise references to image regions (The IoU between masks and their bounding boxes is 56.11 for our data.). Moreover, freeform overlaid visual prompts are suboptimal for region-level understanding tasks, as they often confuse VLMs; models frequently interpret the visual prompts as integral parts of the image's semantic content. Therefore, utilizing masks as region references is an ideal solution for regional understanding tasks.

In this study, we address these limitations from both the model design and dataset construction perspectives. Specifically, we propose a novel VLM, FINECAPTION, capable of multi-grained region compositional attribute captioning. To better capture detailed compositional information and accurately recognize mask-referred regions, we design a new architecture that integrates a mask-aware low-resolution encoder with multiple high-resolution encoders. Table 1 summarizes the distinct capabilities of our model compared to existing models for regional image understanding. For the mask-aware encoding, we follow the method of Alpha-CLIP [35] by introducing an extra convolution layer to the CLIP image encoder, which incorporates binary masks as the alpha channel for RGB images. Our experiments demonstrate that aggregating multiple high-resolution encoders enhances the model's ability to perceive detailed information in image regions. Therefore, we utilize ConvNeXT [28] and the SAM encoder [18] as high-resolution encoders, supporting image encoding at the 1024 x 1024 resolution. Furthermore, we introduce a new human-annotated, high-quality dataset, COMPOSITIONCAP, to improve models' capabilities in **Attribute-Aware Regional Captioning (AARC)**, **Regional Dense Captioning (RDC)**, and **Comprehensive Global Image Captioning (CGIC)**. Our dataset encompasses diverse scenes and introduces the task of compositional aspect-aware regional image captioning, including 18 different compositional attributes. Detailed explanations of these aspects are provided in the appendix.

Unlike the Referring Expression Generation (REG) task – which involves automatically generating a referring expression that uniquely identifies a particular object in an image by distinguishing it from other objects in the same scene – COMPOSITIONCAP emphasizes the generation of multi-grained compositional expressions for regions of interest. In this task, models are required to describe masked regions with detailed and comprehensive compositional information. Our objective is not to differentiate objects from one another but rather to provide rich, attribute-aware descriptions of the specified regions [17, 47]. The goal is to create captions that not only mention objects or regions but also elaborate on their compositional aspects and how they interact or relate to the broader scene.

Empirical results demonstrate the superior performance of FINECAPTION on region compositional image captioning tasks compared to other strong VLMs, especially GPT-4 [1], and LLaMA-3.2 [8] which is recognized for its strong capabilities across a wide range of tasks.

In summary, our contributions are threefold:

- We propose FINECAPTION, a novel vision-language model with enhanced capabilities for mask-referring multi-grained image compositional captioning. FINECAPTION is equipped with a mask-aware image encoder for mask-referring and a high-resolution encoder for fine-grained perception of compositional information. Empirical results demonstrate the superior performance of our model on mask-referring image compositional captioning tasks compared to other strong VLMs.
- We propose COMPOSITIONCAP, a novel, human-annotated, high-quality benchmark for multi-grained mask-referring image compositional captioning. COM-

Model	Mask Referencing	High Resolution	Region Attribute Captioning	Region Dense Captioning
ViP-LLaVA [3]	✗	✗	✗	✗
Ferret [45]	✗	✗	✗	✗
Ferret-v2 [53]	✗	✓	✗	✗
GPT4RoI [54]	✗	✗	✗	✗
VCoder [15]	✓	✗	✗	✗
Osprey [51]	✓	✗	✗	✓
Alpha-CLIP [35]	✓	✗	✗	✗
GLaMM [34]	✓	✗	✓	✗
RegionGPT [9]	✓	✗	✗	✗
OMG-LLaVA [55]	✓	✗	✗	✗
FINECAPTION (ours)	✓	✓	✓	✓

Table 1. Comparison of the capabilities of FINECAPTION and other related VLMs: "Mask Referring" indicates whether the model's encoder can accept a mask input as a reference, "High-Resolution" specifies whether the model supports high-resolution image encoding.

POSITIONCAP encompasses 18 different compositional aspects and offers three levels of captioning granularity: Attribute-Aware Regional Captioning, Regional Dense Captioning, and Comprehensive Global Image Captioning.

- We analyze the capabilities of VLMs in generating compositional aspect-aware captions for detailed region descriptions and in handling input images with region-referring tasks. This analysis highlights key areas where models can improve in capturing compositional attributes – such as color, body gesture, material, and texture – and in accurately distinguishing specific regions. These insights offer directions for future development in both caption generation and recognition tasks.

2. Related Work

2.1. Pre-trained Vision-Language Models

Vision-language models [5, 6, 13, 20, 27, 33, 37, 38, 43] strive for multimodal intelligence by jointly processing visual and linguistic information. Inspired by the remarkable success of recent large language models (LLMs) [7, 11, 39], researchers are now exploring large VLMs that combine pre-trained visual encoders and language decoders to tackle complex multimodal tasks. Flamingo [2] and BLIP-2 [21] are two of the early works that explore the integration of LLMs into vision-language pre-training. These models are trained as VL foundation models. Beginning with LLaVA [27], researchers have used LLM-synthesized instruction-following chat data in VQA format for instruction tuning, achieving significantly improved results [12, 14, 36, 48, 49]. Subsequent studies have expanded to explore the broader capabilities [10, 14, 23, 50] of multimodal LLMs. However, these efforts have placed less emphasis on enhancing models' ability to describe image contents focusing on specific regions and attributes.

2.2. Vision-Language Models for Region-level Image Understanding

Recent works like Kosmos-2 [32], Shikra [4], GPT4RoI [54], Ferret [45], and Sphinx [25] aim to enable region-specific interactions in vision-language models (VLMs). A straightforward approach is to provide bounding box coordinates to the models, as employed by Kosmos-2, Shikra, RegionGPT [9], and Sphinx. Methods such as ViP-LLaVA [3] and Ferret explore recognizing arbitrary freeform visual prompts by overlaying visual prompts onto images. However, this input format can confuse VLMs, as the models often interpret the visual prompts as part of the image's semantic content. Some models, including Alpha-CLIP [35], OMG-LLaVA [55], GLaMM [34], Osprey [51], and VCoder [15], use masks for region referencing, offering more fine-grained control. However, they process images at resolutions ranging from 224×224 to 448×448, limiting their ability to capture fine-grained compositional details. High-resolution image encoding enhances a model's capacity for detailed perception, which is crucial for understanding and describing complex visual information accurately.

3. The COMPOSITIONCAP Dataset

To enhance the capabilities of vision-language models (VLMs) in Attribute-Aware Regional Captioning, Regional Dense Captioning, and Global Dense Captioning, we constructed a new human-annotated, high-quality dataset named COMPOSITIONCAP. We began by collecting high-quality images from various stock image sources such as Adobe Stock Images and iStock. We then obtained entity masks using the SAM model [18]. Human annotators were tasked with describing the attributes of these entities to generate corresponding region compositional attribute descriptions. This process resulted in 14,590 entities across 5,392 images and a total of 186,490 attribute descriptions. To enhance models' capabilities in comprehensive global image captioning, we also construct instruction-tuning data based on DOCCI

[31], a dataset containing long, human-annotated English descriptions for 15,000 images.

In addition, we built out-of-domain test set for COMPOSITIONCAP sourced from the Open Images dataset [19]. We selected 1,000 images featuring diverse and complex scenes for annotation. The test set contains 7,215 masked entities with 19,326 attribute-specific region captions. The attributes in COMPOSITIONCAP include: 1) **Category Name**, 2) **Body Shape**, 3) **Skin Texture and Color**, 4) **Clothing, Shoes, Accessories**, 5) **Interaction with Other Objects**, 6) **Body Pose/Gesture**, 7) **Other Attributes**, 8) **Relative Location with Other Objects**, 9) **Color**, 10) **Materials/Texture**, 11) **Camera Viewpoint**, 12) **Associative Visual Effect**, 13) **Shape**, 14) **Facial Expression**, 15) **Hair**, 16) **Age Range**, 17) **Object Pose for Deformable Objects**, 18) **Style**. Figure 2 shows the resolution distribution of COMPOSITIONCAP, and Figure 3 illustrates the proportion of each attribute in our dataset. More statistical results and examples are included in the appendix, where we also provide a detailed explanation of these attributes.

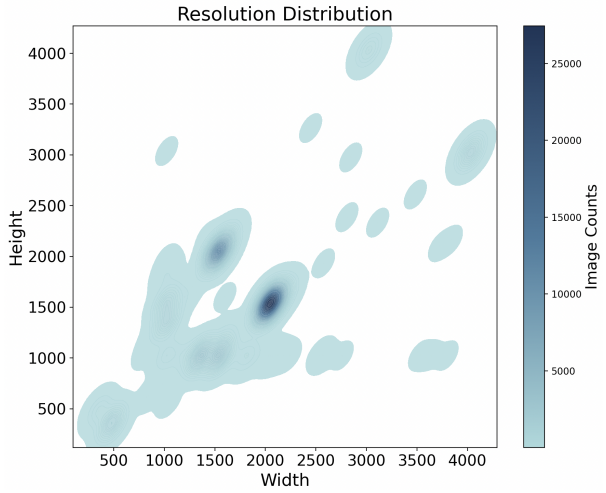


Figure 2. The image resolution distribution across data points in COMPOSITIONCAP reflects the high overall quality of images.

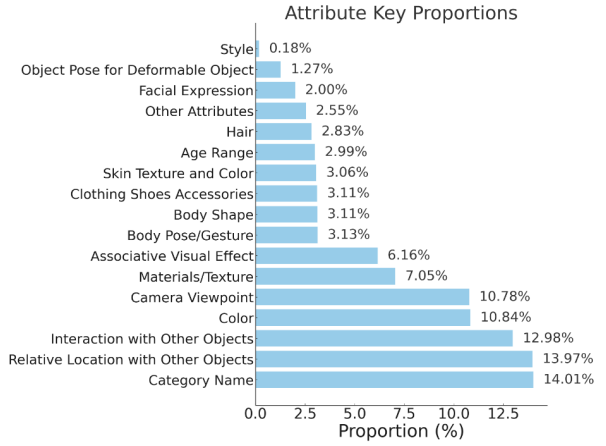


Figure 3. Distribution of attributes in COMPOSITIONCAP.

4. Methodology

Our proposed model integrates mask-aware and high-resolution features through a multi-resolution image encoding framework, as shown in Figure 4. This section details each component, including the mask-aware encoder, high-resolution encoders, channel-wise feature fusion, and integration with a large language model via an adapter module.

4.1. Input Representation

The input of FINECAPTION encoder includes:

1. **Low-resolution image**: $I_{\text{LR}} \in \mathbb{R}^{H_{\text{LR}} \times W_{\text{LR}} \times 3}$, where $H_{\text{LR}} = W_{\text{LR}} = 336$.
2. **High-resolution image**: $I_{\text{HR}} \in \mathbb{R}^{H_{\text{HR}} \times W_{\text{HR}} \times 3}$, where $H_{\text{HR}} = W_{\text{HR}} = 1024$.
3. **Binary mask**: $M \in \mathbb{R}^{H_{\text{LR}} \times W_{\text{LR}} \times 1}$, indicating the region of interest within I_{LR}.

4.2. Mask-Aware Encoding

To align the mask-referred region with the input images, we introduce an additional alpha channel to the CLIP encoder's embedding layer Conv_{α}, following the approach of Alpha-CLIP [35]. This method enables us to encode the mask independently, preserving the original content of the images.

First, the low-resolution image I_{LR} is passed through the encoder's standard patch embedding layer:

$$\mathbf{E}_{\text{patch}} = \text{Conv}_{\text{RGB}}(I_{\text{LR}}), \tag{1}$$

where $\text{Conv}_{\text{patch}}(\cdot)$ maps I_{LR} into patch embeddings $\mathbf{E}_{\text{patch}} \in \mathbb{R}^{C \times H' \times W'}$ (C is the number of output channels, H' and W' are the spatial dimensions of the patch embeddings).

Simultaneously, the mask M is processed by the additional convolutional layer:

$$\mathbf{E}_{\text{mask}} = \text{Conv}_{\alpha}(M), \tag{2}$$

where $\text{Conv}_{\alpha}(\cdot)$ is a convolutional layer with the same parameters as Conv_{RGB} except for the input channels, which is set to 1 for the mask. This layer outputs mask embeddings $\mathbf{E}_{\text{mask}} \in \mathbb{R}^{C \times H' \times W'}$.

The patch embeddings and mask embeddings are then combined and flattened to form the input sequence:

$$\mathbf{E}_{\text{seq}} = \text{Flatten}(\mathbf{E}_{\text{patch}} + \mathbf{E}_{\text{mask}})^{\top} \in \mathbb{R}^{N \times C}, \tag{3}$$

where $N = H' \times W'$ is the number of patches.

We then add class embeddings $\mathbf{E}_{\text{class}}$ and positional embeddings $\mathbf{E}_{\text{pos}}$:

$$\mathbf{E} = [\mathbf{E}_{\text{class}}; \mathbf{E}_{\text{seq}}] + \mathbf{E}_{\text{pos}}, \tag{4}$$

resulting in input embeddings for the encoder.

The mask-aware feature map $\mathbf{F}_M$ is obtained by passing $\mathbf{E}$ through the mask-aware CLIP encoder:

$$\mathbf{F}_M = \text{Encoder}_M(\mathbf{E}), \tag{5}$$

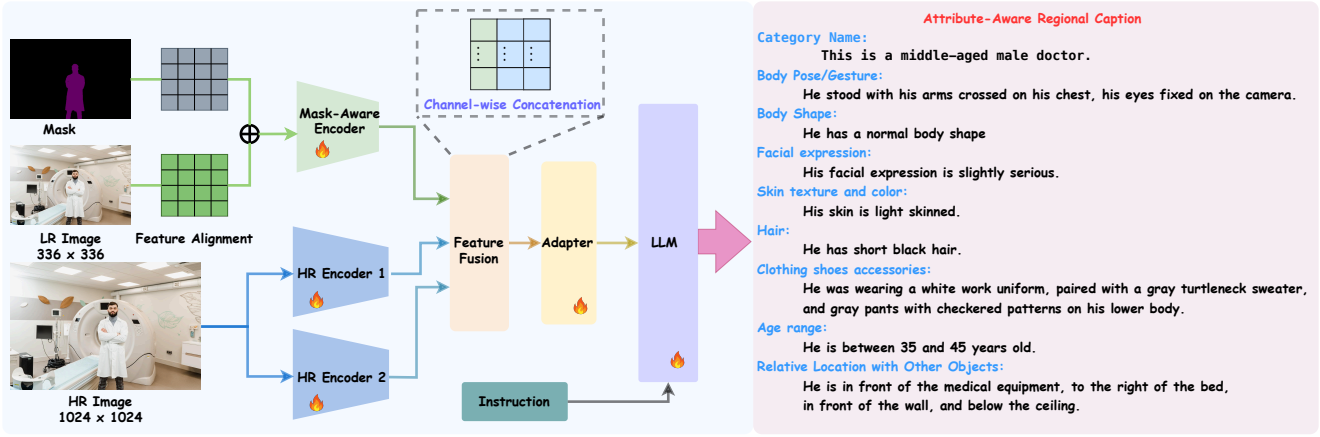


Figure 4. Overview of FINECAPTION: The model incorporates a mask-aware visual encoder and two high-resolution encoders (ConvNext and SAM), enabling precise recognition of mask references and the perception of detailed compositional and spatial information for images.

where $\mathbf{F}_M \in \mathbb{R}^{N \times C_M}$ captures the region information of the low-resolution image and the mask.

4.3. High-Resolution Encoding

To capture more detailed spatial information for language models, we employ two high-resolution encoders to extract fine-grained features: ConvNext [28] and SAM [18] encoders.

$$\mathbf{F}_{\text{HR1}} = \text{Encoder}_{\text{ConvNeXT}}(I_{\text{HR}}), \tag{6}$$

$$\mathbf{F}_{\text{HR2}} = \text{Encoder}_{\text{SAM}}(I_{\text{HR}}), \tag{7}$$

where $\mathbf{F}_{\text{HR1}} \in \mathbb{R}^{N' \times C_{\text{HR1}}}$ and $\mathbf{F}_{\text{HR2}} \in \mathbb{R}^{N' \times C_{\text{HR2}}}$ ($N' = H'' \times W''$) are the resulting feature maps from each encoder.

4.4. Feature Fusion

The mask-aware and high-resolution features are unified through a channel-wise fusion module. First, the mask-aware feature map $\mathbf{F}_M$ is resized as $\overline{\mathbf{F}}_M$ and then interpolated to match the spatial dimensions of the high-resolution feature maps and finally flattened into a sequence:

$$\mathbf{F}'_M = \text{Flatten}(\text{Interpolate}(\overline{\mathbf{F}}_M)), \tag{8}$$

where $\mathbf{F}'_M \in \mathbb{R}^{N' \times C_M}$.

The three feature maps are concatenated along the channel dimension:

$$\mathbf{F}_{\text{fusion}} = [\mathbf{F}'_M; \mathbf{F}_{\text{HR1}}; \mathbf{F}_{\text{HR2}}], \tag{9}$$

resulting in $\mathbf{F}_{\text{fusion}} \in \mathbb{R}^{H'' \times W'' \times C_{\text{fusion}}}$, where $C_{\text{fusion}} = C_M + C_{\text{HR1}} + C_{\text{HR2}}$.

4.5. Adapter and Language Model Integration

The fused feature map is then mapped into the LLM's word embedding space by an adapter:

$$\mathbf{F}_{\text{adapted}} = \text{Adapter}(\mathbf{F}_{\text{fusion}}), \tag{10}$$

where $\mathbf{F}_{\text{adapted}} \in \mathbb{R}^{N' \times D'}$, with N' and D' representing the adapted sequence length and embedding dimension for the LLM. These adapted features, along with an instruction $\mathbf{I}$, guide the LLM's response generation:

$$y = \text{LLM}(\mathbf{F}_{\text{adapted}}, \mathbf{I}), \tag{11}$$

where y is the output generated by the LLM, conditioned on the visual features and task-specific instructions.

4.6. Training Objective

We then train the model end-to-end using negative log-likelihood loss:

$$\mathcal{L} = -\sum_{\mathcal{D}} \sum_{i=1}^{N} \log p(y_i \mid \mathbf{F}_{\text{adapted}}, \mathbf{I}, y_{<i}), \tag{12}$$

where $\mathcal{D}$ represents the training dataset, N is the total number of tokens or features in each sequence, and y_i represents the language tokens in the generated output.

4.7. Training Strategy

Stage 1: Pre-training. Similar to LLaVA [27], this stage aims to optimize the projector to align the visual features with the word embeddings of the language model (LLM). Thus, the image encoders and the LLM are kept frozen during pre-training. We use LLaVA-Pretrain [27] dataset for training, with the mask in this stage highlighting all regions of the image.

Stage 2: Image-Mask Alignment Pre-training. The goal of this stage is to align the image and mask features for the mask-aware encoder. We use the data including COMPOSITIONCAP, GranD [34], RefCOCO, RefCOCO+ [17], and RefCOCOg [47] for training. In this stage, only the mask-aware encoder is trainable.

Stage 3: Region Attribute-Aware Instruction Tuning. In the final stage, we use the COMPOSITIONCAP training set,

Model	Region Referral			Semantic Evaluation				
	Visual Prompt	Resolution	# Image Token	ROUGE-L↑	BLEU-4↑	METEOR↑	CIDEr↑	BERT Score↑
Zero-Shot Learning								
Kosmos-2 [32]	Bbox	224	256	9.21	0.14	1.98	1.07	37.69
Alpha-CLIP-13B [35]	Mask	336	576	13.89	0.51	5.94	2.68	42.01
Qwen2-VL-7B [41]	Bbox	AnyRes	-	14.12	0.57	6.18	2.74	42.97
Ferret-13B [45]	MContour	336	576	15.01	1.06	5.86	3.12	43.82
ViP-LLaVA-13B [3]	MContour	336	576	15.47	1.48	5.76	3.84	44.29
LLaMA-3.2-11B-Vision-Instruction [8]	Bbox	-	-	15.64	1.59	9.73	3.95	44.53
LLaMA-3.2-90B-Vision-Instruction [8]	Bbox	-	-	16.21	1.75	11.70	4.53	48.29
InternVL-2-40B [8]	Bbox	1792	4096	16.21	1.79	11.91	4.63	48.38
GPT-4o [1]	Bbox	-	-	17.87	3.21	12.87	6.49	49.85
Supervised Learning								
Qwen2-VL-7B [41]	Bbox	AnyRes	-	31.59	9.11	13.56	90.32	75.86
LLaVA-1.6-13B [27]	Bbox	AnyRes	576	31.72	9.35	13.64	90.71	75.89
VILA1.5-8B [24]	Bbox	336	144	31.87	9.03	13.79	90.01	75.95
ViP-LLaVA-13B [3]	MContour	336	576	32.42	9.97	14.82	91.44	76.77
Alpha-CLIP-13B [35]	Mask	336	576	35.68	10.96	16.11	93.85	77.66
LLaVA-HR-X [29]	Bbox	1024	1024	35.97	11.25	16.57	95.12	78.08
LLaMA-3.2-11B-Vision [8]	Bbox	-	-	38.14	12.87	18.31	99.11	78.94
FINECAPTION-8B (ours)	Mask	1024	1024	**41.05**	**14.46**	**22.01**	**127.95**	**80.97**

Table 2. Comparison of the capabilities of FINECAPTION and other related VLMs including both open-sourced models and API-based models . The column "Visual Prompt" indicates the format of input region-referral, "Resolution" indicates the encoding resolution for input images. We evaluate the models' performance under both the zero-shot and supervised learning settings.

allowing all parameters to be trainable. This stage fine-tunes the model to accurately handle compositional referring expressions, enhancing its ability to produce detailed, attribute-aware captions.

5. Experiments

In this section, we comprehensively evaluate FINECAPTION against state-of-the-art VLMs on mask-referring, multi-grained image compositional captioning.

5.1. Baseline Models

We selected a range of strong VLMs as baselines for both zero-shot and supervised learning settings to highlight FINECAPTION's performance in mask-referring, multi-grained image compositional captioning. These include: Kosmos-2 [32], Alpha-CLIP [35], Qwen2-VL [41], Ferret [45], ViP-LLaVA [3], LLaMA-3.2[8], GPT-4o [1], ViP-LLaVA [3], LLaVA-1.6 [26], LLaVA-HR [29], and the API-based model GPT-4o [1].

5.2. Implementation Details

In our mixture-of-encoder architecture, we incorporate CLIP-ViT-L/14@336p [33] with an additional alpha channel – similar to Alpha-CLIP [35] – for mask-aware low-resolution encoding. For high-resolution image encoding, we use ConvNeXt-XXL@1024p [28] and SAM [18] encoders. As the language model decoder, we employ LLaMA-3.1-8B-Instruction [39]. All encoder and decoder parameters are trainable, and we adopt the default hyperparameters from LLaVA-1.6 [26]. We trained our model on 8 Nvidia-A100 GPUs, and the whole training process took 38 hours.

Model	AARC	RDC
	GPT4-as-a-Judge ↑	Grounding Score (Acc@0.5) ↑
ViP-LLaVA-13B [3]	38.46	74.37
Alpha-CLIP-13B [35]	43.89	77.61
LLaVA-HR-X [29]	45.76	79.15
LLaMA-3.2-11B-Vision [8]	50.26	81.01
FINECAPTION-8B (ours)	**56.84**	**83.49**

Table 3. GPT-4-as-a-Judge score and Grounding score for Attribute-Aware Regional Captioning (AARC) and Regional Dense Captioning (RDC) tasks. The GPT-4-as-a-Judge score indicates the binary accuracy for correct attribute description, and the Grounding score is IoU Acc@0.5.

5.3. Experimental Results on COMPOSITIONCAP

We evaluate models under both the zero-shot and supervised learning settings across multiple metrics. Table 2 presents the main results of FINECAPTION alongside baseline models. From the zero-shot learning results, we observe that most models struggle to recognize both region referrals and attribute instructions precisely. Even strong open-source models like InternVL-2-40B and LLaMA-3.2-90B perform suboptimally on tasks introduced by COMPOSITIONCAP, achieving only 11.7 and 11.9 in METEOR, and 5.5 and 4.63 in CIDEr, respectively. Additionally, the performance of the API-based model GPT-4 is still unsatisfactory in this context. In contrast, the supervised learning results demonstrate that models trained with our data exhibit significantly stronger performance. FINECAPTION outperforms all other models due to the incorporation of mask-aware encoders and multiple high-resolution encoders, which enhance its ability to recognize region referrals and attribute-specific instructions effectively.

5.4. Fine-Grained Evaluation

In addition to traditional image captioning evaluation methods, we designed fine-grained evaluation techniques for the Attribute-Aware Regional Captioning (AARC) and Regional Dense Captioning (RDC) tasks. Following previous work [26, 45], we use GPT-4 as a judge for the AARC task, where GPT-4 is fed the model's predictions, the image, and the ground truth to assess whether the model properly describes the attributes of the referred region. To better evaluate performance on the RDC task, we trained an LLaVA-1.6-13B model using the COMPOSITIONCAP test set to predict bounding boxes for the region descriptions. We calculate Acc@0.5 to assess the quality of the predicted regional dense captions. From Table 3, we observe that FINECAPTION achieves 56.84 on the AARC task and 83.49 on the RDC task, significantly outperforming the baseline models and indicating the superior capability of our model in these tasks. We provide the evaluation details in the appendix.

5.5. Results on Referring Expression Generation Tasks

To further evaluate our model's ability to recognize region referrals accurately, we conduct experiments on traditional REG tasks. Following the evaluation framework of previous works [32, 51], we use the RefCOCOg test set and calculate METEOR and CIDEr scores for evaluation. Table 4 presents the performance of different models compared to FINECAPTION. Our model significantly outperforms previous models, indicating its superior capability in region description.

Model	Region Referral	RefCOCOg	
		METEOR	CIDEr
GRIT [30]	Bbox	15.2	71.6
Kosmos-2 [32]	Bbox	14.1	62.3
OMG-LLaVA [55]	Mask	15.3	-
GLaMM [34]	Bbox	16.2	105.0
Osprey [51]	Mask	16.6	108.3
Alpha-CLIP+LLaVA [35]	Mask	16.7	109.2
RegionGPT [9]	Mask	16.9	109.9
ControlCap [56]	Bbox	17.0	111.4
FINECAPTION-8B (ours)	**Mask**	**17.5**	**118.2**

Table 4. Region captioning performance evaluated on the test set of RefCOCOg.

6. Analysis

6.1. Ablation Study

Models Design. To better evaluate the effectiveness of our proposed model, we conduct an ablation study from the perspective of vision encoder fusion and different visual referring inputs. The ablation includes two parts, model design and region referral format. In model design, we investigate the effects of different visual encoder architectures and feature fusion strategies. From Table 5 we can observe that models with mask-aware low-resolution encoders only achieve lower performance compared to those that incorporate high-resolution encoders. Specifically, the FINECAPTION with LR Encoding Only variant attains a CIDEr score of 97.62, which is significantly less than the full model's score of 127.95. This indicates that relying solely on low-resolution features is insufficient for capturing the detailed information necessary for high-quality caption generation.

When investigating different high-resolution encoder architectures, we find that FINECAPTION w/ ConvNeXt Encoder outperforms FINECAPTION w/ SAM Encoder, achieving a higher CIDEr score of 109.25 versus 106.74. This indicates that the combination of LR and HR encoder architecture significantly impacts model performance. In addition, the fusion strategy plays a crucial role: the FINECAPTION + Self-Attention Fusion variant yields better results than the FINECAPTION + Sequence Append Fusion. The latter underperforms with a CIDEr score of 93.26, suggesting that simply appending sequences is a less effective method for feature fusion.

The Impact of Referring Format. please make this paragraph more brief: In the region referral format ablation, replacing the precise segmentation masks with bounding boxes (FINECAPTION + Bounding Box) and mask contours (FINECAPTION + Mask Contour) leads to noticeable performance drops, with CIDEr scores of 96.85 and 95.97, respectively. This decline underscores the importance of using detailed mask information for accurately grounding visual content within the image.

Overall, the ablation study confirms that incorporating both low-resolution and high-resolution encoders, along with effective fusion strategies and precise region referrals, substantially enhances the model's ability to generate detailed and accurate captions. These results validate the effectiveness of our model.

6.2. Qualitative Analysis

Figure 5 compares the performance of different models. The examples illustrate that while GPT-4o can recognize some attribute instructions, it struggles with identifying fine-grained region references and often includes extraneous information unrelated to the specified attributes or regions. In contrast, our model, FINECAPTION, excels in both attribute recognition and precise region localization, generating accurate and succinct descriptions focused exclusively on the specified regions and attributes, thereby reflecting a superior understanding of fine-grained compositional details.

7. Conclusion

In this paper, we present FINECAPTION, a vision-language model designed for mask-referring, multi-grained image compositional captioning. By combining a mask-aware encoder with high-resolution encoders, FINECAPTION effec-

Method	ROUGE-L	BLEU-4	METEOR	CIDEr	BERT Score
FINECAPTION	**41.05**	**14.46**	**22.01**	**127.95**	**80.97**
Model Architecture					
FINECAPTION w/ LR Encoding Only	37.92	11.67	17.86	97.62	78.11
FINECAPTION w/ ConvNeXt	39.87	13.42	20.97	109.25	79.83
FINECAPTION w/ SAM	38.97	12.95	20.01	106.74	79.36
FINECAPTION + Self-Attn Fusion	38.21	12.26	19.73	104.85	79.24
FINECAPTION + Sequence Append Fusion	36.11	10.13	16.07	93.26	77.25
Region Referral					
FINECAPTION + Bbox	37.10	11.12	17.21	96.85	77.91
FINECAPTION + MContour	36.59	10.62	16.88	95.97	77.03

Table 5. Performance comparison of different model designs and region referrals.

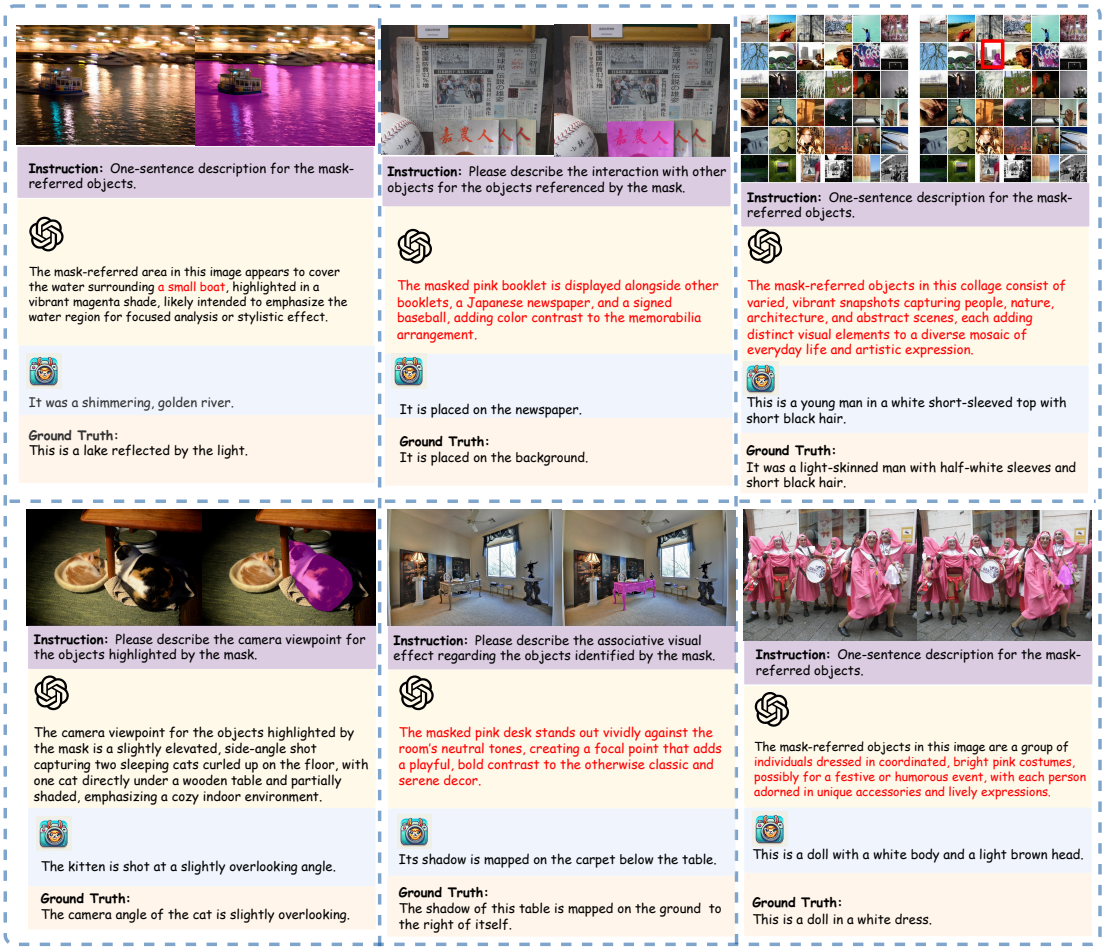


Figure 5. FINECAPTION provides accurate and concise descriptions focused on specified attributes and regions, while GPT-4o often misses fine-grained references and includes irrelevant information.

tively perceives fine-grained details and accurately recognizes masked regions, outperforming models like GPT-4 and LLaMA-3.2 in detailed region-level tasks. To support our model, we created COMPOSITIONCAP, a human-annotated dataset featuring 18 compositional aspects across diverse scenes and offering three levels of captioning granularity. This dataset fills a gap by emphasizing rich, attribute-aware descriptions of specified regions. Our work lays a foundation for advanced region-level understanding in vision-language models. We hope FINECAPTION and COMPOSITIONCAP become valuable resources for future research in detailed image perception and captioning.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 6

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 3

[3] Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. Vip-llava: Making large multimodal models understand arbitrary visual prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12914–12923, 2024. 2, 3, 6

[4] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 2, 3

[5] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 2, 3

[6] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatial-rgpt: Grounded spatial reasoning in vision language model. *arXiv preprint arXiv:2406.01584*, 2024. 3

[7] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, march 2023. *URL https://lmsys. org/blog/2023-03-30-vicuna*, 3(5), 2023. 3

[8] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2, 6

[9] Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, and Sifei Liu. Regiongpt: Towards region understanding vision language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13796–13806, 2024. 3, 7

[10] Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. Promptcap: Prompt-guided image captioning for vqa with gpt-3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2963–2975, 2023. 3

[11] Hang Hua, Xingjian Li, Dejing Dou, Cheng-Zhong Xu, and Jiebo Luo. Improving pretrained language model fine-tuning with noise stability regularization. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 3

[12] Hang Hua, Jing Shi, Kushal Kafle, Simon Jenni, Daoan Zhang, John Collomosse, Scott Cohen, and Jiebo Luo. Finematch: Aspect-based fine-grained image and text mismatch detection and correction. *arXiv preprint arXiv:2404.14715*, 2024. 3

[13] Hang Hua, Yunlong Tang, Chenliang Xu, and Jiebo Luo. V2xum-llm: Cross-modal video summarization with temporal prompt instruction tuning. *arXiv preprint arXiv:2404.12353*, 2024. 3

[14] Hang Hua, Yunlong Tang, Ziyun Zeng, Liangliang Cao, Zhengyuan Yang, Hangfeng He, Chenliang Xu, and Jiebo Luo. Mmcomposition: Revisiting the compositionality of pre-trained vision-language models. *arXiv preprint arXiv:2410.09733*, 2024. 3

[15] Jitesh Jain, Jianwei Yang, and Humphrey Shi. Vcoder: Versatile vision encoders for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27992–28002, 2024. 3

[16] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4565–4574, 2016. 2

[17] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 2, 5

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 3, 5, 6

[19] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020. 4

[20] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 3

[21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2, 3

[22] Jiaxuan Li, Duc Minh Vo, Akihiro Sugimoto, and Hideki Nakayama. Evcap: Retrieval-augmented image captioning with external visual-name memory for open-world comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13733–13742, 2024. 2

[23] Jingyang Lin, Hang Hua, Ming Chen, Yikang Li, Jenhao Hsiao, Chiuman Ho, and Jiebo Luo. Videoxum: Cross-modal

visual and textural summarization of videos. *IEEE Transactions on Multimedia*, 2023. 3

[24] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models, 2023. 2, 6

[25] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023. 3

[26] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 6, 7

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2, 3, 5, 6

[28] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 2, 5, 6

[29] Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. *arXiv preprint arXiv:2403.03003*, 2024. 6

[30] Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. Generative representational instruction tuning. *arXiv preprint arXiv:2402.09906*, 2024. 7

[31] Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, et al. Docci: Descriptions of connected and contrasting images. *arXiv preprint arXiv:2404.19753*, 2024. 2, 4

[32] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 2, 3, 6, 7

[33] 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. 3, 6

[34] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024. 2, 3, 5, 7

[35] Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alpha-clip: A clip model focusing on wherever you want. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13019–13029, 2024. 2, 3, 4, 6, 7

[36] Yunlong Tang, Junjia Guo, Hang Hua, Susan Liang, Mingqian Feng, Xinyang Li, Rui Mao, Chao Huang, Jing Bi, Zeliang Zhang, et al. Vidcomposition: Can mllms analyze compositions in compiled videos? *arXiv preprint arXiv:2411.10979*, 2024. 3

[37] Yunlong Tang, Daiki Shimada, Jing Bi, and Chenliang Xu. Avicuna: Audio-visual llm with interleaver and context-boundary alignment for temporal referential dialogue. *arXiv preprint arXiv:2403.16276*, 2024. 3

[38] 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. 3

[39] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3, 6

[40] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 2

[41] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 6

[42] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. In *European Conference on Computer Vision*, pages 207–224. Springer, 2025. 2

[43] Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, et al. mplug-2: A modularized multi-modal foundation model across text, image and video. In *International Conference on Machine Learning*, pages 38728–38748. PMLR, 2023. 3

[44] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 2

[45] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023. 2, 3, 6, 7

[46] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016. 2

[47] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European*

Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14, pages 69–85. Springer, 2016. 2, 5

[48] 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. 3

[49] Yongsheng Yu and Jiebo Luo. Chain-of-thought prompting for demographic inference with large multimodal models. *arXiv preprint arXiv:2405.15687*, 2024. 3

[50] Yongsheng Yu, Ziyun Zeng, Hang Hua, Jianlong Fu, and Jiebo Luo. Promptfix: You prompt and we fix the photo. *arXiv preprint arXiv:2405.16785*, 2024. 3

[51] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28202–28211, 2024. 2, 3, 7

[52] Zequn Zeng, Yan Xie, Hao Zhang, Chiyu Chen, Bo Chen, and Zhengjue Wang. Meacap: Memory-augmented zero-shot image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14100–14110, 2024. 2

[53] Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, et al. Ferret-v2: An improved baseline for referring and grounding with large language models. *arXiv preprint arXiv:2404.07973*, 2024. 3

[54] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023. 2, 3

[55] Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *arXiv preprint arXiv:2406.19389*, 2024. 2, 3, 7

[56] Yuzhong Zhao, Yue Liu, Zonghao Guo, Weijia Wu, Chen Gong, Qixiang Ye, and Fang Wan. Controlcap: Controllable region-level captioning. In *European Conference on Computer Vision*, pages 21–38. Springer, 2025. 2, 7