ViP-LLaVA: Making Large Multimodal Models Understand Arbitrary Visual Prompts

Mu Cai[1] Haotian Liu[1] Siva Karthik Mustikovela[2] Gregory P. Meyer[2]
Yuning Chai[2] Dennis Park[2] Yong Jae Lee[1,2]
[1]University of Wisconsin–Madison [2]Cruise LLC
https://vip-llava.github.io

Abstract

While existing large vision-language multimodal models focus on whole image understanding, there is a prominent gap in achieving region-specific comprehension. Current approaches that use textual coordinates or spatial encodings often fail to provide a user-friendly interface for visual prompting. To address this challenge, we introduce a novel multimodal model capable of decoding arbitrary (free-form) visual prompts. This allows users to intuitively mark images and interact with the model using natural cues like a "red bounding box" or "pointed arrow". Our simple design directly overlays visual markers onto the RGB image, eliminating the need for complex region encodings, yet achieves state-of-the-art performance on region-understanding tasks like Visual7W, PointQA, and Visual Commonsense Reasoning benchmark. Furthermore, we present ViP-Bench, a comprehensive benchmark to assess the capability of models in understanding visual prompts across multiple dimensions, enabling future research in this domain. Code, data, and model are publicly available.

1. Introduction

Large language models (LLMs) like ChatGPT [21], GPT4 [22], and Bard [9] have recently gained significant attention for their strong reasoning and generalization capabilities, and their ability to chat in a human-like manner. In particular, models such as GPT-4V(ision) [20], which incorporate visual information, have demonstrated human-level perception and reasoning capabilities [36]. This has spurred the development of similar open-source models that aim to replicate or even surpass the proprietary models' performance.

Despite their capabilities, current models, including seminal ones like LLaVA [14, 15] and MiniGPT-4 [42], focus predominantly on whole-image understanding; in other words, they lack the capability to process *region-specific* information in complex scenes. This limitation becomes particularly apparent when attempting to describe specific objects within an image using only language prompts, which can be difficult when there is ambiguity (e.g., when there are multiple people in the image, and the question relates to a specific person), as shown in Figure 1.

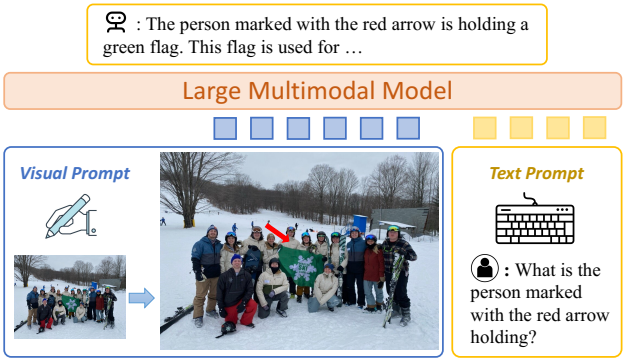

Figure 1. **Main Idea of ViP-LLaVA.** We directly overlay diverse visual prompts (e.g., arrows, boxes, circles, scribbles) onto the original image, and then feed the corresponding visual features along with text embeddings into the large multimodal model for conversational assistance. Here we show an example using a red arrow.

To address this issue, recent work explores spatial references in multimodal models. Existing efforts have primarily focused on using textual representations of coordinates [3, 4, 7, 39], learned positional embeddings [23, 38, 41], or ROI features [26, 38]. However, they often lack user-friendliness, as they are limited to fixed-format visual references like bounding boxes and the spatial coordinates of a mask contour. Most of these approaches, including those by Zhang et al. [38] and Chen et al. [4], only employ bounding box inputs for visual referrals. While effective in structured scenarios, this method proves less versatile in natural, user-driven interactions where the visual prompts may not conform to clean geometric shapes.

In this paper, we propose a simple yet highly effective solution to this problem: a large multimodal model that can

process *arbitrary visual prompts*. This allows a user to intuitively mark up images and interact using natural cues such as a "red bounding box" or "pointed arrow". Our model recognizes these visual prompts, offering a user-friendly way to integrate visual references into the language dialogue. Based on our own observation and prior work [27], which shows that CLIP can understand visual markers, we *directly inject the visual prompts into the original image space* without any additional region-specific model designs. Although our approach is deceptively simple, it yields an unexpected benefit: our model sets new state-of-the-art performances on tasks demanding precise region-specific perception and complex reasoning. It surpasses the capabilities of existing related models with specialized region encoding techniques, as evidenced by our superior performance on region reasoning tasks on Visual7W [43] and PointQA [19].

To further support research in this area, we introduce *ViP-Bench*, a benchmark for evaluating multimodal models' region understanding capabilities with arbitrary visual prompts. By collecting a diverse set of 303 images and questions, we provide a comprehensive assessment of visual understanding capabilities across six aspects at the region level: recognition, OCR, knowledge, math, object relationship reasoning, and language generation. We believe that ViP-Bench will provide a solid foundation for future research into multimodal models with arbitrary visual prompts.

In summary, our main contributions are:

- We introduce a novel multimodal model for intuitive interaction with images using natural language and arbitrary visual prompts, enhancing user accessibility and model flexibility.
- We develop a visual referal approach that overlays visual prompts directly onto images, simplifying the model's architecture without compromising performance.
- Our model, ViP-LLaVA, achieves state-of-the-art results on region understanding tasks on established benchmarks, surpassing specialized region encoding models.
- We introduce ViP-Bench, a benchmark for evaluating visual prompt interpretation, setting a foundational platform for future research.

2. Related Work

Advancements in Large Multimodal Models. Large language models like ChatGPT [21], GPT4 [22], and LLaMA [29] have shown impressive reasoning and generalization capabilities. The landscape of LLMs has been markedly transformed by the recent introduction of models that integrate visual information, such as GPT-4V(ision) [20]. Building upon open-source LLMs [29, 31], a vast number of multimodal vision-language models have made significant strides, spearheaded by LLaVA [14, 15] and MiniGPT-4 [42], which combine LLaMA's [29] language prowess with a CLIP [25] based image encoder. While these models excel at whole-image understanding, a key challenge has been region-specific comprehension within complex visual scenes. This has led to the exploration of spatial referrals in multimodal contexts. Existing models utilize textual coordinate representations [3, 4, 7, 39], learned positional embeddings [23, 38, 41], or Region of Interest (ROI) features [38] to anchor language to specific image regions. However, they often employ rigid visual referral formats that are not as intuitive for users.

Visual Prompting as a User-Friendly Solution. Our focus is on making the interaction with multimodal models more natural and intuitive. Traditional models have employed regular shapes for visual prompting, but our research is motivated by the need for a system that can interpret a wider range of visual prompts. For example, in visual perception, interactive segmentation methods have been proposed that can take in points or scribbles [11, 44]. Drawing inspiration from recent findings that show GPT-4V's ability to understand a variety of markers [32], we advocate for a model that can handle arbitrary visual cues, such as scribbles and arrows. In our model, ViP-LLaVA, we overlay these visual prompts directly onto the image canvas. This is accomplished by fine-tuning on a dataset specifically designed for arbitrary visual prompt instructions.

Evaluating LMM's Region Understanding Capabilities. Existing works [4, 23, 33, 38] evaluates the model's region understanding capabilities on regional multichoice [19, 37, 43] or captioning [12, 35] tasks with metrics such as accuracy, recall, and CIDer [30]. However, these metrics fall short when it comes to evaluating visual dialogue for large multimodal models in an open-world setting. To evaluate LMM's capability in engaging in visual conversations for *image-level* understanding, two families of evaluation are proposed: multiple-choice [16] or using GPT4 as a judge for free-form answers [15, 36]. However, a gap still exists in the evaluation of LMM's capabilities for comprehending arbitrary visual prompts. To address this, we introduce ViP-Bench, a comprehensive benchmark tailored to evaluate how well the LMMs can interpret various visual prompts across multiple dimensions, including recognition, OCR, knowledge, math, relationship reasoning, and language generation.

3. Approach

Our research hinges on the premise that a large multimodal model should not only perceive the visual content of an image but also interpret arbitrary visual markers as part of the user interaction. In this section, we describe our approach that achieves this goal, highlighting the pivotal role of CLIP in understanding visual markers and the construction of a

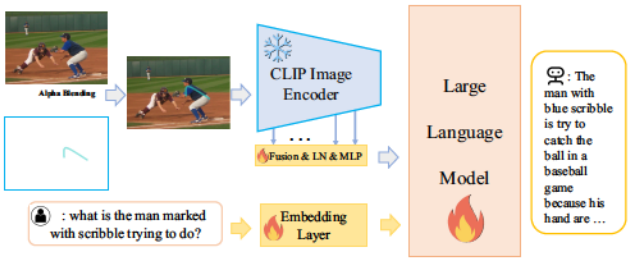

Figure 2. **Model Architecture.** After alpha blending the visual prompts onto the original image, we feed the resulting image into the visual encoder to obtain multi-level visual features. Those features are concatenated and fed into the LayerNorm and MLP layers to form the visual tokens. Then visual tokens and text instruction tokens are fed into the large language model to produce the language response in an auto-regressive manner. The frozen and trainable modules during instruction tuning are annotated.

new instruction tuning dataset tailored to train ViP-LLaVA to understand arbitrary visual prompts.

3.1. Visual Prompt Embedding via CLIP

In contrast to prior work on region understanding [23, 38] which constructs a new module to process visual prompts, we leverage CLIP's [25] existing capabilities to encode both the image and superimposed visual markers. Specifically, CLIP's proficiency in aligning visual and textual data makes it an ideal candidate for this task, as recent studies [27] suggest that it inherently pays attention to marked regions including circles, rectangles, *etc*. As shown in our experiments, we further demonstrate that CLIP can focus the model's attention on a wider variety of visual prompts such as arrows and arbitrary scribbles. To utilize this functionality, we composite the visual prompts $\mathbf{P}_\text{v}$ onto the original image $\mathbf{X}_\text{v}$ through alpha blending, creating a merged representation that highlights the areas of interest:

$$\hat{\mathbf{X}}_\text{v} = \alpha \cdot \mathbf{P}_\text{v} + (1-\alpha) \cdot \mathbf{X}_\text{v}, \tag{1}$$

where $\alpha \in [0, 1]$ denotes the transparency level of the visual prompt, $\mathbf{X}_\text{v}$ is the image, and $\mathbf{P}_\text{v}$ is the image with the visual prompt. Note that we only perform alpha blending for pixels underlying the visual prompt. The composite image $\hat{\mathbf{X}}_\text{v}$ is then fed into the multimodal model.

To effectively recognize the visual prompts, we balance low-level and high-level visual features in ViP-LLaVA.

To address the tendency of CLIP's deeper features to overlook low-level details [40], we selectively extract features from multiple CLIP layers. Specifically, we use one early layer (6-th) to encode detailed geometric shapes and four deeper layers (15, 18, 21, 24-th) to capture broader semantic information. These multi-level features are then concatenated, normalized using LayerNorm [1] for training stability, and finally passed through an MLP layer. This process ensures ViP-LLaVA effectively integrates diverse visual cues, a strategy validated through our ablation studies detailed in Sec. 5.4.

Our design's simplicity of directly overlaying visual prompts offers several advantages. It reduces model complexity by bypassing additional processing modules and aligns closely with natural human interactions, as users often employ diverse and spontaneous visual markers. This flexibility allows ViP-LLaVA to interpret a wide range of user-generated visual cues, enhancing its applicability in real-world scenarios.

To train ViP-LLaVA, we perform autoregressive language modeling; *i.e.*, we maximize the likelihood of generating the tokens of the ground-truth answer $\mathbf{X}_\text{a}$:

$$P(\mathbf{X}_\text{a} \mid \hat{\mathbf{X}}_\text{v}, \mathbf{X}_\text{instruct}) = \prod_{i=1}^{L} P_{\boldsymbol{\theta}}(x_i \mid \hat{\mathbf{X}}_\text{v}, \mathbf{X}_\text{instruct}, \mathbf{X}_{\text{a},<i}) \tag{2}$$

where $\boldsymbol{\theta}$ represents the trainable parameters, $\mathbf{X}_\text{instruct}$ is the text instruction, L is the sequence length of the answer $\mathbf{X}_\text{a}$, and $\mathbf{X}_{\text{a},<i}$ denotes all the answer tokens before the current prediction token x_i, where i denotes the steps during text token generation. Here we omit system messages from the equation for clarity, even though they are part of the conditioning. Figure 2 shows our model architecture.

This training objective enables the model to generate contextually accurate responses by comprehending the visual content, language instruction, and the overlaid prompts. It fosters the model's ability to interpret visual markers in unison with the image, thereby enhancing its proficiency in addressing complex, region-specific language inquiries. This capability is crucial for tasks requiring nuanced understanding of both the visual elements and user intentions conveyed through arbitrary visual prompts.

3.2. Visual Prompting Design

To train the model to recognize and interpret arbitrary visual prompts, we develop a new visual prompt instruction tuning dataset, as there are no prior datasets with arbitrary visual prompts and instruction-output text pairs that we can use.

Our dataset comprises a diverse collection of 520k image-text pairs marked with visual prompts, sourced from publicly available datasets, including (1) single region reasoning data: 80k referring comprehension and generation data from RefCOCOg [35], and 37k object counting data from PointQA-LookTwice [19], (2) two-region reasoning data: 80k triplet relationship data from Visual Genome [12], (3) multi-region reasoning data: 30k grounded image captioning data from Flicker 30k Entities [24], 213K data from Visual Commonsense Reasoning dataset [37], and 82k data from Visual7W [43]. Note that all those data are collected from the training split of the aforementioned datasets.

Figure 3. **Visualization of Visual Prompt Types.** From top-left to bottom-right: mask contour, ellipse, bounding box, triangle, scribble, point, arrow, and mask. Note that the prompts not only have diverse shapes, but they also have diverse colors, transparency values, widths, scales, and directions.

We automatically annotate each image with various visual prompts. For the data that only comes with bounding box annotations, we sample the visual prompts from three possible categories: rectangle, ellipse, and arrow. Here we make sure that the head of the arrow lies within $[(-\frac{W}{2}, -\frac{H}{2}), (\frac{W}{2}, \frac{H}{2})]$ space, where W, H are the width and height of the image, respectively. For ellipse, the lengths along the semi-major and semi-minor axes are inherited from the bounding box size, where we enlarge the ellipse with a ratio between $[1, 1.5]$. On the other hand, for regions that come with ground truth pixel-level mask annotations, we annotate each region with visual prompts sampled from the following 8 possibilities: rectangle, ellipse, point, triangle, mask, mask contour, arrow, and scribble created using Bézier curves; see Figure 3. We make sure that the head of the arrow, entire point, triangle, and scribble lies within the provided mask. These annotations simulate natural human interactions with images, where users often use spontaneous markers to highlight areas of interest.

For scribbles, we simulate human-like drawings using Bézier curves [6]. This process begins by randomly selecting three points within the object mask, which serve as the anchors for the quadratic Bézier curve. The generated Bézier curve is then composited onto the image using the previously mentioned alpha blending technique to produce a merged image with the scribble serving as a visual prompt.

Humans naturally use various markers to highlight objects within their environment. For instance, in educational settings, teachers often use arrows or underlining to draw students' attention to specific parts of an image or text. Similarly, in everyday communication, people might circle items in a photograph to point out something of interest or use scribbles to obscure sensitive information before sharing. Through our design, we create a visual instruction following dataset that mirrors the way humans visually interact with objects, thus fostering a more intuitive and natural interaction with the model.

3.3. Optional Region-level Instruction Tuning Data

Our training data comes from two sources: (i) region-level visual prompting data described in Section 3.2, and (ii) image-level data devoid of visual prompts, sourced from LLaVA-1.5 [14]. This strategy enables ViP-LLaVA to engage in human-like conversations, primarily due to the image-level LLaVA instruction data from Liu *et al*. [15]. Optionally, to further enhance ViP-LLaVA's capability in multimodal conversations at the region-level, we design region-specific instruction data with the help of GPT-4V.

Prior approaches like Shikra [4] attempted to generate region-level instruction data using text-only models like GPT4. However, this method is inherently limiting, particularly in object-level tasks where the model, lacking visual context, cannot accurately reference multiple objects of the same class within a single scene. To overcome this, we develop an instruction data curation method using GPT-4V. Unlike text-only models, GPT-4V can interpret visual prompts displayed in images [32]. Our method involves feeding two images into GPT-4V: the original image and a modified version with annotated visual prompts. Alongside these images, we provide the model with the ground-truth (text) annotation in the original dataset and system messages. This process is used to curate `<visual prompt, text prompt, text output>` triplets for the images in our dataset described in Section 3.2.

We introduce specific textual representations such as `<within red mask>` and `(<within red box>, <within blue box>)` to guide GPT-4V in recognizing the visual prompts in both single-region and multi-region settings. During training, we replace these phrases with the set of eight possible visual prompts described in Section 3.2, significantly enhancing the dataset's versatility. In total, we curate 13k high-quality region-level instruction data points, comprised of 7k single-region and 6k multi-region instances. In the supplementary, we provide specific details of the system messages, input text prompts, and generated text outputs.

Although ViP-LLaVA works well even without this enriched data for standard visual reasoning benchmarks, we find that it helps to further improve the model's ability to have human-like conversations in open-world settings.

4. ViP-Bench for Evaluation

In order to rigorously evaluate the capabilities of multimodal models in interpreting and responding to visual reasoning queries, we introduce ViP-Bench, a benchmarking suite for evaluating multimodal region-understanding capabilities under various visual prompts. ViP-Bench consists of 303 unique image-question pairs, where images are collected from MM-Vet [36], MMBench [16], and Visual Genome [12]. Each pair consists of an image coupled with a diverse visual reasoning question designed to test a model's understanding and interpretation capabilities. We reuse the questions in MM-Vet [36] and MMBench [16] (but make minor adjustments so that they take into account the region-

Method	Generalist?	Accuracy (%)
LSTM-Att [43]	×	56.10
CMNs [10]	×	72.53
12in1 [18]	×	83.35
GPT4ROI-7B [38]	×	81.83
GPT4ROI-13B [38]	×	84.82
Shikra-13B [4]	✓	85.33
ViP-LLaVA-Base-7B	✓	86.04
ViP-LLaVA-Base-13B	✓	87.54
ViP-LLaVA-7B	✓	86.60
ViP-LLaVA-13B	✓	**87.91**

Table 1. Comparison of methods in terms of generality and accuracy on Visual7W [43] test set.

specific visual prompts), while in Visual Genome, we design the questions and answers by ourselves. We use bounding boxes and masks produced by the Segment Anything Model (SAM) [11] to annotate the location of the objects.

Key to the design of ViP-Bench is its comprehensive coverage of six crucial aspects of visual understanding at the region level: recognition, OCR (Optical Character Recognition), knowledge, math, object relationship reasoning, and language generation. This range ensures a holistic assessment of a model's performance in various facets of region-level visual reasoning.

ViP-Bench employs a similar grading mechanism as MM-Vet [36]. We employ the GPT-4 text model, a state-of-the-art language model, to evaluate the responses of multimodal models. Specifically, we feed the response from the multimodal model, the human annotated answer, and several in-context scoring examples to GPT-4. The responses are scored by GPT-4 on a scale from 0 to 10, offering a quantitative measure of the multimodal model's proficiency in understanding and interpreting visual data. This grading system provides a standardized framework for comparing the performance of different models.

ViP-Bench is meticulously annotated by humans. This process involved seven rounds of validation to ensure the accuracy and relevance of the object boxes/masks, questions, and answers. Such rigorous annotation guarantees the reliability of the benchmark as a tool for model evaluation. An illustrative example in Table 6 showcases a scenario where a leading model like GPT-4V misinterprets object localization under ViP-Bench, highlighting the challenges in current multimodal understanding. We present additional visualizations and statistics of ViP-Bench in the supp.

Through ViP-Bench, we provide a valuable tool for the research community, aiding in the development and refinement of multimodal models. By offering a comprehensive and challenging testbed, we believe ViP-Bench can set the stage for future advancements in the field of visual reasoning and multimodal interaction.

Method	Generalist?	Accuracy (%)
Point and ask [19]	×	60.20
LLaVA-1.5-7B [14]	✓	56.19†
LLaVA-1.5-13B [14]	✓	57.93†
Shikra-13B [4]	✓	70.30
ViP-LLaVA-Base-7B	✓	70.86
ViP-LLaVA-Base-13B	✓	**72.15**
ViP-LLaVA-7B	✓	71.31
ViP-LLaVA-13B	✓	71.77

Table 2. Comparison of methods in terms of generality and accuracy on PointQA-LookTwice [19] test set. †zero-shot eval.

5. Experiments

In this section, we compare ViP-LLaVA to state-of-the-art multimodal models, including those that explicitly design region-specific modules, perform in-depth analysis to assess ViP-LLaVA's capabilities, and perform ablation studies.

5.1. Training Setup

Model. For the visual model, we choose CLIP-336px [25] to preserve more information from the raw pixel space. We use Vicuna v1.5 [31] as the language encoder. For the multimodal connector, a 2-layer MLP is utilized.

Training and data. During the initial stage of training, we employ 558k BLIP [5, 15] captioned image-text pairs to pretrain the multimodal connector. The second stage utilizes LLaVA v1.5 [14] instruction data alongside our region-level visual prompting dataset from Section 3.2. Both stages train the model for 1 epoch, with an overall training time of around 20/40 hours for the 7B/13B model using 8 NVIDIA A100 GPUs. Finally, we mix the 13k GPT-4V instruction data with 13k sampled data from stage 2 to get 26k stage 3 training data, and then fine-tune our stage-2 model (referred to as ViP-LLaVA-Base) for one epoch to get our model ViP-LLaVA, which requires approximately 0.5 hours for the 7B model and 1 hour for the 13B model on 8 NVIDIA A100 GPUs.

Visual prompts. ViP-LLaVA uses 8 visual prompts: rectangles, ellipses, points, scribbles, triangles, masks, mask contours, and arrows. Their attributes, such as color, thickness, and alpha value for alpha blending (in [0.5, 1]) are randomized. The arrow's direction and length are randomized, with the endpoint remaining within the mask. For referencing specific regions, we replace the `<region>` text with the color and shape description, such as `red scribble`. The visual prompt type and associated attributes for each region are randomly assigned during training.

5.2. Evaluation on Region Reasoning Benchmarks

We first quantitatively evaluate ViP-LLaVA on three region reasoning benchmarks.

Model	$Q \rightarrow A$ (%)	$QA \rightarrow R$ (%)	$Q \rightarrow A$
ViLBERT [17]	72.4	74.5	54.
Unicoder-VL [13]	72.6	74.5	54.
VLBERT-L [28]	75.5	77.9	58.
ERNIE-ViL-L [34]	78.52	83.37	65.
VILLA-L [8]	78.45	82.57	65.
GPT4RoI-7B [38]	87.4	89.6	78.
ViP-LLaVA-Base-7B	**87.66**	**89.80**	**78.**

Table 3. Validation Accuracy on VCR [37] dataset.

Visual7W. The Visual7W dataset [43] tests models' s tial perception by requiring them to match text desc tions with the correct bounding boxes from a set of choi We differentiate between 'generalist' models, which are specifically trained on the target dataset, and 'specialis models, which are. For a fair comparison, we use image overlays as visual prompts for the LLaVA model and textual coordinates for Shikra's text prompts. The results in Table 1 shows ViP-LLaVA-7B outperforming recent state-of-the-art methods, including GPT4RoI [38] and Shikra [4], despite having fewer parameters, and ViP-LLaVA-13B producing even higher gains. ViP-LLaVA overlays bounding boxes directly onto the image, creating an immediate link between the image and spatial locations. This contrasts with other methods that rely on external embeddings from either textual or newly learned embedding spaces to reference specific regions, proving less effective in this context.

PointQA-LookTwice. PointQA [19] presents a dataset where queries are based on either a specific point or a bounding box within an image. We evaluate ViP-LLaVA under the broad-question scenario using the bounding box type, typified by the prompt `How many of these are there?` This requires the model to first correctly identify the object within the given region and subsequently enumerate instances of the same category across the image—essentially a test of object recognition followed by class-specific counting. In line with our methodology for Visual7W, we use the image overlaid with the bounding box for LLaVA, while for Shikra, we incorporate the bounding box coordinates into the text prompt. Table 2 shows ViP-LLaVA's superior performance on this intricate task, surpassing other multimodal contenders. Our method of overlaying visual prompts ensures the object remains unobscured, effectively combining the original image pixels with visual cues to enhance object recognition and counting accuracy.

Visual Commonsense Reasoning. The Visual Commonsense Reasoning (VCR) dataset [37] is a challenging benchmark designed to evaluate a model's capabilities in high-level cognition and commonsense reasoning in the context of visual information. The dataset presents multiple-choice questions that require an understanding of the scene depicted in an image. Each question (Q) is paired with four potential answers (A), where the model must not only select the correct answer but also provide a rationale (R) that justifies its choice, demonstrating the model's ability to comprehend and rationalize visual elements within a given context.

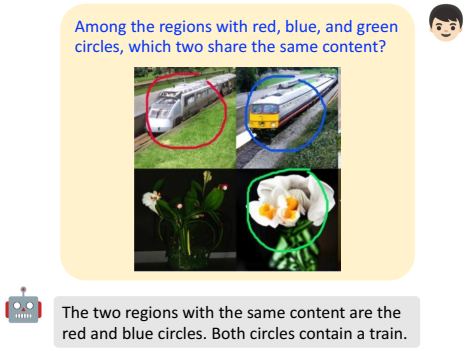

Figure 4. ViP-LLaVA model is able to infer correspondences between multiple objects in the image.

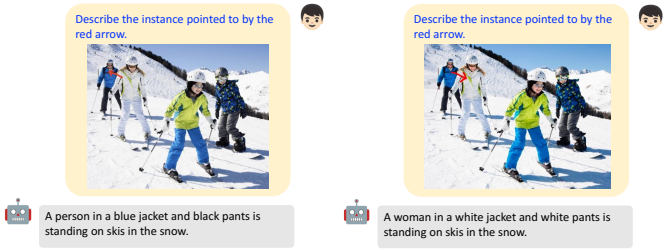

Figure 5. ViP-LLaVA is able to understand the direction of arrows.

We finetune ViP-LLaVA-Base-7B on VCR, similar to the protocol in GPT4RoI [38]. As shown in Table 3, our approach exhibits state-of-the-art performance on the validation set, illustrating its proficiency in visual commonsense reasoning. This success highlights our approach's dual strengths: adeptness in perception tasks and effectiveness in multi-region reasoning. By integrating visual prompts directly into the image, our model more effectively associates spatial locations with semantic understanding, facilitating a better interaction between spatial and semantic reasoning.

5.3. In-depth Analysis

Region reasoning with arbitrary human drawings at test time. ViP-LLaVA, when presented with arbitrarily drawn enclosed regions or arrows by a user, can accurately describe, shown in Figure 4 and 6.

Multi-region understanding capabilities. ViP-LLaVA demonstrates robust multi-region understanding, able to dissect complex visual scenes and infer relationships between various elements. As shown in Figure 4, ViP-LLaVA is able to infer correspondences between multiple objects in the image, and make the correct reasoning that the red and

Model	Visual7W	PointQA	VCR ($Q \rightarrow A$) (%)
Single Layer	82.44	70.23	84.64
Multi-Layer	**86.04**	**70.86**	**87.66**

Table 4. Single layer vs. multi-layer CLIP image encodings.

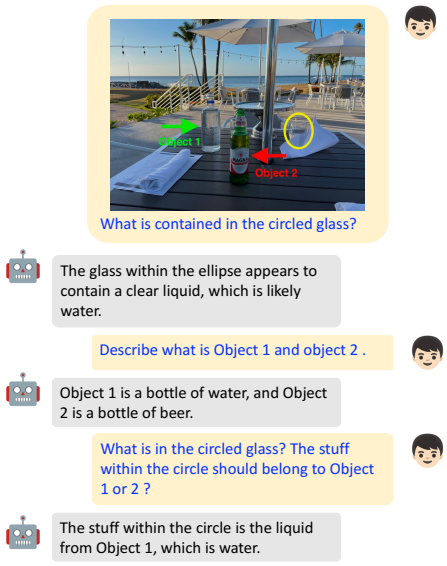

Figure 6. ViP-LLaVA is able to understand the text markers in the image, and reason about them.

blue circles both include the train.

Arrow direction understanding. ViP-LLaVA is able to understand arrows. Here we conduct an ablation study of the arrow direction. Given two arrows that have the same body yet different heads, as shown in Figure 5, ViP-LLaVA is able to understand the direction of the arrows, making correction descriptions about the respective regions.

Generalization to other attributes. ViP-LLaVA also generalizes to untrained attributes, like varying visual prompt thickness or location, showcasing its adaptability beyond what was seen during training. See the supplementary material for examples of different thicknesses.

Figure 6 further shows that ViP-LLaVA is able to conduct OCR first, and then make correspondences between different regions to make a correct prediction about the content of each part.

5.4. Ablation Studies

Impact of overlaying visual prompts on visual information. To assess whether overlaying visual prompts on images obscures visual information, we conduct a comparison by inputting visual tokens from both the original and overlayed images into ViP-LLaVA-Base-7B. Using the VCR dataset, we evaluate the accuracy of the QA task with and without the additional visual tokens from the original image. Results on the VCR validation split shows an accuracy of 81.63% with the original image and overlaid image tokens, compared to 82.47% with the overlaid image tokens only. The similar accuracies suggest that the overlaid prompts do not detract from the visual information processed by our model.

Influence of CLIP multi-layer features. We next explore the impact of using multi-layer visual features from CLIP as opposed to single-layer features, specifically focusing on the second-last layer as implemented in LLaVA [14, 15]. Our ablation study in Table 4 reveals a marked improvement in performance, particularly in scenarios involving multiple visual prompts, as in the Visual7W and VCR datasets. This indicates that leveraging multi-layer visual features significantly enhances the model's ability to localize and recognize visual prompts within images.

6. ViP-Bench Evaluation Results

Finally, we evaluate on ViP-Bench using a set of image-level and region-level LMMs, including InstructBLIP [5], GPT-4V [20], LLaVA v1.5 [15], Qwen-VL [2], Shikra [4], GPT4ROI [38] and Kosmos-2 [23]. For open-source models, we evaluate with greedy decoding (temperature=0). As shown in Table 5, we first see that the performance of all models, including GPT-4V, is far from perfect, demonstrating the difficulty of ViP-Bench. An illustrative case in Table 6 depicts a scenario where GPT-4V and LLaVA incorrectly predict object localization. Overall, ViP-LLaVA outperforms other models, except GPT-4V, demonstrating greater adaptability to various visual perception and reasoning tasks. By training on images overlaid with visual prompts, ViP-LLaVA becomes adept at understanding arbitrary visual cues and mimicks the natural human method of referring to objects in images. This enables it not only to better identify and interpret visual prompts but also to integrate these prompts into its reasoning process, enhancing its overall comprehension and response accuracy.

Visual prompting is superior to other representations. In zero-shot evaluation, when visual prompts are represented as a simple list of four textual numerical values, models like Qwen-VL and LLaVA underperform compared to ViP-LLaVA. This underscores the effectiveness of visual prompts over basic textual representations.

Language tasks: A challenge for current LMMs. The ViP-Bench results reveal that, compared to GPT-4V, open-source LMMs show a significant gap in OCR, math, and language generation tasks, while they perform decently in recognition, knowledge, and object relationship reasoning. This suggests that future VLM developments should prioritize enhancing language reasoning capabilities. For OCR, the results indicate a need for higher resolution inputs or a more robust backbone model, moving beyond the existing capabilities of models like CLIP.

Model	Format	Synthesized visual prompts (tight bounding box)							Visual prompts from human (*e.g.* arrow, circle)						
		Rec	OCR	Know	Math	Rel	Lang	All	Rec	OCR	Know	Math	Rel	Lang	All
GPT-4V-turbo-detail:high [22]	VP	58.1	69.8	59.5	71.0	61.4	51.9	60.7	56.9	69.7	63.7	80.6	61.1	45.6	59.9
GPT-4V-turbo-detail:low [22]	VP	53.2	50.3	55.6	67.7	57.5	57.5	52.8	51.7	50.3	59.3	60.3	55.0	43.8	51.4
InstructBLIP-7B [5]	VP	36.9	16.3	34.2	22.3	26.8	7.5	31.7	38.9	17	35.4	9.7	29.3	17.5	33.3
Shikra 7B [4]	Coor	40.2	10.0	28.0	3.5	18.9	20.6	33.7	–	–	–	–	–	–	–
GPT4ROI 7B [38]	ROI	35.6	16.7	29.7	9.7	32.5	13.8	35.1	–	–	–	–	–	–	–
Kosmos-2 [23]	Dis	29.5	14.2	18.5	9.7	7.5	21.9	26.9	–	–	–	–	–	–	–
LLaVA-1.5-7B [15]	Coor	52.7	20.7	44.7	14.5	44.6	30.6	44.8	–	–	–	–	–	–	–
LLaVA-1.5-7B [15]	VP	50.8	12.4	49.2	6.5	51.8	23.8	41.6	49.1	13	42.9	9.7	50	27.5	40.2
Qwen-VL-Chat [2]	Coor	52.6	22.0	40.0	12.9	47.1	26.9	45.3	–	–	–	–	–	–	–
Qwen-VL-Chat [2]	VP	43.0	30.4	40.2	9.7	25.7	28.7	39.2	48.7	22.1	41.2	6.5	48.2	25	41.7
ViP-LLaVA-Base-7B	VP	54.8	18.8	52.9	9.7	53.9	42.5	45.5	55.3	17.6	45.9	8.1	44.6	33.1	46.8
ViP-LLaVA-7B	VP	56.7	19.4	49.7	10.0	50.4	33.8	**48.4**	56.7	21.2	47.1	12.3	50.4	36.2	**48.3**
InstructBLIP-13B [5]	VP	42.5	12.2	37.5	3.2	33.2	12.5	35.8	41.7	13.6	35.9	3.2	27.9	18.8	35.2
LLaVA-1.5-13B [15]	Coor	53.2	26.1	45.9	9.7	52.5	31.9	47.1	–	–	–	–	–	–	–
LLaVA-1.5-13B [15]	VP	48.1	21.8	40.0	6.1	45	28.1	41.8	48.8	21.3	47.3	15.8	44.6	31.2	42.9
ViP-LLaVA-Base-13B	VP	54.4	27.8	51.2	16.1	51.1	46.9	48.2	55.3	21.1	46.8	9.7	45.4	38.8	47.0
ViP-LLaVA-13B	VP	56.3	24.6	53.4	15.5	50.0	53.8	**48.3**	55.4	26.9	49.3	15.5	48.6	41.9	**48.2**

Table 5. **ViP-Bench Evaluation Results.** This table presents the performance of various models under ViP-Bench, utilizing different visual prompt formats. The evaluation includes both synthesized and human-drawn prompts, providing insights into the models' maximum potential and real-world applicability, respectively. Formats include VP (visual prompts), Coor (coordinates as visual prompts), Dis (discrete positional tokens for vocabulary expansion), and ROI (CLIP region of interest features with positional embedding). The assessed dimensions are Recognition (Rec), OCR, Knowledge (Know), Math, Relationship (Rel), and Language Generation (Lang).

Visual input example for Spatial Reasoning:	
	
User	Between Object 1: the object within the red mask contour, Object 2: the object within the blue mask contour, and Object 3: the object within the green mask contour, which one has something on top of it?
Ground Truth	Object 2
GPT-4V	Object 3, the object within the green mask contour, has something on top of it. It appears to be a coffee pot placed on top of what could be a coffee machine.
LLaVA-1.5-13B	Object 3, which is the coffee maker, has something on top of it.
ViP-LLaVA-13B	Object 2: the object within the blue mask contour has something on top of it.

Table 6. An example in ViP-Bench where GPT-4V makes a wrong prediction. The correct answer should be Object 2.

Overfitting Concerns in Region-Level LMMs. Current region-level LMMs, including Shikra [4], GPT4ROI [38] and Kosmos-2 [23], tend to struggle with tasks involving mathematics, relationship reasoning, and language generation. This trend suggests a potential overfitting issue with these models to existing public region-level datasets, which predominantly feature brief descriptions.

7. Conclusion

In summary, ViP-LLaVA shows that visual prompts are promising for region-specific image understanding. By integrating arbitrary visual prompts, we bridge the gap between user-friendly interfaces and the precision required for region comprehension. ViP-LLaVA's intuitive design leverages natural linguistic interactions coupled with visual markers, simplifying the process of image annotation while enhancing the clarity of visual references. Our state-of-the-art performance on established benchmarks including Visual7W, PointQA, and VCR, underlines the efficacy of ViP-LLaVA. Notably, the introduction of ViP-Bench as a comprehensive evaluative platform sets a new standard for assessing multimodal models' region reasoning abilities. ViP-LLaVA establishes a foundation for further exploration in the field of intelligent visual systems. We believe that ViP-LLaVA can motivate how visual and linguistic modalities are integrated, enabling more sophisticated and nuanced human-machine interactions.

Acknowledgements. This work was supported in part by NSF CAREER IIS2150012, and Institute of Information & communications Technology Planning & Evaluation(IITP) grants funded by the Korea government(MSIT) (No. 2022-0-00871, Development of AI Autonomy and Knowledge Enhancement for AI Agent Collaboration). (No. RS2022-00187238, Development of Large Korean Language Model Technology for Efficient Pre-training), and Microsoft Accelerate Foundation Models Research Program.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 3

[2] 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. *arXiv preprint arXiv:2308.12966*, 2023. 7, 8

[3] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechu Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 1, 2

[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. 1, 2, 4, 5, 6, 7, 8

[5] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 5, 7, 8

[6] Gerald Farin. *Curves and Surfaces for Computer-Aided Geometric Design: A Practical Guide*. Academic Press, 2014. 4

[7] Jon Ferraiolo, Fujisawa Jun, and Dean Jackson. *Scalable vector graphics (SVG) 1.0 specification*. iuniverse Bloomington, 2000. 1, 2

[8] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628, 2020. 6

[9] Google. Google bard. https://bard.google.com/chat/, 2023. 1

[10] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1115–1124, 2017. 5

[11] 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. *arXiv preprint arXiv:2304.02643*, 2023. 2, 5

[12] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 2, 3, 4

[13] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11336–11344, 2020. 6

[14] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 1, 2, 4, 5, 7

[15] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv:2304.08485*, 2023. 1, 2, 4, 5, 7, 8

[16] 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. 2, 4

[17] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 6

[18] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10437–10446, 2020. 5

[19] Arjun Mani, Nobline Yoo, Will Hinthorn, and Olga Russakovsky. Point and ask: Incorporating pointing into visual question answering. *arXiv preprint arXiv:2011.13681*, 2020. 2, 3, 5, 6

[20] OpenAI. Gpt-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf, 2023. 1, 2, 7

[21] OpenAI. Chatgpt. https://openai.com/blog/chatgpt/, 2023. 1, 2

[22] OpenAI. Gpt-4 technical report. 2023. 1, 2, 8

[23] 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. 1, 2, 3, 7, 8

[24] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 3

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 5

[26] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. *arXiv preprint arXiv:2311.03356*, 2023. 1

[27] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. *ICCV*, 2023. 2, 3

[28] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-

linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 6

[29] 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. 2

[30] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 2

[31] Vicuna. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://vicuna.lmsys.org/, 2023. 2, 5

[32] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 2023. 2, 4

[33] 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

[34] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3208–3216, 2021. 6

[35] 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, 3

[36] 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. 1, 2, 4, 5

[37] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3, 6

[38] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023. 1, 2, 3, 5, 6, 7, 8

[39] Liang Zhao, En Yu, Zheng Ge, Jinrong Yang, Haoran Wei, Hongyu Zhou, Jianjian Sun, Yuang Peng, Runpei Dong, Chunrui Han, et al. Chatspot: Bootstrapping multimodal llms via precise referring instruction tuning. *arXiv preprint arXiv:2307.09474*, 2023. 1, 2

[40] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[41] Qiang Zhou, Chaohui Yu, Shaofeng Zhang, Sitong Wu, Zhibing Wang, and Fan Wang. Regionblip: A unified multi-modal pre-training framework for holistic and regional comprehension, 2023. 1, 2

[42] 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, 2

[43] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7W: Grounded Question Answering in Images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2, 3, 5, 6

[44] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *NeurIPS*, 2023. 2