VP-Bench: A Comprehensive Benchmark for Visual Prompting in Multimodal Large Language Models

Mingjie Xu[1*], Jinpeng Chen[2*], Yuzhi Zhao[2†], Jason Chun Lok Li[3], Yue Qiu[4], Zekang Du[4], Mengyang Wu[5], Pingping Zhang[2], Kun Li[2], Hongzheng Yang[5], Wenao Ma[5], Jiaheng Wei[1], Qinbin Li[4], Kangcheng Liu[6], Wenqiang Lei[7]

[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]City University of Hong Kong
[3]The University of Hong Kong
[4]Huazhong University of Science and Technology
[5]The Chinese University of Hong Kong
[6]Hunan University
[7]Sichuan University
mingjiexu@hkust-gz.edu.cn, jinpeng.chen@my.cityu.edu.hk, yzzhao2-c@my.cityu.edu.hk

Abstract

Multimodal Large Language Models (MLLM) have enabled a wide range of advanced vision-language applications, including fine-grained object recognition and contextual understanding. When querying specific regions or objects in an image, human users naturally use "Visual Prompts" (VP) like bounding boxes to provide reference. However, no existing benchmark systematically evaluates the ability of MLLMs to interpret such VPs. This gap raises uncertainty about whether current MLLMs can effectively recognize VPs, an intuitive prompting method for humans, and utilize them to solve problems. To address this limitation, we introduce VP-Bench, aiming to assess MLLMs' capability in VP perception and utilization. VP-Bench employs a two-stage evaluation framework: Stage 1 examines models' ability to perceive VPs in natural scenes, utilizing visualized prompts spanning 8 shapes and 355 attribute combinations. Stage 2 investigates the impact of VPs on downstream tasks, measuring their effectiveness in real-world problem-solving scenarios. Using VP-Bench, we evaluate 28 MLLMs, including proprietary systems (e.g., GPT-4o) and open-source models (e.g., InternVL-3 and Qwen2.5-VL). In addition, we provide a comprehensive analysis of factors affecting VP understanding, such as variations in VP attributes, question arrangement, and model scale. VP-Bench establishes a new reference framework for studying MLLMs' ability to comprehend and resolve grounded referring questions.

Datasets — https://github.com/Endlinc/VP-Bench

Introduction

The emergence of multimodal large language models (MLLM) (OpenAI 2023, 2024; Liu et al. 2023) has spurred research into their applications across diverse downstream tasks. For example, a user can verbally instruct the model to recognize furniture in indoor 3D scenes (Zhou et al. 2024b; Zhang et al. 2024c; Zhou et al. 2024a), ground wild animals in their natural environment as depicted in an image (Rasheed et al. 2024; Zhang et al. 2024a), use external knowledge to determine a piece of furniture's brand and price, and provide insights into the habits of those wild animals. Beyond object detection via natural-language queries, researchers have explored MLLMs' ability to interpret user-drawn annotations (e.g., freehand regions) and identify interactions involving target instances in context (Cai et al. 2024; Fu et al. 2024). Consequently, visual prompts (VP), graphical cues such as bounding boxes and alphabet tags, have emerged to direct model attention, offering an intuitive alternative to verbal descriptions. However, MLLMs still underperform human annotators in grounded referring tasks with VPs. To quantify this gap, ViP-Bench (Cai et al. 2024) was introduced, comprising 303 image–question pairs across eight VP types in practical scenarios (e.g., OCR, mathematical reasoning). Although ViP-Bench provides valuable insights into region reasoning, it does not assess, first, how perceptible different VP styles are to models. For example, while bounding boxes are ubiquitous, their low-contrast or overly thin edges may be difficult for models to detect, limiting task accuracy. Moreover, it remains unclear whether adding distinctive corner markers to a bounding box VP would further enhance performance. Second, how variations in VP design affect downstream performance. For example, consider a lung region that is suspected to contain a malignant lesion. Is the application of an additional overlaying VP on this area more effective in directing the model's attention compared to specifying the region's location within the verbal instruction, while still maintaining the contextual information?

To address these questions, we conduct a comprehensive review of prior studies (Cai et al. 2024; Fu et al. 2024; Zhang et al. 2024a; Rasheed et al. 2024; Yang et al. 2023) to catego-

[*]These authors contributed equally.
[†]Corresponding Author.

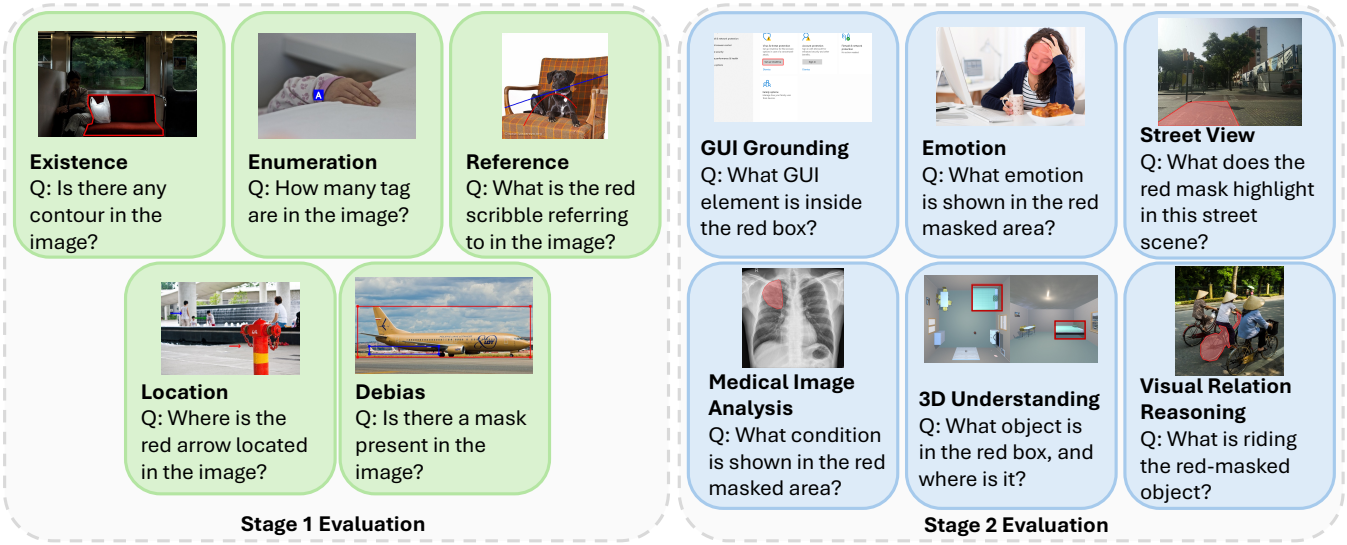

Figure 1: Overview of the VP-Bench Dataset. VP-Bench introduces a two-stage evaluation framework: (1) **Model Perception**, which assesses general VP recognition capabilities using visualized VPs spanning five question types; and (2) **VP Effect on Downstream Tasks**, which evaluates the impact of visual prompts on various downstream applications. All questions follow a multiple-choice format, but the full list of options is not displayed due to space limitations.

Benchmark	# Images	# VP	Domains	Debias
SoV	119	≈ 5	1	No
ViP-Bench	303	8	1	No
VP-Bench	**34 267**	**355**	4	Yes

Table 1: Comparison of three VP-related benchmarks. We list the number of images, number of VP attribute combinations, the image domains covered, and whether each dataset includes debias questions.

rize commonly employed VP shapes and their corresponding variants, which we subsequently define as attributes. We classify VP shapes into eight categories: tag, bounding box, arrow, mask, contour, oval, point, and scribble. Each category shares general attributes such as line width and color, and may also incorporate shape-specific attributes, including style or content format. A detailed overview of the VP shapes and their associated attributes is provided in the supplementary materials. In total, our proposed VP-Bench encompasses 355 unique VP attribute permutations. The benchmark consists of 38,932 distinct image–question pairs, constructed to evaluate core VP capabilities, including existence, enumeration, referencing, and spatial localization. This component constitutes Stage 1 of VP-Bench. Furthermore, to examine the influence of VP on downstream task performance, we introduce six tasks that incorporate VP reasoning as a second stage, thereby reflecting the model's ability to leverage VP reasoning in real-world problem-solving contexts. Additional examples from the benchmark are provided in Figure 1, and a concise comparison between VP-Bench and other VP-related benchmarks is presented in Table 1.

We evaluate a range of popular MLLMs on our VP-Bench, including proprietary systems (e.g., GPT-4o) and open-source models (e.g., InternVL-3 and Qwen2.5-VL). In Stage 1, we analyze the influence of factors such as instruction arrangement, model parameters, VP shapes, and attributes on model performance. In Stage 2, we examine the models' capabilities across various VP-enabled downstream tasks, evaluating their ability to integrate visual cues with textual context, distinguish fine-grained object features, and leverage domain-specific knowledge. Moreover, we complement quantitative metrics with qualitative analyses to highlight each model's interpretability and reliability in real-world scenarios. This comprehensive evaluation not only benchmarks current capabilities but also identifies potential areas for improvement, thereby informing future research in multimodal interaction and model interpretability.

To summarize, our contributions are as follows:

- We introduce VP-Bench, a two-stage evaluation framework for assessing MLLMs. Stage 1 measures VP perception capability in natural scenes, while Stage 2 evaluates the ability to integrate VP understanding for practical problem-solving. Compared to existing benchmarks, our evaluation is significantly more comprehensive and includes over 100 times as many images.
- In stage 1, we examine the effects of 355 attribute combinations across eight VP shapes, covering a scale more than 40 times larger than previous studies. In Stage 2, we assess six VP-enabled downstream tasks, offering a comprehensive reference for real-world applications.
- Our results highlight the critical role of VP shape in model performance. Regular shapes (e.g., bounding boxes, ovals) are generally more efficient than irregular ones (e.g., masks, contours), even though the latter pro-

vide finer spatial details. Including VP shape descriptions in prompts further improves contextual understanding. Selecting VP shapes suited to the application scenario is more impactful than relying on the model's preferred shape in a general scenario.

Related Works

Multimodal Large Language Models

Building on the recent success of neural language processing, particularly through LLM approaches, researchers have increasingly integrated visual understanding and reasoning to expand these models' capabilities (Chen et al. 2022; Huang et al. 2023; OpenAI 2023; Li et al. 2023; Liu et al. 2023). Several studies have introduced visual instruction tuning (Liu et al. 2023) and specialized architectures (Li et al. 2023) for MLLMs, leading to significant advancements in image comprehension and common-sense reasoning. However, many existing MLLMs lack dedicated models or data designs for referring expressions or location-based referencing. Consequently, some researchers have adopted mask encoders to focus attention on specific regions (Guo et al. 2024), while others craft targeted text prompts to highlight region (Wang et al. 2024b). For instance, RegionGPT (Guo et al. 2024) and LLaVA-Grounding (Zhang et al. 2024a) employ additional mask encoders to improve location comprehension, though this adds computational overhead and demands retraining when introducing newer base models or additional VP shapes. Alternatively, approaches such as SoM (Yang et al. 2023), ViP-LLaVA (Cai et al. 2024), and ControlMLLM (Wu et al. 2024) demonstrate that sketching VPs directly on images can substantially enhance various downstream region-referring tasks. Yet, evaluations of these models' handling of visually marked images have largely been qualitative, and comprehensive quantitative assessments across diverse region-referring tasks remain limited.

MLLM Benchmarks

Benchmarking MLLMs is crucial for exposing model limitations and guiding future development (Yue et al. 2024; Yu et al. 2024; Meng et al. 2025; Ying et al. 2024; Liu et al. 2024; Guan et al. 2024). Although many existing benchmarks assess perception and reasoning, they largely emphasize image-level tasks. A few incorporate referring expression questions (Wei et al. 2024; Zhang et al. 2024a; Li et al. 2025), yet often neglect the role of VPs in visual understanding. For instance, RefCOCO (Kazemzadeh et al. 2014) evaluates referring expression capabilities but lacks VP-oriented image design, while HC-RefLoCo (Wei et al. 2024) extends expression length without addressing the contribution of VPs to regional comprehension in MLLMs. Recently, researchers have worked on developing VP-oriented benchmarks, such as the SoV (Zhang et al. 2024b) validation dataset and ViP-Bench (Cai et al. 2024), to provide a more thorough evaluation of VP cognition. However, these benchmarks still fall short when it comes to assessing the impact of VP on MLLM awareness and its effect on downstream tasks. For instance, while ViP-Bench offers a relatively comprehensive evaluation of VPs, it does not fully address the variations in VP shapes, attributes, and their effectiveness. These factors are crucial to understanding how different VPs influence MLLMs' visual comprehension and their potential to improve downstream task performance.

In contrast, our VP-Bench combines a broader range of VP shapes, attributes, and colors with a thorough examination of how VPs affect MLLM performance on various tasks. This comprehensive approach enables a more detailed assessment of model capabilities, providing insights into how VPs can be optimized for better task outcomes.

Key Statistics	
Statistic	**Number**
Total samples	38,932
Total images	34,267
VP Properties	
Shapes	8
Attributes	78
Colors	5
Debias Question	
Without Visualized VP	4.5%
Incorrect Model VP Instruction	8%

Table 2: Key statistics for the dataset. This table provides an overview of the total number of samples, images, and tasks, as well as the debias question amount over the total benchmark.

VP-Bench

Overview

VP-Bench is designed to assess the perception of VPs by MLLMs and to evaluate the impact of VPs on downstream task performance. Specifically, VP-Bench comprises 34,267 images and 38,932 multiple-choice questions. In constructing VP-Bench, we meticulously categorized VP shapes and attributes and developed a two-stage evaluation protocol. The Stage 1 data curation process assesses VP perception, while the Stage 2 data curation process examines the influence of VPs on models' regional perception in downstream tasks.

Compared to existing VP-related benchmarks, VP-Bench introduces several key improvements (see Table 1). First, by incorporating a substantially larger set of test samples, VP-Bench covers an extensive range of VP attribute combinations and diverse grounded referring expression tasks. In Stage 1, it evaluates MLLMs' perception across all 355 VP attribute combinations, exceeding the scope of other benchmarks by more than 40 times. This diversity compels models to develop a deeper understanding of visual information across various VP types. Second, in Stage 2, we determine the optimal VP combinations for each model and employ them to evaluate six downstream tasks within grounded referring scenarios, a critical gap that previous VP-focused

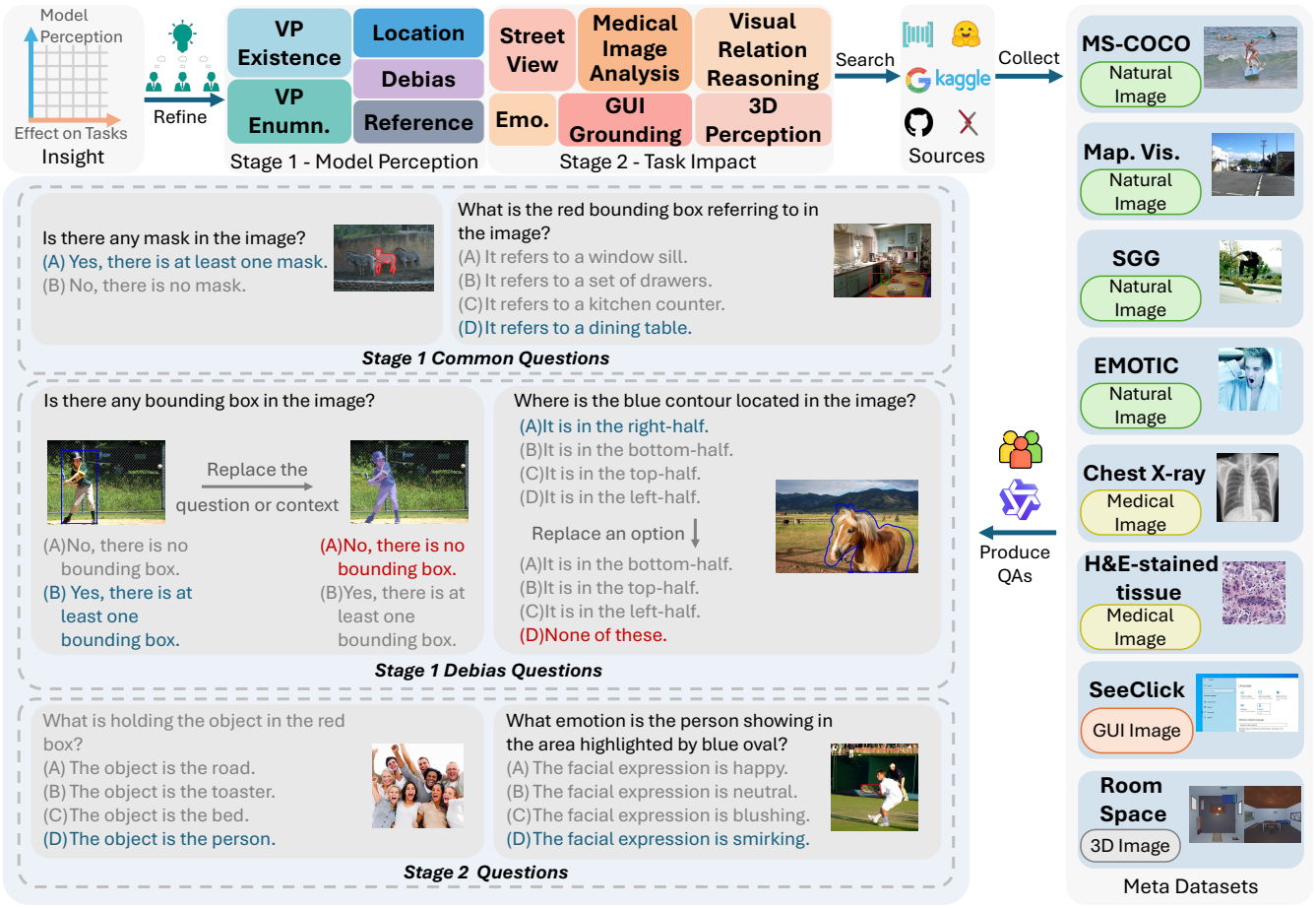

Figure 2: An illustration of our data collection and evaluation pipeline. Stage 1 constructs VP perception data by placing diverse VPs with different shapes, attributes, and colors, and generating multiple-choice questions to test existence, enumeration, localization, and referring capabilities, including debiased variants. Stage 2 builds VP-enabled downstream tasks across domains such as medical image, 3D recognition, natural image, and GUI understanding.

frameworks have yet to fully address. Additionally, we introduce debias questions, where either the images lack the VP type mentioned in the question or all substantively descriptive answer choices are incorrect (see Figure 2) to mitigate unreliable conclusions arising from MLLMs' hallucinations regarding VP presence. Detailed statistics of VP-Bench are provided in Table 2.

Leveraging the rich data available in VP-Bench, our analytical framework enables a comprehensive evaluation of both VP perception and VP-based referring tasks. The top-down hierarchical structure of VP attributes allows for comparative analyses of models' perception capabilities across various attributes. Moreover, the diverse downstream referring expression tasks facilitate performance estimation in practical scenarios, helping to delineate in-domain from out-of-domain tasks. Finally, the evaluation samples can be used to assess task learning difficulty, offering valuable insights for optimizing model training and dataset design. A detailed data curation breakdown is provided in the complementary material.

Visual Prompt Description

The VP description is designed to explicitly verbalize the spatial cues conveyed by visual prompts, rather than relying on the model to infer their meaning solely from the image. As summarized in Table 3, each VP shape encodes a distinct type of spatial signal (e.g., region extent, outline). However, without a corresponding textual description, these cues remain implicit and may be underutilized by current MLLMs. To address this gap, we propose a unified VP description scheme that converts each visual prompt into a short, structured phrase appended to the instruction. For example, a Bounding Box is described as "the red box outlining the target region". This mapping makes the semantics of each VP shape explicit in language. This design offers two key benefits. First, it reduces ambiguity by explicitly stating whether a mark denotes a point or region. Second, it improves vision–language alignment by mirroring the same spatial cue in both the image and the text, guiding the model to the correct region. Empirically, we observe consistent gains from adding VP descriptions, especially for

VP Shape	Context
Tag	A tag is a small label located at the center of a target, displaying a number or letter. It may be in red, blue, green, or black, and can be circular or square in shape.
Bounding Box	A bounding box is a rectangular frame that marks a target, which may be in red, blue, green, or black.
Arrow	An arrow is a symbol that points to a target, which may be in red, blue, green, or black.
Mask	A mask is a filled area used to indicate a target region, which may be in red, blue, green, or black.
Contour	A contour is the outline of a target, which may be in red, blue, green, or black. It can be drawn precisely along the outline or may resemble a loosely hand-drawn line.
Oval	An oval is an elliptical shape that encircles a target, which may be in red, blue, green, or black.
Point	A point is a square or circular dot that represents a target, located at the center of the marked target, which may be in red, blue, green, or black.
Scribble	A scribble is a random hand-drawn line that indicates a target, which may be in red, blue, green, or black.

Table 3: Definitions of the eight visual prompt (VP) shapes used in our benchmark. Each shape provides a distinct spatial cue (e.g., location, extent, or outline of the target) that is embedded into the model instructions.

complex shapes, showing that explicit textual VP semantics are key to fully leveraging visual prompts in MLLMs.

Stage 1 Data Curation Process

VPs play a pivotal role in guiding model attention and addressing complex problems. In Stage 1 evaluation, we treat VPs as visual cues defined by their intrinsic shapes, attributes, and colors, enabling more targeted information retrieval. We design the VP attributes according to a top-down hierarchical structure. Initially, we categorize VPs by shape, such as tag, bounding box, arrow, mask, contour, oval, point, and scribble, and further subdivide them by finer-grained attributes and colors. The specific VP shapes and their corresponding attributes are listed in the supplementary materials. The data in Stage 1 primarily comes from the MS-COCO dataset (Lin et al. 2014) due to its segmentation and bounding box annotations. These annotations are essential for generating VPs and their associated question-answer pairs to assess the capabilities of MLLMs.

To define the properties of each VP shape, we begin with a detailed breakdown. A **tag** has attributes including: (1) an alphabet or digit label at its center, with a circular or square shape, and (2) font size specifying the label's dimensions. A **bounding box** is defined by: (1) line width, representing the thickness of its outline, and (2) vertex shape, such as small squares or circles at the corners. An **arrow**, pointing from a selected direction to the target, has: (1) line width, and (2) pointer shape, which could be triangular or wedge-shaped. A **mask** is a filled region indicating the target, possibly accompanied by an outline, and consists of: (1) line width, and (2) style, which could either be a filled region with or without an outline. A **contour** captures the target's outline, characterized by: (1) line width, and (2) style, which may be precise or hand-drawn. An **oval** is an elliptical shape encircling the target, defined by the line width. A **point** is a small square or circular dot located at the target's center, described by: (1) point size (the area it occupies), and (2) point shape (square or circle). Finally, a **scribble** is a free-form line used to indicate a target, defined by the line width of the scribble.

To evaluate the effect of VPs on visual perception, fine-grained multiple-choice questions were generated for each VP type, covering presence, location, enumeration, and referring objects, with up to four answer options depending on the question type. First, question templates were created manually with guidance from GPT-4o (OpenAI 2024), and annotated answers from metadata were inserted into these templates. Second, distractors were produced either through manually crafted rules or by employing Qwen2-VL-72B (Wang et al. 2024a) with carefully designed prompts to ensure plausibility and quality. For instance, in "reference" questions, Qwen2-VL-72B generated realistic but incorrect options, whereas for "VP spatial location" questions, distractors were sampled based on randomized canvas positions. To enhance evaluation robustness, one debiased sample was included for every six entries. A debiased question consists of a visual image that excludes the VP referenced in the question, together with its corresponding QA pair, as illustrated by the *debias* example in the Stage 1 evaluation in Figure 1.

Overall, the Stage 1 data curation process establishes a comprehensive foundation for evaluating MLLMs' VP perception capabilities in a controlled setting. By systematically defining VP shapes, attributes, and associated QA pairs, this stage isolates the model's ability to interpret visual cues independent of downstream reasoning tasks. The resulting dataset enables fine-grained benchmarking of core VP competencies and serves as the basis for the Stage 2 evaluation, where VP understanding is integrated into complex, task-oriented scenarios.

Stage 2 Data Curation Process

Stage 1 evaluates the performance of different VP types and models in natural scenes. Following this, a new question arises: Can VPs contribute to a broader range of VP-related downstream tasks? How well do existing models perform in this regard? Hence, in Stage 2 evaluation, we carefully select several widely used application scenarios, such as natural recognition, medical image analysis, human facial recognition, street view recognition, traffic sign recognition, visual relation reasoning, GUI recognition, and 3D understanding, which are further categorized into 6 tasks. To investigate whether VPs enhance MLLMs' regional perception under downstream task scenarios, we evaluate both open-source

MLLMs and proprietary systems, using each model's best-performing VP from Stage 1 results.

For data curation, the target scenarios were first defined, followed by the specification of downstream tasks for each scenario. Relevant datasets were collected through searches on Google, Papers with Code, and Kaggle, and each dataset's suitability and relevance were systematically evaluated. The data were then organized into a standardized metadata format that includes task descriptions, questions, answers, input contexts, regional annotations, and images. This standardized format facilitated the construction of visual question–answer pairs, and the accuracy of the information was manually verified to ensure its compatibility with multiple-choice question generation. To maintain efficiency, each task was limited to a maximum of 200 randomly selected images with relevant entries, unless the dataset contained fewer samples. A detailed description of the metadata format is provided in the supplementary materials. The final collection comprises datasets including SZ-CXR (Stirenko et al. 2018), Gleason2019 (Nir et al. 2018), SD-100 (Li, Hogg, and Cohn 2024), Emotic (Kosti et al. 2020), MapillaryVistas (Neuhold et al. 2017), SeeClick (Cheng et al. 2024), and PSG (Yang et al. 2022). These were grouped into six downstream tasks: Medical Image Analysis (MIA) using SZ-CXR and Gleason2019; 3D object recognition using SD-100; facial emotion recognition using Emotic; street view recognition using MapillaryVistas; GUI element recognition using SeeClick; and scene graph generation (SGG) using PSG.

For question and answer generation, we adapt multiple-choice visual questions (with up to four options), drawing from each sample's metadata. We either craft rules manually or use Qwen2-VL-72B (Wang et al. 2024a) with carefully designed prompts to ensure efficient and high-quality generation. For instance, in 3D question-answering tasks, Qwen2-VL-72B generates plausible but incorrect distractors based on the question and the correct answer, while in visual relation reasoning tasks, we randomly select misleading item classes from the metadata as alternative options.

In summary, the Stage 2 dataset extends the evaluation of VPs beyond controlled natural scenes to diverse, task-driven scenarios. By leveraging the best-performing VP configurations from Stage 1, this stage enables a systematic assessment of how visual prompting contributes to regional perception and task-specific reasoning across multiple application domains. The curated datasets and standardized question–answer pairs provide a robust foundation for benchmarking MLLMs in realistic downstream contexts, thereby bridging the gap between VP perception and practical deployment.

Experiment

Experiment Setup

Compared Models. In this study, we selected 3 proprietary models along with 25 open-source MLLMs across various categories. These include popular visual models (e.g., Qwen2.5-VL, InternVL-3, and Llama-3.2-Vision), specialized models (e.g., MiniCPM-V 2.6, CogVLM2, and GLM-4V), and recently introduced vision-language architectures (e.g., NVLM, DeepSeek-VL2, Ovis2, and Molmo). To further illustrate the impact of model parameter scale and architecture on visual understanding, we incorporated models of different scales from the InternVL-3 and Qwen2.5-VL families.

Model	Enum.	Exist.	R.Loc.	Ref.	Avg.
Human Baseline					
Human	90.73	94.36	97.68	84.26	90.03
Proprietary Models					
GPT-4o	60.44	87.03	57.83	67.74	68.80
Doubao-Seed-1.6	62.47	64.04	<u>95.34</u>	**80.53**	70.37
Qwen-VL-Max	81.80	88.11	92.13	76.82	82.63
Pre-trained Models					
CogVLM2-19B	71.87	57.97	73.26	64.13	64.88
LLaVA-1.5-13B	83.17	**98.77**	60.17	60.58	75.47
Llama-3.2-90B	77.89	81.51	79.23	71.76	78.45
DeepSeek-VL2	73.92	91.41	80.86	71.37	77.94
LLaVA-O.V.-7B	79.80	84.99	91.47	69.43	79.58
GLM-4V-9B	**91.37**	91.44	70.77	74.20	82.36
Qwen2.5-VL-72B	81.84	88.26	92.01	75.79	82.80
Ovis2-34B	84.72	88.19	92.49	75.59	83.24
LLaVA-v1.6-34B	82.62	94.97	93.86	72.24	84.97
NVLM-D-72B	83.82	93.99	92.06	76.50	85.39
Molmo-72B	87.18	<u>97.24</u>	92.46	70.08	85.61
InternVL3-78B	<u>88.59</u>	92.93	95.24	<u>79.65</u>	**87.97**
Fine-tuned with Visual-Prompt Data					
ViPLLaVA-13B	80.36	61.43	87.70	57.28	68.12
ViPLLaVA-7B	77.51	84.84	83.01	53.58	72.58
Average	79.92	85.62	84.03	70.21	78.61

Table 4: Accuracy (%) on question-type subtasks: Enumeration (Enum.), Existence (Exist.), Rough Localization (R.Loc.), and Visual-Prompt Reference (Ref.). Best non-human result for each column is highlighted in **bold**, and the second-best is <u>underlined</u>.

Evaluation Metrics. Our proposed VP-Bench is comprised of two stages, both stages are in multi-choice question format, e.g., "Where is VP located? Options: (A) In the bottom, (B) In the top". Generally, we follow the VLMEvalKit (Duan et al. 2024) procedure to evaluate models' performance. Accuracy is the primary metric.

Evaluation Main Results

This section evaluates MLLMs on VP-Bench alongside Human baselines. We report the overall score for all perceptual tasks in Table 4 as well as the best performance on each downstream task in Table 5. Various instruction arrangements for all tasks are investigated. We summarize the key findings as follows.

In Stage 1 evaluation, we present the average accuracy of all models across eight VP shapes and four question types in Table 4, along with the human baseline. "Avg." denotes the average accuracy across all QA samples. In terms of overall accuracy, InternVL3-78, InternVL3-38B, and Molmo-72B rank among the top three, with accuracy around 87%. Other

Model	Best VP Attr. Comb.	MIA	SD	Emo.	M.V.	S.C.	SGG	Avg.
Proprietary Models								
Doubao-Seed-1.6	bbox(contrast, round, thick)	59.60	**89.33**	70.48	76.08	97.22	**96.18**	81.48
GPT-4o	Tag(digit, blue, round, small)	<u>60.60</u>	72.24	66.04	57.89	97.00	81.60	72.56
Qwen-VL-Max	bbox(contrast, none, medium)	44.32	81.27	73.03	62.68	96.89	95.44	75.61
Pre-trained Models								
CogVLM2-19B	oval(contrast, thin)	5.40	31.33	70.71	33.49	75.31	83.30	49.92
DeepSeek-VL2	tag(alphabet, green, round, large)	20.90	23.00	63.76	24.88	67.50	83.30	47.22
GLM-4V-9B	contour(contrast, contour, thick)	18.19	59.33	64.94	53.59	96.25	86.32	63.10
InternVL3-78B	bbox(red, none, medium)	**53.40**	87.00	<u>73.65</u>	<u>67.46</u>	97.75	95.03	**79.05**
Llama-3.2-90B	bbox(green, square, medium)	59.50	58.00	65.65	51.20	98.50	88.99	70.31
LLaVA-v1.5-7B	tag(digit, red, square, medium)	35.30	31.00	33.65	13.88	83.50	73.87	45.20
LLaVA-v1.5-13B	tag(alphabet, red, square, large)	30.60	24.00	39.76	17.22	80.25	83.30	45.85
LLaVA-O.V.-7B	oval(blue, thick)	16.72	63.67	64.39	47.37	98.50	92.01	63.78
LLaVA-v1.6-34B	oval(contrast, thin)	37.60	52.33	64.05	48.33	96.30	91.30	64.98
MiniCPM-V-2_6	tag(digit, red, square, large)	15.30	38.00	67.06	26.32	92.75	89.08	54.75
Molmo-72B	bbox(contrast, square, thin)	62.80	78.00	67.38	60.77	96.30	91.83	<u>76.18</u>
NVLM-D-72B	bbox(red, square, thick)	58.10	77.00	72.71	49.76	97.00	93.43	74.67
Ovis2-34B	bbox(contrast, square, thin)	37.30	83.33	73.10	50.72	**99.69**	**94.58**	73.12
Qwen2.5-VL-72B	bbox(contrast, square, thin)	43.80	83.33	71.43	61.72	97.53	<u>95.47</u>	75.55
Fine-tuned with Visual-Prompt Data								
ViP-LLaVA-7B	contour(blue, contour, medium)	36.80	36.33	46.59	11.00	80.25	83.84	49.13
ViP-LLaVA-13B	mask(red, fill, medium)	33.60	28.00	59.53	33.97	82.50	88.72	54.39
Average		37.39	61.96	64.96	47.35	93.23	90.01	

Table 5: Performance comparison on VP-related tasks using BVP scores. "Best VP Attr. Comb." indicates the Stage 1 attribute combination. Best in each task is **bold** and second-best is <u>underlined</u>.

statistics are as follows.

Across all **question types**, most models accurately detect, count, and localize VP, achieving over 90% accuracy on existence queries, around 85% on enumeration queries, and over 92% on rough-location queries. Their performance on referring-expression resolution remains substantially lower: mean accuracy hovers around 70%, and the community favored Qwen-2.5-VL-72B reaches only 75.79%. When benchmarked against human annotators, MLLMs still exhibit an around 10% deficit.

An appropriate **VP shape** can significantly enhance a model's ability to detect both the prompt and the highlighted region. For example, the bounding box shape is generally a better comparison to the point shape in Table 4, as models can achieve an average of 87.49% with the bounding box shape, but only 67.22% accuracy with the point shape. As well as reflecting most of the models' best recognized **VP attributes** combination, there are variations of the bounding box shape in Table 5. The results indicate that a contrast color bounding box with medium thickness is optimal for more than half of the models, while a contrast color oval with thin edges is also the most effective for many models, where the contrast color is selected as one of the most distinctive red, green, or blue hues relative to the background. Overall contrast color emerges as the preferred choice for most models. This suggests that color plays a crucial role in distinguishing the region of interest from the background. Regarding scale, thin to medium scale VP yields better results, implying that MLLMs exhibit greater perceptual sensitivity to thin prompts while preserving more contextual information.

Results Analysis

To further interrogate our preliminary findings, we carried out a series of supplementary experiments whose results yielded several salient statistical patterns. These additional tests not only corroborate the baseline trends but also uncover nuanced insights into the models' behavior under varying prompt conditions. The remainder of this section highlights the principal observations before delving into a detailed, case-by-case analysis.

Shape regular VPs are more readily perceived by MLLMs than shape irregular ones. As shown in Table 4, models recognize regular VPs like Tag, Arrow, Bounding Box, Oval, at around 80% accuracy on average, whereas for irregular VPs like Mask, Point, Scribble, they only reach less than 70%. Moreover, the gap relative to the human baseline is larger on those irregular shapes. A per-model breakdown shows that InternVL3-78B, while still 9.79% behind humans on Point, differs by only about 2% on Mask and Scribble, and even exceeds human performance on irregular contours (91.56% compared to 87.68%). Notably, DeepSeek-VL2 and InternVL3-14B suffer their worst scores on hand-drawn scribbles. This pattern suggests that training data are biased toward regular geometric forms, with fewer examples of irregular shapes, which in turn impairs MLLMs' ability to detect those VP forms.

An explicit description of VP shapes is critical for enabling MLLMs to interpret them in context. To examine whether models truly understand VPs, we compared two

Model	Tag	Arrow	BBox	Contour	Mask	Oval	Point	Scribble	Avg.
InternVL3-78B									
w/o VP description	91.69	85.45	94.06	90.19	50.77	92.49	69.21	76.52	81.30
w. VP description	93.87	85.77	94.25	91.56	80.01	95.81	81.59	80.89	87.97
Qwen2.5-VL-72B-Instruct									
w/o VP description	85.93	81.96	92.88	87.81	41.38	82.80	72.63	69.76	76.89
w. VP description	92.26	82.57	92.88	88.62	68.68	92.83	69.58	74.77	82.77

Table 6: Comparison of model performance with and without VP descriptions on Stage 1 evaluation. **Tag** is the mean of Alphabet and Digit. **Avg.** is the mean over {Tag, Arrow, BBox, Contour, Mask, Oval, Point, Scribble}.

settings: with and without a VP-shape description inserted into the instruction. As shown in Table 6, experiments were conducted on InternVL3-78B and Qwen2.5-VL-72B, incorporating VP descriptions consistently enhanced the models' comprehension of VPs, though the effect varied across VP types. For instance, when VP descriptions were included in the **Arrow** and **Contour** scenarios, InternVL3-78B achieved only marginal improvements of 0.32% and 1.37%, respectively, while Qwen2.5-VL-72B improved by just 0.61% and 0.81%. In contrast, substantial gains were observed in the **Mask** and **Point** scenarios: InternVL3-78B improved by 29.24% and 12.38%, respectively, whereas Qwen2.5-VL-72B improved by 27.3% in the **Mask** scenario but declined by 3.05% in the **Point** scenario. These differences are likely influenced by the distribution of training data, where the frequency of each VP type affects the extent to which descriptive prompts enhance model performance. Collectively, these results suggest that providing an explanatory language description of the VP can help models more accurately identify and interpret the corresponding visual cue in the image.

VPs that a model perceives most accurately are generally the best choices for downstream tasks but not always. To investigate this, we devised two selection schemes:

1. **Random Best VP (R-BVP):** is randomly selected from the set of VP attribute combinations that achieved the highest perception performance across all models in the Stage 1 evaluation.
2. **Best VP (BVP):** always uses the single VP attribute combination the current model perceives most accurately.

As shown in Table 5, for Qwen-2.5-VL-72B, the performance gap between BVP and R-BVP stays within $\pm$ 5%, and in over half of the tasks BVP outperforms R-BVP. Likewise, InternVL3-78B shows slightly better results with BVP in five tasks, for example, a +5.85% gain on MIA and +1.00% on SD-100, and only trails by 0.26% on SeeClick. By contrast, DeepSeek-VL2 exhibits much larger disparities: in most downstream evaluations, BVP underperforms R-BVP (–11.71% on MapillaryVistas and –19.43% on SeeClick), yet it exceeds R-BVP by +1.68% on MIA. A similar pattern appears with GPT-4o: BVP is +6.35% better on MIA but –10.89% worse on MapillaryVistas. These findings suggest that, although a model's perception accuracy is generally the main criterion for choosing a VP, a robust VP-selection strategy is also critical for maximizing downstream performance.

Simply training models with VP data offers no clear benefit. ViP-LLaVA extends the LLaVA architecture by incorporating VP-related data during the instruction tuning stage to enhance VP perception, while LLaVA-1.5 serves as the baseline model in our experiments for comparison against ViP-LLaVA trained with VP-enriched datasets. In Stage 1, LLaVA-1.5-7B achieved an average accuracy of 64.33%, whereas ViP-LLaVA-7B reached 73.01%. However, in Stage 2 downstream tasks, ViP-LLaVA-7B underperformed LLaVA-1.5-7B by 2.88% on MapillaryVistas and by 3.25% on SeeClick. Furthermore, this straightforward training approach led to degradation at larger scales: ViP-LLaVA-13B's Stage 1 accuracy dropped by 4.13% relative to ViP-LLaVA-7B and by 6.72% relative to LLaVA-1.5-13B. These findings indicate that downstream performance depends more on a model's robust foundational capabilities and that improving VP perception without compromising these core abilities requires more balanced data composition and refined training strategies.

Conclusion

We introduce VP-Bench, a two-stage evaluation framework for assessing the capabilities of MLLMs in perceiving VP and solving grounded referring queries. In Stage 1, we construct a dataset from 34,267 images, covering 8 distinct VP shapes and 355 attribute combinations, to evaluate a model's general understanding of VPs. Our results show that while MLLMs perform well in VP recognition and object counting, they struggle with spatial localization and fine-grained understanding. Additionally, existing models exhibit a preference for bounding boxes and tags and show heightened sensitivity to red VPs. In Stage 2, we introduce 6 VP-related downstream tasks to evaluate how well models integrate VP perception for practical problem-solving. Experimental results suggest that VPs offer certain advantages over text-based spatial prompts for these tasks. However, model performance remains largely dependent on domain knowledge. Overall, this work aims to highlight the need to refine VP attribute representations and enhance spatial reasoning, ultimately improving model interpretability and real-world applicability.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 62502174).

References

Cai, M.; Liu, H.; Mustikovela, S. K.; Meyer, G. P.; Chai, Y.; Park, D.; and Lee, Y. J. 2024. ViP-LLaVA: Making large multimodal models understand arbitrary visual prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12914–12923.

Chen, J.; Guo, H.; Yi, K.; Li, B.; and Elhoseiny, M. 2022. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18030–18040.

Cheng, K.; Sun, Q.; Chu, Y.; Xu, F.; YanTao, L.; Zhang, J.; and Wu, Z. 2024. SeeClick: Harnessing GUI Grounding for Advanced Visual GUI Agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9313–9332.

Duan, H.; Yang, J.; Qiao, Y.; Fang, X.; Chen, L.; Liu, Y.; Dong, X.; Zang, Y.; Zhang, P.; Wang, J.; et al. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, 11198–11201.

Fu, X.; Hu, Y.; Li, B.; Feng, Y.; Wang, H.; Lin, X.; Roth, D.; Smith, N. A.; Ma, W.-C.; and Krishna, R. 2024. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, 148–166. Springer.

Guan, T.; Liu, F.; Wu, X.; Xian, R.; Li, Z.; Liu, X.; Wang, X.; Chen, L.; Huang, F.; Yacoob, Y.; et al. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14375–14385.

Guo, Q.; De Mello, S.; Yin, H.; Byeon, W.; Cheung, K. C.; Yu, Y.; Luo, P.; and Liu, S. 2024. Regiongpt: Towards region understanding vision language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13796–13806.

Huang, S.; Dong, L.; Wang, W.; Hao, Y.; Singhal, S.; Ma, S.; Lv, T.; Cui, L.; Mohammed, O. K.; Patra, B.; Liu, Q.; Aggarwal, K.; Chi, Z.; Bjorck, N.; Chaudhary, V.; Som, S.; SONG, X.; and Wei, F. 2023. Language Is Not All You Need: Aligning Perception with Language Models. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 72096–72109. Curran Associates, Inc.

Kazemzadeh, S.; Ordonez, V.; Matten, M.; and Berg, T. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 787–798.

Kosti, R.; Alvarez, J. M.; Recasens, A.; and Lapedriza, A. 2020. Context Based Emotion Recognition Using EMOTIC Dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(11): 2755–2766.

Li, F.; Hogg, D. C.; and Cohn, A. G. 2024. Reframing Spatial Reasoning Evaluation in Language Models: A Real-World Simulation Benchmark for Qualitative Reasoning. In Larson, K., ed., *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 6342–6349. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, Y.; Huang, H.; Chen, C.; Huang, K.; Huang, C.; andZhiyuan Liu, Z. G.; Xu, J.; Li, Y.; Li, R.; and Sun, M. 2025. Migician: Revealing the Magic of Free-Form Multi-Image Grounding in Multimodal Large Language Models. arXiv:2501.05767.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, 740–755.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 34892–34916. Curran Associates, Inc.

Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. 2024. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, 216–233. Springer.

Meng, F.; Li, C.; Wang, J.; Lu, Q.; Tian, H.; Yang, T.; Liao, J.; Zhu, X.; Dai, J.; Qiao, Y.; et al. 2025. MMIU: Multimodal Multi-image Understanding for Evaluating Large Vision-Language Models. In *The Thirteenth International Conference on Learning Representations*.

Neuhold, G.; Ollmann, T.; Rota Bulo, S.; and Kontschieder, P. 2017. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, 4990–4999.

Nir, G.; Hor, S.; Karimi, D.; Fazli, L.; Skinnider, B. F.; Tavassoli, P.; Turbin, D.; Villamil, C. F.; Wang, G.; Wilson, R. S.; et al. 2018. Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts. *Medical image analysis*, 50: 167–180.

OpenAI. 2023. GPT-4V(ision) system card. https://openai.com/index/gpt-4v-system-card/. Accessed: 2023.

OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 2024.

Rasheed, H.; Maaz, M.; Shaji, S.; Shaker, A.; Khan, S.; Cholakkal, H.; Anwer, R. M.; Xing, E.; Yang, M.-H.; and Khan, F. S. 2024. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13009–13018.

Stirenko, S.; Kochura, Y.; Alienin, O.; Rokovyi, O.; Gordienko, Y.; Gang, P.; and Zeng, W. 2018. Chest X-Ray Analysis of Tuberculosis by Deep Learning with Segmentation and Augmentation. In *2018 IEEE 38th International Conference on Electronics and Nanotechnology (ELNANO)*, 422–428.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Fan, Y.; Dang, K.; Du, M.; Ren, X.; Men, R.; Liu, D.; Zhou, C.; Zhou, J.; and Lin, J. 2024a. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. arXiv:2409.12191.

Wang, W.; Shi, M.; Li, Q.; Wang, W.; Huang, Z.; Xing, L.; Chen, Z.; Li, H.; Zhu, X.; Cao, Z.; Chen, Y.; Lu, T.; Dai, J.; and Qiao, Y. 2024b. The All-Seeing Project: Towards Panoptic Visual Recognition and Understanding of the Open World. In *The Twelfth International Conference on Learning Representations*.

Wei, F.; Zhao, J.; Yan, K.; Zhang, H.; and Xu, C. 2024. A large-scale human-centric benchmark for referring expression comprehension in the LMM era. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 69566–69587.

Wu, M.; Cai, X.; Ji, J.; Li, J.; Huang, O.; Luo, G.; Fei, H.; Jiang, G.; Sun, X.; and Ji, R. 2024. ControlMLLM: Training-Free Visual Prompt Learning for Multimodal Large Language Models. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 45206–45234. Curran Associates, Inc.

Yang, J.; Ang, Y. Z.; Guo, Z.; Zhou, K.; Zhang, W.; and Liu, Z. 2022. Panoptic scene graph generation. In *European conference on computer vision*, 178–196. Springer.

Yang, J.; Zhang, H.; Li, F.; Zou, X.; Li, C.; and Gao, J. 2023. Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V. arXiv:2310.11441.

Ying, K.; Meng, F.; Wang, J.; Li, Z.; Lin, H.; Yang, Y.; Zhang, H.; Zhang, W.; Lin, Y.; Liu, S.; et al. 2024. MMT-Bench: A Comprehensive Multimodal Benchmark for Evaluating Large Vision-Language Models Towards Multitask AGI. In *International Conference on Machine Learning*, 57116–57198. PMLR.

Yu, W.; Yang, Z.; Li, L.; Wang, J.; Lin, K.; Liu, Z.; Wang, X.; and Wang, L. 2024. MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities. In *International Conference on Machine Learning*, 57730–57754. PMLR.

Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9556–9567.

Zhang, H.; Li, H.; Li, F.; Ren, T.; Zou, X.; Liu, S.; Huang, S.; Gao, J.; Leizhang; Li, C.; et al. 2024a. Llava-grounding: Grounded visual chat with large multimodal models. In *European Conference on Computer Vision*, 19–35. Springer.

Zhang, Q.; Wang, Z.; Zhang, D.; Niu, W.; Caldwell, S.; Gedeon, T.; Liu, Y.; and Qin, Z. 2024b. Visual Prompting in LLMs for Enhancing Emotion Recognition. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 4484–4499.

Zhang, S.; Huang, D.; Deng, J.; Tang, S.; Ouyang, W.; He, T.; and Zhang, Y. 2024c. Agent3d-zero: An agent for zero-shot 3d understanding. In *European Conference on Computer Vision*, 186–202. Springer.

Zhou, S.; Chang, H.; Jiang, S.; Fan, Z.; Zhu, Z.; Xu, D.; Chari, P.; You, S.; Wang, Z.; and Kadambi, A. 2024a. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21676–21685.

Zhou, X.; Ran, X.; Xiong, Y.; He, J.; Lin, Z.; Wang, Y.; Sun, D.; and Yang, M.-H. 2024b. GALA3D: Towards Text-to-3D Complex Scene Generation via Layout-guided Generative Gaussian Splatting. In *International Conference on Machine Learning*, 62108–62118. PMLR.