Progressive Multi-granular Alignments for Grounded Reasoning in Large Vision-Language Models

Quang-Hung Le[1], Long Hoang Dang[2], Ngan Hoang Le[3], Truyen Tran[1], Thao Minh Le[1]

[1]Applied Artificial Intelligence Institute (A^2I^2), Deakin University, Australia
[2]Posts and Telecommunications Institute of Technology, Vietnam
[3]University of Arkansas, USA
q.le@deakin.edu.au, longdh@ptit.edu.vn, thile@@uark.edu, truyen.tran@deakin.edu.au, thao.le@deakin.edu.au

Abstract

Existing Large Vision-Language Models (LVLMs) excel at matching concepts across multi-modal inputs but struggle with compositional concepts and high-level relationships between entities. This paper introduces Progressive multi-granular Vision-Language alignments (PromViL), a novel framework to enhance LVLMs' ability in performing grounded compositional visual reasoning tasks. Our approach constructs a hierarchical structure of multi-modal alignments, ranging from simple to complex concepts. By progressively aligning textual descriptions with corresponding visual regions, our model learns to leverage contextual information from lower levels to inform higher-level reasoning. To facilitate this learning process, we introduce a data generation process that creates a novel dataset derived from Visual Genome, providing a wide range of nested compositional vision-language pairs. Experimental results demonstrate that our PromViL framework significantly outperforms baselines on various visual grounding and compositional question answering tasks.

Introduction

Large Vision-Language Models (LVLMs) (Liu et al. 2024; Li et al. 2023b), pre-trained on vast amounts of image-text data, hold great promise in solving complex vision-language (V-L) tasks. However, current LVLMs still fall short in compositional reasoning – the ability to answer complex queries composed of smaller elements (Ma et al. 2023; Zhao et al. 2022). Compositionality is a pervasive phenomenon, seen in language with syntax trees, in vision with scenes and objects, and indeed wherever "the meaning of the whole is a function of the meanings of its parts" (Cresswell 2016).

When faced with a complex visual query, humans can recursively decompose it into smaller components while simultaneously grounding them to corresponding visual elements (Bottou 2014; Hupkes et al. 2020; Janssen and Partee 1997). Current LVLMs do not yet possess that ability. Most existing models (Huang et al. 2024; Li et al. 2023b) process whole textual prompts and entire images, relying on associations between holistic image embeddings and textual embeddings. These methods disregard interactions between sentence parts and image components, leading to poor grounding of textual information (Li et al. 2023a) and subsequently to suboptimal performance in reasoning tasks.

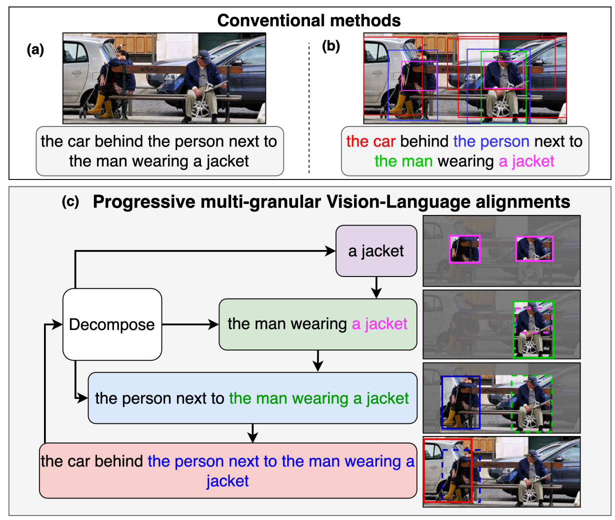

Figure 1: **Comparison with Existing LVLMs:** (a) Coarse-grained: Whole image/region with full text, lacks object details. (b) Fine-grained: Simple phrases and bounding boxes, lacks relational context. (c) PromViL employs hierarchical multi-granular associations, progressively utilizing simpler concepts as cues to understand more complex ones.

Others, like Kosmos-2 (Peng et al. 2024), Pink (Xuan et al. 2024), and CoVLM (Li et al. 2023a) address these drawbacks by infusing location information corresponding to visual entities into the language generation process to enhance grounding ability. As a result, they can ground simple concepts involving individual objects, but still struggle with scenarios involving multiple objects and complex relationships. This limitation may stem from their underlying grounding processes. The coarse-grained processes (Fig. 1.a) align complex expressions (e.g., "the car behind the person next to the man with a jacket") to a region but ignore object-level alignments, which is crucial for tasks like visual reasoning and image captioning (Zeng, Zhang, and Li 2021). The fine-grained processes (Fig. 1.b), on the other hand, align concepts with single objects but struggle to understand relations among multiple objects, especially in ambiguous scenarios (e.g., identifying "man with a jacket" when multiple men are present). Both approaches inadequately capture nuanced relationships between textual descriptions and visual elements, limiting effectiveness in tasks requiring sophisticated compositional reasoning.

Addressing these limitations, we propose a novel compositional reasoning framework that leverages off-the-shelf LVLMs in an 'agentic flow' style. Our approach, applicable to any grounded LVLM, integrates *multi-granular language-vision training* with *progressive reasoning*: >We consolidate the model's grounding ability through a dataset construction pipeline that leverages the Visual Genome dataset (Krishna et al. 2017) to curate nested compositional V-L pairs. This enables training on multiple complexity levels, allowing models to learn visual relations of unbounded complexity. Unlike previous methods, it is not constrained by the number of concepts in textual information or limited to specific levels of visual information. >Rather than inputting the entire question and image at once, we decompose the text into object-centric components with gradually increasing complexity. We then prompt the model to perform step-by-step reasoning, progressing from simple to complex. Visual information from each step is fed back to the model in subsequent steps, gradually guiding it through grounding and reasoning processes. We call the method PromViL (**Pro**gressive **m**ulti-granular **Vi**sion-**L**anguage alignments) and show its behaviors in Fig. 1.

We conduct extensive experiments to demonstrate the effectiveness of our framework: Compared to other models of the same size, PromViL, with only 4.9% tunable parameters and 60K fine-tuning data samples, shows significant improvements: an approximately 9.0 point increase on our benchmark; up to 5.5 on zero-shot grounding tasks, respectively; and nearly 5 point and 10 point increases in accuracy and validity, respectively, on zero-shot compositional reasoning task. Importantly, PromViL surpasses larger models like CoVLM (2.4B) and Pink (7B) on grounding tasks, and exceeds baselines fine-tuned with twice the VQA data on compositional reasoning tasks.

Our main contributions are threefold: (1) We propose PromViL, a novel framework integrating multi-granular language-vision training with progressive reasoning, allowing models to ground and reason in scenarios with intricate textual information and multiple visual relations. (2) We introduce a dataset construction pipeline to create a new dataset of nested compositional V-L pairs curated from Visual Genome, enabling training on multiple complexity levels. (3) We conduct extensive experiments demonstrating PromViL's effectiveness in handling complex visual scenes and linguistic descriptions, outperforming existing approaches on various benchmarks. Our experiments are reproducible in academia, using only public data and models. The model can be trained on consumer GPUs with 32GB memory. We will release our code and datasets to support further research in this field.

Related Works

Large Vision-Language Models (LVLMs) have shown impressive capabilities in tasks like image captioning and visual question answering (Liu et al. 2024; Bai et al. 2023; Wang et al. 2022). Recent methods enable phrase-region alignments, allowing models to perceive image regions and ground text to visual entities (Chen et al. 2023; Li et al. 2023a; Peng et al. 2024; Xuan et al. 2024). However, these approaches often struggle with compositional tasks involving concepts of varying complexity attended by multiple objects and relations (Yuksekgonul et al. 2023; Li et al. 2023a). Our method addresses this limitation by introducing hierarchical multi-granular V-L alignments and progressive reasoning, enabling multi-step grounding and reasoning from simple to complex tasks. Previous studies have shown the benefits of utilizing multi-granularity information in training LVLMs in improving textual and visual alignments (Zeng, Zhang, and Li 2021; Le et al. 2020; Dang et al. 2021; Gao et al. 2022; Chen et al. 2024). However, existing approaches often use single granularity levels or obtain information in embedding space, contradicting the varying structure of text and images. Our method derives granularity levels based on language structure, creating a suitabletailored hierarchical representation for each input. This enables the model to learn multimodal alignments across granularities, enhancing its ability to reason across textual and visual inputs effectively.

Grounding datasets used in LVLMs can be broadly categorized into two groups: >The coarse-grained group such as RefCOCO/+ (Kazemzadeh et al. 2014), RefCOCOg (Mao et al. 2016), and GRIT (Peng et al. 2024), provide pairs of phrases and corresponding image bounding boxes. However, for phrases containing multiple objects, they lack bounding boxes for individual elements. As a result, models trained on these datasets struggle to learn fine-grained alignments between vision and language, e.g., object level (Zeng, Zhang, and Li 2021). >The fine-grained group such as Flickr30K (Plummer et al. 2015) and Objects365 (Shao et al. 2019) provide simple phrases describing single objects (e.g., "the person", "the jacket") with corresponding bounding boxes. While these datasets enable object-centric feature learning, they often fail to represent relationships among multiple objects effectively (e.g., "the person with jacket"). This becomes particularly problematic when dealing with ambiguous images that require relational context to locate correct entities (e.g., multiple "persons", as illustrated in Fig. 1).

Several LVLMs have utilized both groups of datasets (Chen et al. 2023; Xuan et al. 2024), but they still fall short as each sample maintains a single granularity level, and the fine-grained and coarse-grained data remain unrelated to each other within samples. Our data generation approach addresses this limitation, resulting in a new dataset named CompoVL, based on the Visual Genome dataset (Krishna et al. 2017). CompoVL provides multi-grained data for each data instance, offering richer information for training and aligning well with our method's objectives.

Preliminaries

Most state-of-the-art LVLMs such as LlaVa (Liu et al. 2024), Flamingo (Alayrac et al. 2022), and BLIP-2 (Li et al. 2023b), are extensions of language-based LLMs designed to generate text responses to input images and text prompts.

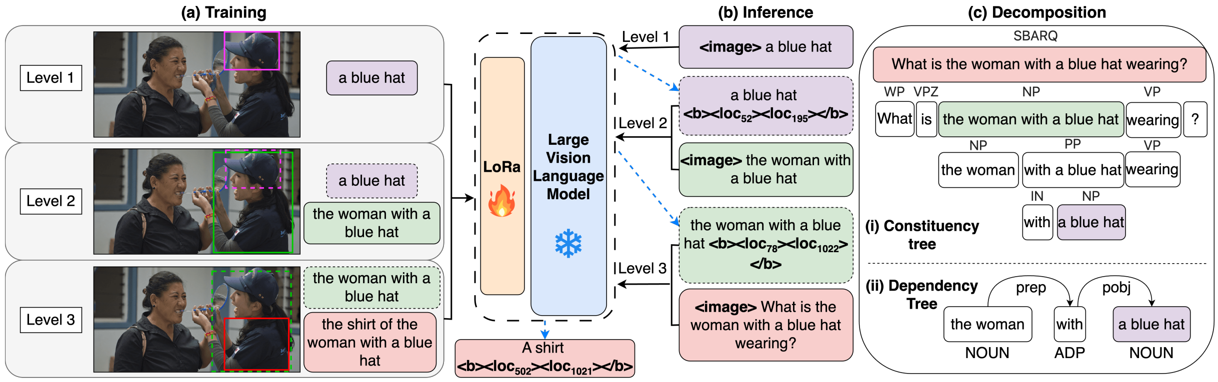

Figure 2: Overview of our PromViL framework. (a) Training: Learn multi-level visual entities-textual expression associations. (b) Inference: Progressively prompt from simple to complex, using prior responses as clues. (c) Decomposition: Extract nested subsequences based on (i) constituency parsing (simplified for illustration) and (ii) dependency parsing.

While capable of handling diverse vision-to-language tasks, these models only scratch the surface of holistic visual scene understanding and fall short in fine-grained comprehension of specific visual regions of interest.

A new family of LVLMs has emerged to address these drawbacks by producing visual answers such as bounding boxes or segmentation masks. Examples include Kosmos-2 (Peng et al. 2024), Pink (Xuan et al. 2024), and MiniGPTv2 (Chen et al. 2023). By locating specific image regions, these models alleviate ambiguities of text-only descriptions and offer insights into the decision-making processes of LVLMs.

To achieve this, these models process multimodal token sequences combining text spans and their corresponding spatial locations (e.g., bounding box coordinates), which are placed next to each other. Trained on vast grounded image-text datasets using next-token prediction, they are able to generate both text descriptions of the region of interest and spatial tokens indicating object location within the image. For instance, for the query ”The woman with a blue hat” and its bounding boxes correspondence, Kosmos-2 employs a markdown-like input format: ”*<s><img>Image Embedding</img> <grounding> <p>The woman</p> <loc*$_{78}$*><loc*$_{1022}$*> with <p>a blue hat</p> <loc*$_{52}$*><loc*$_{159}$*> </s>*”. *<s>* and *</s>* are start and end tokens of the text sequence; *<img>* and *</img>* refer to image embedding; *<p>* and *</p>* indicate the boundaries of non-relational concepts within input text sequence, and ** and ** refer to their corresponding bounding box locations. These models' output include both text tokens describing the visual content of the region of interest and spatial location tokens locating the object's position within the visual scene.

Compositional visual reasoning requires understanding two-way interactions between sentence parts and corresponding image regions. However, current LVLMs focus on the alignments of either overly fine-grained or coarse-grained concepts and image regions. Fine-grained concepts include non-relational individual concepts, such as “a woman” or “a blue hat”, while high-order relational concepts might be “the shirt of the woman with a blue hat”. This limitation partly stems from a lack of intermediate connections in existing datasets. We propose a novel mechanism to obtain such connections, bridging the generalization gap from fine-grained concepts to coarse-grained concepts.

Methods

Progressive multi-granular V-L Alignments

Given a compositional input sentence, we decompose it into a series of *nested subsequences*. This input sentence and elements in the nested subsequences are later referred to as “expressions” for the remainder of this paper. These expressions cover concepts of varying complexity, ranging from individual concepts (e.g., “a woman”, “a blue hat”) to high-order relational concepts (e.g., “the woman with a blue hat”, “the shirt of the woman with a blue hat”).

Our approach, termed ***Pro****gressive* ***m****ulti-granular* ***Vi****sion-****L****anguage alignments (PromViL)*, leverages the alignments between vision-language pairs of increasing levels of complexity to create a *progressive chain of reasoning steps* for understanding complex compositional expressions. Assuming we have access to nested vision-language (V-L) pairs, our task is to direct LVLMs to iteratively leverage feedback from lower levels to properly align more complex expressions with their corresponding visual regions (See Fig. 2). The next section details our method for generating nested V-L pairs from existing data.

Multi-granular Compositional V-L Dataset

Nested Vision-Language Pairs Generation: To generate nested V-L pairs , we utilize Visual Genome (VG) (Krishna et al. 2017) annotations and an open-source LLM for text generation (Mixtral8x7B (Jiang et al. 2024) in our implementation). Our data generation pipeline (See Fig. 3) begins by assigning the level of complexity of an input expressions based on the depth of relational steps needed to arrive at the main entity of interest. Level-one expressions include non-relational concepts like ”the woman” or ”a horse” , while level-two expressions combine these with a relationship, such as ” the woman riding a horse” . Our level-one and

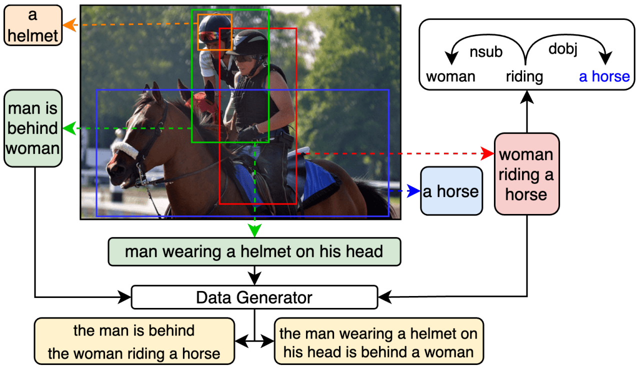

Figure 3: Multi-granular Compositional V-L Data Generation. We create a novel dataset with rich, multi-granular V-L data using existing VG annotations.

level-two pairs use direct VG annotations. For higher levels, we instruct the LLM to generate text descriptions from VG predicates <*subject, relation, object*> that share entities in common. For example, in Fig. 3, to generate the expression “the man is behind the woman riding a horse”, we first retrieve predicates <*man, behind, woman*> and <*woman, riding, horse*>. Then, we instruct the LLM to produce more complex textual descriptions by combining and extending information from these two input predicates. To reduce the effects of hallucinations by LLMs, we direct them to strictly adhere to the genuine objects and relationships in provided input predicates and not to invent any new objects and relationships (See Appendix for our prompts used). To assign corresponding visual bounding boxes to these generated descriptions, we employ a dependency parser (Nivre 2008) to identify the main object of interest within a given expression. The visual bounding boxes are then chosen as bounding boxes of the identified main entities in VG. For instance, the bounding box corresponding to ” the man” is chosen as the bounding box for the phrase ” the man is behind the woman riding a horse” .

From these generated descriptions and VG annotations, we construct a hierarchical series of nested compositional V-L pairs, progressing from simple to complex. Lower-level expressions form building blocks for higher-level ones by extending them as referential components. This results in a list of nested expressions such as: ”a horse” (level-one), ”the woman riding a horse” (level-two), and ”the man is behind the woman riding a horse” (level-three) (Fig. 3). Multiple expressions can exist at each level. In total, we generate 29K such lists of nested expressions, comprising up to 115K individual V-L pairs. *We refer to each list of a compositional V-L pair and its associated nested subsequences as one data instance in our dataset*.

To diversify spatial relationships in our dataset beyond VG, we incorporate annotations from VSR (Liu, Emerson, and Collier 2023). Given VSR’s annotations in the format of <*subject, relation, object*> predicates, we use GroundingDINO (Liu et al. 2023) to obtain bounding boxes for all the involved visual entities, adding 1.2K data instances to our dataset. As LVLMs are often trained to perform various tasks, we also include 8K and 22K data instances randomly sampled from VG-VQA and LLaVA-Instruct150K (Liu et al. 2024), respectively. Our combined dataset, multi-granular Compositional Vision-Language (CompoVL), contains a total of 60.3K instances.

To assess the limitations of state-of-the-art LVLMs on compositional visual grounding, we also provide a CompoVL subset, named *CompoVL-hard,* containing only level-two and higher V-L pairs. This comprises 6K image-expression pairs in total. Given these pairs, LVLMs are required to output bounding boxes indicating the location of visual regions corresponding to the given expressions. Compared to RefCOCOg (Mao et al. 2016), our dataset has a higher average object count per linguistic expression (**2.70** vs. 2.29) and a greater average complexity level per expression (**2.56** vs. 2.26), offering more challenging compositional scenarios. We will later empirically demonstrate current models’ limitations on this subset (See Sec.).

Annotation’s reliability: To ensure data quality, we randomly sampled 2% of CompoVL-hard for three independent human evaluations. Evaluators are asked to answer questions about the generated text descriptions in terms of: naturalness, ambiguity of visual answers and bounding box accuracy. The questions asked include: “Does the generated image caption sound natural?”, “Does the caption refer to a unique object in the image? “, “Is the bounding box correct for the caption?”. According to assessments, 92.5% of the generated text descriptions sounds natural while 87.61% of them refer to unique visual objects, and 92.48% have correct bounding boxes. We also measured inter-annotator agreement using *Kappa scores (Cohen 1960)* and found an average score of *0.68* for question 1, *0.81* for question 2, and *0.76* for question 3. These scores indicate substantial or almost perfect agreement among annotators (Hallgren 2012). Details on the evaluation interface are in the Appendix.

Algorithm 1: Progressive Multi-granularity Decoding

Input: V: visual embeddings of image I; $E = \{E_c, E_{c-1}, \ldots, E_1\}$: series of nested subsequence expressions, c: complexity level, max_length: maximum output length
Output: y_c: generated spatial tokens y_c

```
for i = 1 to c do
    Initialize y_i ← []
    for j = 1 to max_length do
        y_{i,j} ← PromViL(V, E_i, (E_{i-1}, y_{i-1}), y_i)
        add y_{i,j} to y_i
    end for
end for
y_c ← y_i

return y_c
```

Training and Inference

Training Our PromViL framework takes as input an image I, represented by visual embeddings V and a series of nested subsequence expressions $E = \{E_1, ..., E_{c-1}, E_c\}$, with E_c is a complex compositional expression at level-c complexity ($c > 1$). Multiple expressions can exist at

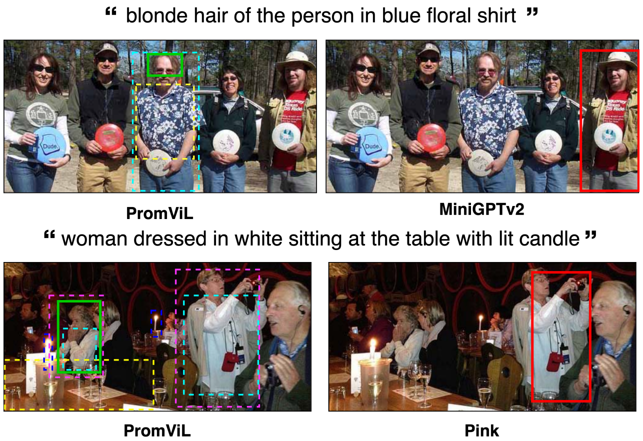

Figure 4: Qualitative results on CompoVL-hard. Solid green: correct boxes, solid red: incorrect. Existing methods struggle with complex descriptions or multiple similar objects (e.g., two ” women dressed in white”). PromViL leverages simpler expressions (dashed boxes) to accurately locate complex targets. More qualitative results in Appendix.

each level of complexity. We maintain the standard language modeling objective of predicting the next word token based on the preceding context. However, we explicitly train PromViL to visually ground, particularly through utilizing the nested subsequences in the CompoVL dataset. The model is trained to generate spatial tokens y_i referred to by text expression E_i of level-i complexity in consideration feedback received from lower levels of complexity. Indicating $< E_{i-1}, y_{i-1} >$ as responses from the previous level of complexity, our PromViL framework is trained by optimizing an averaged autoregressive next token prediction loss across expressions of all levels of complexity:

$$\mathcal{L} = \frac{1}{\mid E \mid}\sum_{i=1}^{c} \log P(y_i|V, E_i, < E_{i-1}, y_{i-1} >);$$

For data without nested subsequences (e.g., VQA, instruct-follow) we follow to the common practices employed in other LVLMs such as LlaVa (Liu et al. 2024) and Kosmos-1 (Huang et al. 2024), where y_i are language-based responses to a generic prompt.

In practice, our PromViL fine-tunes existing LVLMs on the CompoVL dataset with input representations for language expression E_i of level-i complexity as below: *<s><img> V </img> <grounding> We can see in the image: <p>* $\{E_{i-1}\}$ *</p>*$\{y_{i-1}\}$. *Based on that, we can locate: <p>*$\{E_i\}$ *</p>*$\{y_i\}$ *</s>.*

Inference During inference, we do not have access to nested subsequences of a given input expression E_c of level-c complexity. Therefore, we use a constituency parser (Kitaev, Cao, and Klein 2019) to extract nested subsequences from E_c by selecting noun phrase constituents from leaves toward the root (showed in Fig. 2). We then use a dependency parser to identify the semantic dependancies between entities within the subsequences, ultimately to remove level-one expressions that do not satisfy referential entities. This is to ensure our decomposed nested subsequences to have the same structure with our data generation process in CompoVL. Once these nested subsequences $E = \{E_c, E_{c-1}, ..., E_1\}$ are available, we progressively prompt the model where the generated response to a prior level provides a clue for the next level. Algorithm 1 describes PromViL's decoding process during inference.

Model	Params	Test Acc. (Ref-COCOg)	Acc. (CompoVL-hard)
Pink*	7B	83.70	62.03
MiniGPTv2*	7B	84.66	62.57
PromViL‡	7B	**85.32**	**68.92**
Kosmos-2	1.6B	61.65	55.37
F-Kosmos-2	1.6B	62.69	59.12
PromViL	1.6B	**65.28**	**64.07**

Table 1: Existing LVLMs excel in understanding simple concepts in RefCOCOg but struggle with complex expressions in CompoVL-hard. *F-Kosmos-2:* Kosmos-2 finetuned on the CompoVL excluding nested subsequences. PromViL‡: our framework finetuned with MiniGPTv2. (*) are supervised on region-caption data from VG.

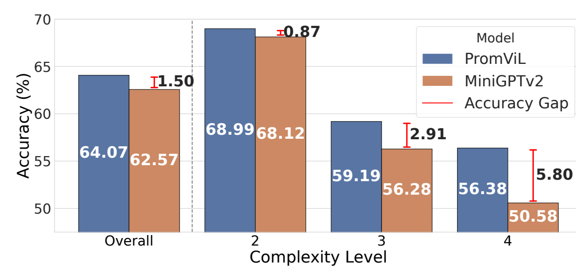

Figure 5: Accuracy comparison between PromViL and MiniGPTv2 on the CompoVL-hard dataset.

Experiments

Implementation Details

Unless otherwise stated, our PromViL is achieved by fine-tuning Kosmos-2 (Peng et al. 2024) on the CompoVL dataset, adhering to its default hyperparameters. We perform LoRA (Hu et al. 2021) tuning with r=64, learning rate 1e-4, warm-up ratio 0.1, and batch size 4. LoRA targets all linear layers in the model (85M parameters, ~4.9% of Kosmos-2's). Fine-tuning takes around 7 hours on a single NVIDIA V100 GPU. We use spaCy (Honnibal et al. 2020) for dependency parsing and Berkeley Neural Parser (Kitaev, Cao, and Klein 2019) for constituency parsing. We evaluate the effectiveness of PromViL on various visual grounding and

	Model	Params	RefCOCOg		RefCOCO+			RefCOCO		
			val	test	val	testA	testB	val	testA	testB
Supervised	Pink	7B	83.70	83.70	81.40	**87.50**	73.70	88.30	91.70	84.00
	MiniGPTv2	7B	84.44	84.66	79.97	85.12	74.45	88.69	91.65	85.33
	PromViL‡	7B	**85.91**	**85.32**	**81.65**	87.20	**75.62**	**89.92**	**92.76**	**86.77**
Zero-shot	CoVLM	1.4B	60.87	61.91	47.62	50.93	44.16	48.19	53.17	43.18
	CoVLM	2.8B	61.23	62.33	48.87	52.51	44.71	49.32	53.67	44.49
	Pink	7B	59.1	60.1	43.9	50.7	35.0	54.1	61.2	44.2
	Kosmos-2	1.6B	60.57	61.65	45.48	50.73	42.24	52.32	57.42	47.26
	PromViL	1.6B	**64.44**	**65.28**	**49.54**	**53.46**	**44.92**	**57.89**	**62.75**	**52.89**

Table 2: Comparison on visual grounding tasks. PromViL‡: our model finetuned with MiniGPTv2.

Model	GQA val							GQA-OOD		
	acc. ↑	bin. ↑	open ↑	consis. ↑	valid. ↑	plaus. ↑	dist. ↓	acc_all ↑	acc_tail ↑	acc_head ↑
Kosmos-8K	40.55	43.86	37.45	64.12	74.82	71.24	13.71	32.65	31.23	33.53
Kosmos-16K	41.44	44.70	38.38	66.19	75.14	71.61	**13.33**	33.83	32.36	34.74
PromViL	**45.07**	**52.15**	**38.44**	**68.40**	**83.47**	**79.80**	14.49	**37.84**	**33.77**	**40.33**

Table 3: Performance of PromViL in comparison with two finetuned versions of Kosmos-2 on the GQA validation split and the GQA-OOD dataset. Abbreviations for results on GQA: acc. - accuracy, bin. - binary, consis. - consistency, valid. - validity, plaus. - plausibility and dist. - distribution. Abbreviations for results on GQA-OOD: acc_all - overall accuracy, acc_tail - out-of-distribution accuracy, and acc_head - in-distribution accuracy.↑: the higher the better. ↓: the lower the better.

Model	Test Acc.	Val. Acc.
Kosmos-8K	47.24	46.82
Kosmos-16K	47.96	47.74
PromViL	**50.28**	**50.31**

Table 4: Performance of PromViL in comparison with other finetuned versions of Kosmos-2 on Visual7W.

downstream tasks, including zero-shot referring expression grounding tasks and VQA tasks.

Experimental Results

LVLMs on CompoVL-hard subset: To provide insights into the failure of existing LVLMs in compositional visual grounding, we experiment with various state-of-the-art models, including Pink (Xuan et al. 2024), MiniGPTv2 (Chen et al. 2023), Kosmos-2 (Peng et al. 2024), on our CompoVL-hard subset. We report *top-1 accuracy* of bounding box generation.

Tab. 1 and Fig. 4 show these models struggle on this subset both quantitatively and qualitatively. While these models could understand simple cross-entity relations (typically two-entity relations) in most situations, such as those in the RefCOCOg (Mao et al. 2016), they perform poorly on more complex compositional expressions in our CompoVL-hard. For example, Pink and MiniGPTv2 experience a drastic drop from in accuracy from over 80.0% to just over 62.0%. Smaller models like Kosmos-2 tend to face greater challenges in generalizing their understanding of primitive concepts to complex relational concepts.

We also finetune these models on our newly established CompoVL dataset to enhance their compositional visual grounding capabilities (Tab. 1, F-Kosmos-2). Results indicate that even with exposure to training data containing complex compositional expressions, the improvement is marginal. Kosmos-2 achieves only 59.12% after fine-tuning compared to 55.37% previously. PromViL benefits from both having access to complex expressions in CompoVL and advancements in modeling, reaching 64.07% accuracy.

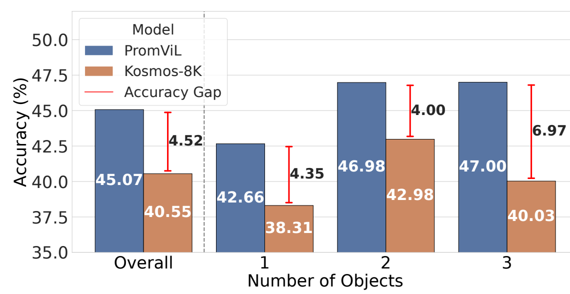

Figure 6: Accuracy comparison between PromViL and Kosmos-8K on the GQA_val dataset.

Referring expression tasks: To validate our proposed method PromViL's generalization capability on visual grounding task, we conduct evaluations on well-established benchmarks: RefCOCOg (Mao et al. 2016) and RefCOCO/RefCOCO+ (Kazemzadeh et al. 2014). For zero-shot

Model	CompoVL -hard	RefCOCOg	
		val	test
All levels included	**64.07**	**64.44**	**65.28**
Intermediate levels removed	61.22	63.48	63.44
Only highest level included	59.12	61.25	62.69
Only simplest level included	53.26	59.63	59.92

Table 5: Ablation studies on the RefCOCOg dataset and CompoVL-hard subset.

settings, our PromViL is based on Kosmos-2 while for supervised settings, we finetune MiniGPT2.

PromViL consistently outperforms all recent LVLMs across all datasets in both zero-shot and supervised settings (See Tab. 2). In zero-shot settings, it improves upon Kosmos-2 by nearly 4.0 points and over 5.5 points on average for RefCOCOg and RefCOCO, respectively. Among these datasets, RefCOCO contains a significantly larger number of same-type objects (Yu et al. 2016), thus benefiting the most from the PromViL's multi-granularity V-L alignment. Compared to other LVLMs, PromViL surpasses all of them with a considerably smaller model size. Performance gaps are even more pronounced when comparing PromViL to other models of comparable sizes. Similar results are observed on the supervised settings. These results strongly support the necessity of constructing a *chain of reasoning cues* to effectively handle complex compositional expressions.

Zero-shot compositional VQA tasks: We also validate PromViL's ability to perform compositional VQA tasks on the GQA (Hudson and Manning 2019) and GQA-OOD (Kervadec et al. 2021) datasets. Additionally, we challenge PromViL with zero-shot grounded VQA tasks using the Visual7W (Zhu et al. 2016) dataset, requiring the model to select the correct bounding box from four options based on a given question. Refer to the Appendix for implementation details and prompts for all models on these tasks.

We compare against two finetuned versions of the original Kosmos-2: (1) *Kosmos-8K*: finetuned on the CompoVL dataset excluding all nested subsequences (2) *Kosmos-16K*: also finetuned on the CompoVL dataset excluding all nested subsequences, but it is added with additional 8K VQA pairs randomly sampled from the VQA data of the VG dataset.

Tab 3 details experimental results on the two chosen datasets. As shown, PromViL outperforms Kosmos-8K and Kosmos-16K with accuracy gaps of approximately 4.0 points and 3.5 points, respectively. More importantly, PromViL is considerably better than these two baselines in terms of validity and plausibility. This suggests that clues from lower levels help constrain the model's search space, ultimately leading to more relevant answers to input queries. Furthermore, the differences in results of Kosmos-8K and Kosmos-16K indicate that the increasing size of VQA training data by twice yields minimal improvements, suggesting inherent limitations in the modeling of Kosmos.

Regarding the GQA-OOD dataset, PromViL surpasses both Kosmos-8K and Kosmos-16K in both out-of-distribution (acc_tail) and in-distribution (acc_head) settings. This indicates that our PromViL model is more effective in reducing the reliance on spurious linguistic correlations between input queries and generated answers.

On zero-shot grounded VQA setting (See Tab. 4), we observe that while increasing the size of VQA samples in the training set does not bring much effect (Row 1 vs. Row 2), the benefits of using V-L grounding progressively from low-level complexity expressions to higher-level ones are substantial. This highlights the crucial role of improved visual grounding capabilities in solving these tasks.

Model Analysis and Ablation Studies

To quantify the the relations between performance improvements and the complexity levels of input expressions in our PromViL model, we analyze its performance on each group of input expressions with the same level of complexity (See Fig. 5). As shown, the performance gap between our PromViL and MiniGPTv2 widens as input expressions become more complex. This clearly reaffirms the benefits of leveraging reasoning cues at lower levels to pave a reasoning path toward solving more complex tasks.

Similar trends are observed in the GQA dataset evaluation (See Fig. 6). PromViL consistently improves as scene complexity increases, achieving 42.66%, 46.98%, and 47.0% accuracy when the number of objects in visual scenes grows from 1 to 3. In contrast, Kosmos-8K struggles with increasing object count, dropping from 42.98% accuracy with 2 objects to 40.03% with 3 objects.

Moreover, we conduct an extensive set of ablation studies on CompoVL-hard and RefCOCOg to assess the contribution of each model component to the overall accuracy. Our experiments include: 1) *All levels included*: a full model, which features the chain of reasoning at all levels, ranging from the simple to complex. 2) *Intermediate levels removed*: all intermediate-level subsequences are removed. The model is trained using only expressions of level-one and the highest level of complexity. 3) *Only highest level included*: only includes expressions having the highest level of complexity in a series. 4) *Only simplest level included*: only includes expressions at the lowest level of complexity. Tab. 5 confirms that relying on either fine-grained processes and coarse-grained processes are not sufficient, especially when facing complex compositional queries as in CompoVL.

Conclusion

This paper introduced PromViL, a hierarchical framework for enhancing LVLMs' compositional reasoning capabilities. By progressively aligning visual and textual information at multiple granularities, PromViL enables effective learning of complex compositional vision-language tasks. We also introduced CompoVL, a dataset containing nested compositional vision-language pairs with varying levels of complexity, which can be readily used by existing LVLMs to enable progressive reasoning ability. Experimental results demonstrated PromViL's superior performance on various benchmarks, advancing the state-of-the-art in grounded compositional visual reasoning.

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Bottou, L. 2014. From machine learning to machine reasoning: An essay. *Machine learning*.

Chen, G.; Shen, L.; Shao, R.; Deng, X.; and Nie, L. 2024. Lion: Empowering multimodal large language model with dual-level visual knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Chen, J.; Zhu, D.; Shen, X.; Li, X.; Liu, Z.; Zhang, P.; Krishnamoorthi, R.; Chandra, V.; Xiong, Y.; and Elhoseiny, M. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*.

Cresswell, M. J. 2016. *Logics and languages*. Routledge.

Dang, L. H.; Le, T. M.; Le, V.; and Tran, T. 2021. Hierarchical object-oriented spatio-temporal reasoning for video question answering. *arXiv preprint arXiv:2106.13432*.

Gao, Y.; Liu, J.; Xu, Z.; Zhang, J.; Li, K.; Ji, R.; and Shen, C. 2022. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *Advances in neural information processing systems*.

Hallgren, K. A. 2012. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*.

Honnibal, M.; Montani, I.; Van Landeghem, S.; and Boyd, A. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, S.; Dong, L.; Wang, W.; Hao, Y.; Singhal, S.; Ma, S.; Lv, T.; Cui, L.; Mohammed, O. K.; Patra, B.; et al. 2024. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36.

Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Hupkes, D.; Dankers, V.; Mul, M.; and Bruni, E. 2020. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*.

Janssen, T. M.; and Partee, B. H. 1997. Compositionality. In *Handbook of logic and language*. Elsevier.

Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Hanna, E. B.; Bressand, F.; et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

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)*.

Kervadec, C.; Antipov, G.; Baccouche, M.; and Wolf, C. 2021. Roses are red, violets are blue... but should vqa expect them to? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Kitaev, N.; Cao, S.; and Klein, D. 2019. Multilingual Constituency Parsing with Self-Attention and Pre-Training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*.

Le, T. M.; Le, V.; Venkatesh, S.; and Tran, T. 2020. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Li, J.; Chen, D.; Hong, Y.; Chen, Z.; Chen, P.; Shen, Y.; and Gan, C. 2023a. Covlm: Composing visual entities and relationships in large language models via communicative decoding. *arXiv preprint arXiv:2311.03354*.

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

Liu, F.; Emerson, G.; and Collier, N. 2023. Visual spatial reasoning. In *Transactions of the Association for Computational Linguistics*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *Advances in neural information processing systems*.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Ma, Z.; Hong, J.; Gul, M. O.; Gandhi, M.; Gao, I.; and Krishna, R. 2023. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Nivre, J. 2008. Algorithms for Deterministic Incremental Dependency Parsing. *Computational Linguistics*.

Peng, Z.; Wang, W.; Dong, L.; Hao, Y.; Huang, S.; Ma, S.; Ye, Q.; and Wei, F. 2024. Grounding multimodal large language models to the world. In *The Twelfth International Conference on Learning Representations*.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*.

Shao, S.; Li, Z.; Zhang, T.; Peng, C.; Yu, G.; Zhang, X.; Li, J.; and Sun, J. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*.

Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; and Yang, H. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International conference on machine learning*.

Xuan, S.; Guo, Q.; Yang, M.; and Zhang, S. 2024. Pink: Unveiling the power of referential comprehension for multi-modal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*.

Yuksekgonul, M.; Bianchi, F.; Kalluri, P.; Jurafsky, D.; and Zou, J. 2023. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*.

Zeng, Y.; Zhang, X.; and Li, H. 2021. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*.

Zhao, T.; Zhang, T.; Zhu, M.; Shen, H.; Lee, K.; Lu, X.; and Yin, J. 2022. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*.

Zhu, Y.; Groth, O.; Bernstein, M.; and Fei-Fei, L. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.