arXiv:2403.18260v1 [cs.CV] 27 Mar 2024

Toward Interactive Regional Understanding in Vision-Large Language Models

Jungbeom Lee[1*] **Sanghyuk Chun**[2†] **Sangdoo Yun**[2†]
[1]Amazon [2]NAVER AI Lab
jungbeol@amazon.com, {sanghyuk.c, sangdoo.yun}@navercorp.com

Abstract

Recent Vision-Language Pre-training (VLP) models have demonstrated significant advancements. Nevertheless, these models heavily rely on image-text pairs that capture only coarse and global information of an image, leading to a limitation in their regional understanding ability. In this work, we introduce **RegionVLM**, equipped with explicit regional modeling capabilities, allowing them to understand user-indicated image regions. To achieve this, we design a simple yet innovative architecture, requiring no modifications to the model architecture or objective function. Additionally, we leverage a dataset that contains a novel source of information, namely Localized Narratives, which has been overlooked in previous VLP research. Our experiments demonstrate that our single generalist model not only achieves an interactive dialogue system but also exhibits superior performance on various zero-shot region understanding tasks, without compromising its ability for global image understanding.

1 Introduction

Vision-Language Pre-training (VLP) models (Radford et al., 2021; Li et al., 2022a, 2023b; Alayrac et al., 2022) have shown significant progress in recent years. A notable advancement is the emergence of zero-shot capabilities, which turn VLP models into generalist models, particularly when combined with large language models (LLMs). These models are now capable of solving various vision-language (VL) downstream tasks, including visual question answering (VQA) and image captioning, without the need for task-specific fine-tuning. The general knowledge enabling such zero-shot capabilities can be attained through training with massive image-text pair datasets (Schuhmann et al., 2021, 2022; Gadre et al., 2023).

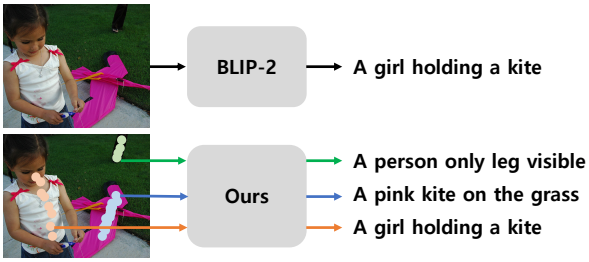

Figure 1: Conceptual comparison between BLIP-2 and our model. While BLIP-2 generates a single caption based on the entire image, our model can generate multiple captions corresponding to regions explicitly indicated by users.

However, VLP models still face a significant challenge: their limited ability to comprehend the fine-grained semantics of specific regions within an image. This stems from the nature of their training datasets. Existing image-text pairs, typically obtained by web crawling, tend to focus on the salient information of the image and fail to provide an explicit indication of the area of the image the text is describing. As a result, existing VLP models tend to focus on the implicit global information of the image, lacking the ability to understand the image region explicitly indicated by a user.

In this paper, we introduce RegionVLM, equipped with the **regional understanding** capability with explicit, or interactive, indications from users. We argue its significance for the following reasons. First, the regional understanding ability broadens the versatility of VLP models. This facilitates the execution of additional vision-language tasks that require explicit indications of regions, such as referring image segmentation (Yu et al., 2016) and visual commonsense reasoning (Zellers et al., 2019). Second, we can tackle inherent ambiguity (or multiplicity) in VL tasks (Gao et al., 2022; Chun et al., 2022; Chun, 2024; Chun et al., 2021) by employing regional understanding. An image can inherently be described by numerous text descriptions, *e.g.*, describing the visual attribute of

[*] Work done while doing visiting researcher at NAVER
† Corresponding authors

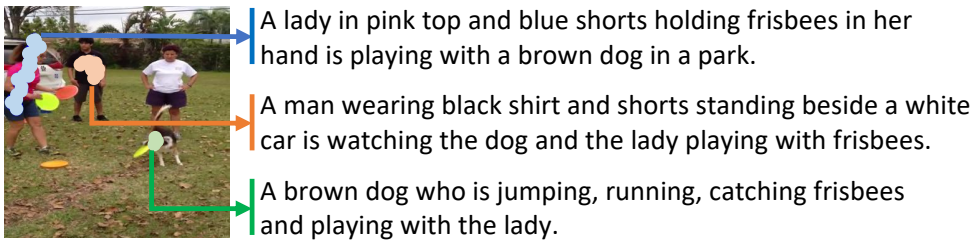

Figure 2: Examples of trajectories and their corresponding captions provided by the Localized Narratives dataset.

the salient object, explaining the background, etc. However, current VLP models are trained to associate an image with a single given caption rather than encompassing all possible explanations. By empowering a model to focus on specific regions with explicit indications, we can better manage this ambiguity. Third, integrating regional understanding enhances the interactivity between the model and users. As demonstrated by commercial services such as GPT-4V (OpenAI, 2023; Yang et al., 2023), enabling users to specify regions of interest in an image can lead to more precise and relevant interactions.

There have been several attempts to make VLP models endow region understanding ability. One possible direction is to develop a specialized model based on CLIP (Radford et al., 2021) for segmentation and detection by directly using fully supervised labels for each task (Zhou et al., 2023b; Gu et al., 2021; Yun et al., 2023; Li et al., 2023a). Several recent studies aim to develop a generalist model with the capability of region understanding by leveraging datasets containing image regions and their corresponding captions (Zhang et al., 2023; Wang et al., 2023b; Jin et al., 2023; Zhou et al., 2023a). However, the datasets used for these methods exhibit inherent drawbacks. For example, the visual grounding dataset (Yu et al., 2016) contains region-text pairs only for a limited set of object classes; the captions of Visual Genome (Krishna et al., 2017) are relatively short and only depict a limited relationship between objects. To achieve generality and scalability in VLP models, we need a dataset containing diverse regions with various open-world objects as well as expressive captions.

In the paper, we propose to exploit regional textual information from diverse narratives of images. Specifically, we utilize the Localized Narratives (LN) datasets (Pont-Tuset et al., 2020; Voigtlaender et al., 2023), which provide narrative descriptions by annotators and their mouse trajectory over the described region. The LN dataset includes expressive free-form captions depicting multiple open-world objects in a single image (see Figure 2), and thus, it can provide general and meaningful regional information to the VLP model. As shown in Figure 1, unlike BLIP-2 which can generate captions only for the entire image, our model can generate captions for multiple regions of the image through explicit indications.

We introduce a simple technique allowing a model to accept the regional information of LN. More specifically, we directly convert the 2D coordinates of the trajectory points to the sequence of strings (*e.g.*, "[", "19", "44", "]", "[", "23", "55", "]", as shown in Figure 4) and simply use them as the input of VLP models. Finally, our model is trained to generate a caption corresponding to image regions associated with each trajectory, resulting in an ability to understand regional information. Our approach does not require architectural modifications or redefinition of the objective function, ensuring seamless alignment with the original scheme that takes the entire image and text as input.

Our RegionVLM can incorporate various appealing aspects into the existing model while preserving its original capabilities. Our experiments demonstrate that our generalist model can achieve the interactive dialogue system by understanding the explicit region indication from a user. In addition, we show that our model can perform various zero-shot regional understanding tasks that were beyond the capability of the conventional BLIP-2. Furthermore, our model achieves better performance than the recent state-of-the-art methods.

2 Related Works

2.1 Vision-Language Pretraining

Vision-language pre-training (VLP) aims to learn meaningful multi-modal representations, enabling zero-shot ability and few-shot adaptation for various VL tasks. CLIP (Radford et al., 2021) and its variants (Li et al., 2021b; Mu et al., 2022;

Geng et al., 2023) align the vision and language representations obtained from independent vision and language encoders. The unified architecture, which learns multi-modal joint representation, is also popularly adopted (Li et al., 2022a, 2021a; Chen et al., 2020; Wang et al., 2023c; Kim et al., 2021) and shows powerful performance on various vision-language tasks. Recently, the attempts to inject visual information into large language models (LLMs) have been proposed (Koh et al., 2023; Tsimpoukelli et al., 2021; Li et al., 2023b; Alayrac et al., 2022). They can fully exploit the generality power of LLMs so that they have zero-shot, few-shot adaptation, and in-context learning abilities. All the models mentioned above are trained only on image-text pairs, so they primarily concentrate on global image information, with a limited understanding of the local regions of the image.

2.2 Region Modeling for VLP

To equip VLP models with region-specific information, a dataset explicitly matching image regions to their corresponding texts is essential. However, due to the lack of publicly available datasets and the high costs of creating such datasets, researchers often rely on various forms of supervision, though these methods have their limitations. For example, datasets which provide object bounding boxes or masks annotated with their class names, such as MS-COCO (Lin et al., 2014; Chen et al., 2015) and OpenImages (Kuznetsova et al., 2020), have been widely utilized (Li et al., 2022b; Wang et al., 2023b; Zang et al., 2023; Zhong et al., 2022; Zhang et al., 2022). However, the text descriptions in the datasets are short and simple object class names, which have limited ability to capture the relationships between objects in an image. Visual Genome (Krishna et al., 2017) provides dense captions of various objects and attributes in an image. Still, its captions are relatively short and simple, falling short in modeling the complex inter-object relationships. The visual grounding datasets, such as RefCOCO (Yu et al., 2016) or visual common-sense reasoning (VCR) dataset (Zellers et al., 2019), have also been utilized (Lai et al., 2023; Yao et al., 2022; Zhang et al., 2023). However, their region-text pairs still provide limited contexts (*e.g.*, 80 class categories for RefCOCO, and person-centric categories for VCR). We utilize the Localized Narratives dataset (Pont-Tuset et al., 2020; Voigtlaender et al., 2023), a comprehensive large-scale dataset that includes expressive captions corresponding to various regions associated with open-world objects.

Image-level prompting is another line of research for enabling regional understanding of VLP models. Wang et al. (2023a) crop the target image region and feed the cropped image into the model. Shtedritski et al. (2023) propose drawing a red circle around the target object in the image, which can direct the model's attention to a specific region. These methods can provide regional information to VLP models. However, they involve manual manipulation of the original image, potentially contaminating or eliminating the crucial context around the target object. In contrast, our method does not interrupt the image, preserving all contexts.

3 Proposed Method

In this section, we describe our training dataset and the proposed model, which addresses the limitations of previous methods introduced in Section 2. We first revisit our base model, BLIP-2 (Li et al., 2023b) in Section 3.1. We then introduce our dataset and model in Sections 3.2 and 3.3, respectively. Finally, we present how our method performs various VL downstream tasks in Section 3.4.

3.1 Revisiting BLIP-2

In this paper, we use BLIP-2 (Li et al., 2023b) as our base model due to its training efficiency and versatility. BLIP-2 aims to bridge a frozen pre-trained visual encoder and a frozen large language model (LLM) through a Q-former module, which effectively allows the LLM to comprehend images while maintaining its overall versatility. Given the input image, the frozen pre-trained visual encoder produces the image feature I. The Q-former module introduces N learnable input queries Z. These input queries are subsequently updated by interacting with each other through self-attention layers and interacting with the image features I through cross-attention layers. After a linear projection, output embedding $\hat{Z} \in \mathbb{R}^{N \times d}$ is obtained: $\hat{Z} = \texttt{Linear}(\texttt{Q-Former}(Z; I))$, where d is the dimension of the text embedding of the LLM. Given $\hat{Z}$ to the LLM, the Q-former is trained so that the frozen LLM generates the caption of the given image through language modeling loss. The model is trained with image-text pair datasets such as MS-COCO and LAION.

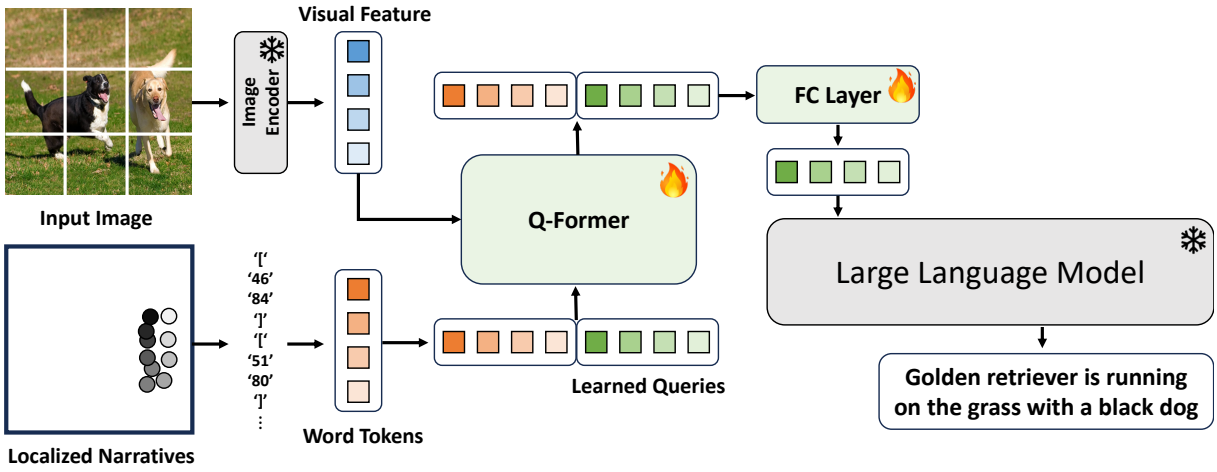

Figure 3: Overall architecture of our proposed model. Our model converts a set of trajectory points from Localized Narratives into word tokens. The word tokens and visual features are passed to the Q-Former, generating a soft prompt. This allows the frozen LLM to generate captions corresponding to the indicated regions.

3.2 Dataset Construction

To achieve a generalist model that has zero-shot capabilities with region understanding ability, it is essential to have a dataset containing diverse regions indicating various open-world objects and expressive captions. We explore a new dataset in terms of VLP, the Localized Narratives dataset (Pont-Tuset et al., 2020; Voigtlaender et al., 2023). This dataset includes images accompanied by narrative descriptions from annotators, along with their mouse trajectories over the corresponding regions. For a single image, an annotator describes the various situations and relationships among objects in the image using several sentences (see Figure 2). Therefore, we can split the given caption into multiple sentences based on periods (.) and commas (,) and associate each sentence with the corresponding trajectory points. From now on, in this context, we will refer to the trajectory points as *scribbles*. The Localized Narratives dataset is large-scale, contains expressive free-form captions depicting open-world objects, and provides multiple scribble-caption pairs for a single image. These properties enable the model to learn general region-aware multi-modal representation.

3.3 Grounding Image Regions to LLM

We propose a simple yet intuitive technique to convey regional information obtained from Localized Narratives to the frozen LLM. The overall architecture of our model is presented in Figure 3. Suppose an image with multiple scribbles $\{S\}$ and their associated captions $\{T\}$. We randomly choose one of the scribble-caption pairs, namely $\{S_i, T_i\}$. From S_i, we randomly sample K points, which can be represented as a 2-dimensional list $P = [[x_1, y_1], [x_2, y_2], \cdots, [x_K, y_K]]$, where x and y indicate relative positions of the image (*i.e.*, $0 \leq x, y \leq 1$).

To inject the regional information into the model, we convert P into text. However, directly using the 2-dimensional list can result in unnecessarily long input tokens. We introduce two tricks to simplify the text string. First, we removed unnecessary redundant word tokens, including those for the outermost brackets and intermediate commas. Second, we multiplied each coordinate by 100 and rounded them to ensure they have integer values. This allows us to omit the repeated "0." string. For example, when $K = 2$, $P = [[0.324, 0.643], [0.369, 0.622]]$ is converted to a string "`[32 64] [37 62]`". We then tokenize the string by using a tokenizer, resulting in a set of word tokens $W \in \mathbb{R}^{L \times d}$, where L is the length of word tokens in W.

In the Q-former module, the learnable input queries Z are concatenated with the word tokens W. This enables the queries Z to engage in cross-attention mechanisms with the visual feature I, while being conditioned by W through self-attention layers. As a result, the Q-former produces the output query embeddings $\hat{Z}$ and the output word embeddings $\hat{W}$, where $[\hat{Z}, \hat{W}] = \texttt{Linear}(\texttt{Q-Former}([Z, W]; I)) \in \mathbb{R}^{(N+L) \times d}$. We expect that the output query embeddings $\hat{Z}$ contain the semantics corresponding to the regions indicated by W, as the attention mechanism with W enables Z to direct its focus toward the regions.

We provide an empirical analysis showing that

Figure 4: Examples of cross-attention maps between learnable queries Z and image features I by varying the W for a single image. The examples demonstrate that the queries successfully attend to the regions indicated by W, as denoted by yellow stars.

the text-form input W can effectively guide the Q-former queries Z to focus on the regions indicated by W. We investigate which image regions the queries attend to by analyzing the attention scores of a cross-attention layer between queries and image features. We visualize the cross-attention maps between Z and I by varying W for a single image in Figure 4. It demonstrates that the queries appropriately attend to the regions that the text prompts W actually indicate, which are noted by yellow stars.

We now provide the output query embeddings $\hat{Z}$ obtained from the Q-former to the frozen LLM. Note that we exclusively use $\hat{Z}$ while excluding $\hat{W}$ to maintain the original scheme of BLIP-2. We train the Q-Former using the language modeling loss so that the frozen LLM generates the text T_i corresponding to the region associated with S_i. However, focusing mainly on modeling the local region may lead to a loss of global image understanding ability. Therefore, half of the training mini-batch samples are sampled from the Localized Narrative dataset, while the remaining half is sourced from global image-text pairs originating from the existing dataset (*e.g.*, LAION), following the standard training procedure of BLIP-2. Note that we set S as an empty string for the global image-text pairs, denoted by "".

3.4 Downstream Vision-Language Tasks

Our generalist model can be utilized in a range of VL downstream tasks that demand either global or local image comprehension abilities, or both.

Visual Question Answering (VQA): It requires the ability for global image understanding. Given an image, the Q-former generates the output query embeddings $\hat{Z}$ using S="", ensuring that $\hat{Z}$ contains the global image information. We attach the word tokens of a question text prompt ("Question: {} Answer:") following $\hat{Z}$ as input to the LLM. We follow BLIP-2 (Li et al., 2023b) for the answer generation process.

Referring Image Segmentation (RIS): It aims to segment the object based on the provided language description. Zero-shot RIS can directly showcase the model's region modeling ability. To implement zero-shot RIS, we first obtain several object proposals $\{M_i\}$ using the Segment Anything Model (SAM) (Kirillov et al., 2023). Among these proposal masks, our goal is to select a mask whose caption, generated by our model, is mostly similar to the given language description Y. We generate K random points inside each M_i and convert these points into the text prompt W_i. The W_i is then fed into our trained model along with the image features I, resulting in the likelihood of the generated text y_i. We compare each y_i with Y and select the final output mask M_{i^*}, where $i^* = \arg\min_i \texttt{dist}(y_i, Y)$, and $\texttt{dist}$ is a distance metric. Since we have no access to supervision for RIS, there is little chance that the generated captions closely resemble Y as provided by the RIS datasets. Therefore, we define $\texttt{dist}$ as the language modeling loss. We can expect improved performance by exploring more advanced distance metrics, but this is beyond our current scope.

Visual Commonsense Reasoning (VCR): Given a set of object masks in the image, VCR aims to answer the questions related to those objects

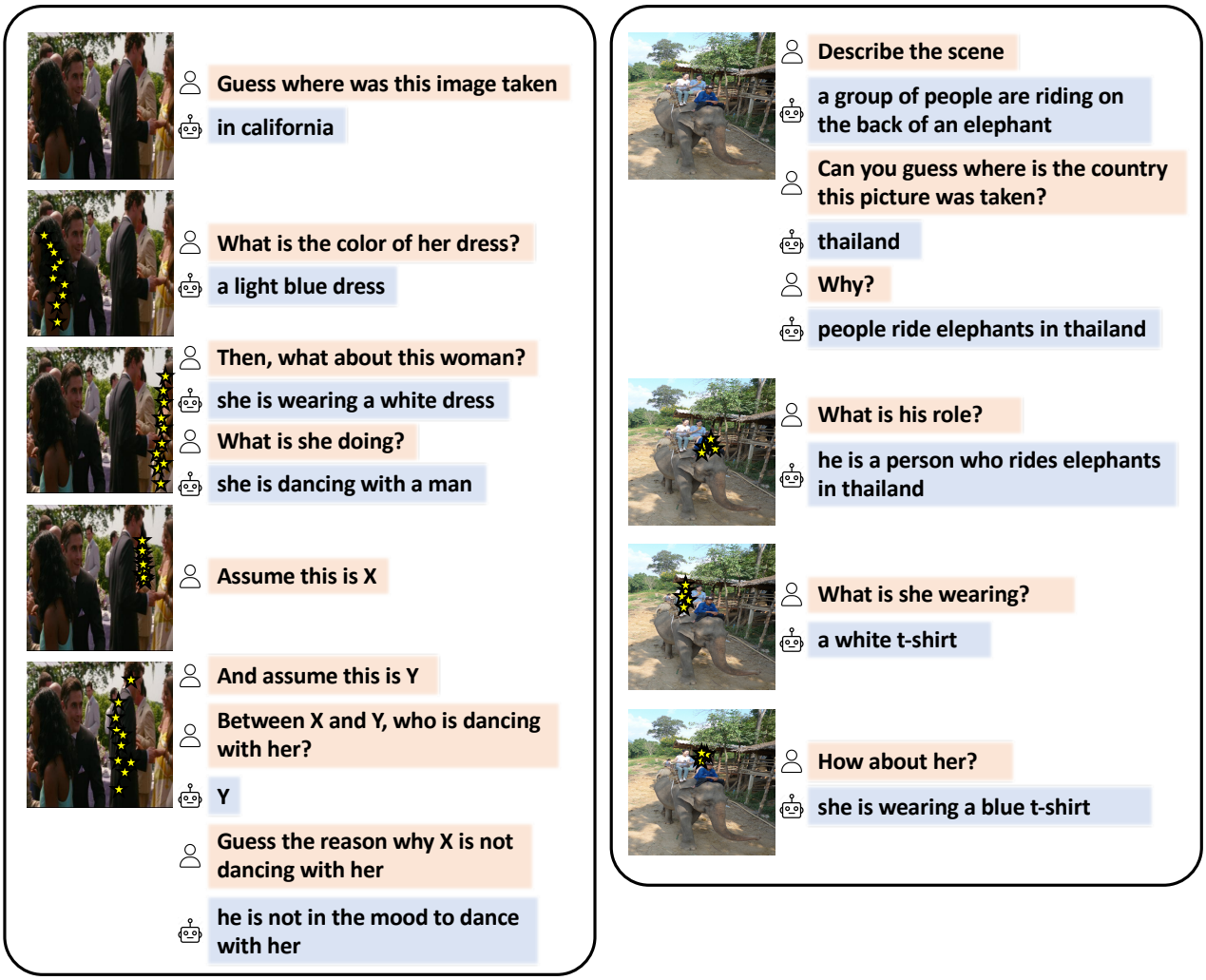

Figure 5: Selected examples of interactive dialogue using our model. The regions indicated by a user are noted as yellow stars. The examples illustrate a wide range of abilities for interacting with users, reasoning, guessing, question answering, etc. Note that the series of dialogues in one column is obtained from a single process.

(Figure 7). Moreover, beyond merely answering about the question, the model is required to provide a justification of the generated answer. Therefore, VCR demands advanced cognition and commonsense reasoning abilities, as well as both global and local image understanding abilities. Since VCR requires the information for multiple objects, we implement VCR as follows: for each given object mask M_i, we generate K random points inside it and obtain $\hat{Z}_i$ through the Q-Former. We then create a prompt to let the LLM understand which object index is associated with each query embedding, and we append the question to this prompt. The resulting example input to the LLM is as follows: "[0]: $\hat{Z}_0$ [1]: $\hat{Z}_1$, What is [0] here to do? 1. [1] is here to steal gold from [0]. 2. $\cdots$ 3. $\cdots$ 4. $\cdots$". Our model will respond with one of choices 1 through 4, which it considers the most appropriate.

4 Experiments

4.1 Experimental Setup

Our base model is pre-trained BLIP-2 (Li et al., 2023b) equipped with FlanT5$_{\text{XL}}$ (Chung et al., 2022). For a visual encoder, we use ViT-g/14 from EVA-CLIP (Fang et al., 2023). We finetune the Q-former and the linear layer of BLIP-2 for 10 epochs with a learning rate of 5×10^{-6} and a batch size of 64. We set K to 10 and N to 32. We follow the configuration of BLIP-2 for other settings regarding optimization. For experiments, we use 8 NVIDIA Tesla V100 (32GB) GPUs. For the training dataset, we use a mixture of Localized Narratives datasets built upon images (Pont-Tuset et al., 2020) and videos (Voigtlaender et al., 2023). Additionally, we utilize the Visual Genome dataset (Krishna et al., 2017) that provides bounding boxes

Table 1: Comparison with recent state-of-the-art weakly supervised and zero-shot methods for referring image segmentation on three benchmarks. Our results are obtained by a single experiment run.

Method	RefCOCO *val*	RefCOCO *testA*	RefCOCO *testB*	RefCOCO+ *val*	RefCOCO+ *testA*	RefCOCO+ *testB*	GRef *val*
Supervision: Weakly Supervised							
TSEG (Strudel et al., 2022)	25.44	-	-	22.01	-	-	22.05
Liu et al. (2023a)	31.17	32.43	29.56	30.90	30.42	30.80	36.00
Kim et al. (2023)	34.76	34.58	35.01	28.48	28.60	27.98	28.87
Lee et al. (2023)	31.06	32.30	30.11	31.28	32.11	30.13	32.88
Supervision: Zero-Shot							
Yu et al. (2023)	26.70	24.99	26.48	28.22	27.54	27.86	32.79
RegionVLM (Ours)	**38.74**	**39.40**	**37.59**	**31.47**	**33.99**	**30.22**	**33.94**

Figure 6: Selected examples of referring image segmentation on RefCOCO (*left*) and RefCOCO+ (*right*).

Table 2: Comparison with recent state-of-the-art methods for zero-shot visual commonsense reasoning. Our results are obtained by a single experiment run.

	Q→A	QA→R	Q→AR
Random	25.0	25.0	6.3
VL-T5 (Cho et al., 2021)	28.2	27.5	8.2
FewVLM (Jin et al., 2021)	27.0	26.1	7.4
GRILL (Jin et al., 2023)	40.6	39.3	16.2
UniFine (Sun et al., 2023)	**58.3**	51.3	-
RegionVLM (Ours)	52.4	**54.6**	**29.3**

Table 3: Comparison with BLIP-2 (Li et al., 2023b) on zero-shot visual question answering.

	OK-VQA	GQA	VQAv2
BLIP-2	41.08	**43.92**	63.12
RegionVLM (Ours)	**41.88**	43.50	**63.22**

along with their captions. To generate the set of points P for this dataset, we sample random K points inside the bounding box. For global image-text pairs, we use 115M images from the LAION-400M dataset (Schuhmann et al., 2021) filtered by Li et al. (2022a).

4.2 Experimental Results

Interactive Dialogue System: Our model can enable an interactive dialogue system with the generality power from the frozen LLM, realized by appending the previous chat history in front of the new query. If the image is provided, we append the Q-former queries computed from the image with user region indication in front of the text prompts. In Figure 5, we provide examples illustrating its capacity to comprehend the region indicated by the user (interactivity) and its ability for reasoning, guessing, and answering questions. It's worth noting that our method also retains the original BLIP-2's capability to process and understand the entire image.

Zero-shot RIS: Table 1 compares our method with recent state-of-the-art weakly supervised and zero-shot RIS methods on three benchmarks: RefCOCO (Yu et al., 2016), RefCOCO+ (Yu et al., 2016), and G-Ref (Mao et al., 2016). We adopt mean Intersection-over-Union (mIoU) as an evaluation metric. As shown in Table 1, our generalist method achieves significantly better performance than Yu et al. (2023), which is a current state-of-the-art specialized zero-shot RIS method. Additionally, our method demonstrates competitive performance compared to recent weakly supervised RIS methods (Strudel et al., 2022; Liu et al., 2023a; Kim et al., 2023; Lee et al., 2023), which have access

Figure 7: Selected examples of visual commonsense reasoning.

to image-text pairs for each target dataset. Figure 6 presents examples of RIS results obtained by our model.

Zero-shot VCR: Table 2 compares our method with recent state-of-the-art zero-shot visual commonsense reasoning on the VCR dataset (Zellers et al., 2019). The benchmark comprises three evaluation tasks: question answering (Q $\rightarrow$ A), rationale prediction given a question and answer pair (QA $\rightarrow$ R), and answer and rationale prediction given a question (Q $\rightarrow$ AR). We use accuracy as an evaluation metric. As shown in Table 2, our method demonstrates comparative performance compared to the recent zero-shot VCR methods. It outperforms GRILL (Jin et al., 2023) by 11.8%p on Q $\rightarrow$ A, 15.3%p on QA $\rightarrow$ R, which suggests that our model exhibits particularly strong reasoning ability. UniFine (Sun et al., 2023) shows superior Q $\rightarrow$ A performance compared to ours, whereas our model outperforms UniFine in QA $\rightarrow$ R, also indicating our stronger reasoning abilities. Our method focuses on generative modeling with a frozen LLM, known for its strong reasoning ability. On the other hand, UniFine (Sun et al., 2023) utilizes frozen CLIP (Radford et al., 2021), Roberta (Liu et al., 2019), and OFA (Wang et al., 2022), which may exhibit weaker reasoning capabilities compared to recent strong LLMs. Additionally, our fine-grained modeling captures meaningful relationships between objects, thereby contributing to enhanced reasoning capabilities. Figure 7 presents examples of VCR results obtained by our model. Our model generally possesses a commonsense reasoning ability but tends to struggle with reasoning about open-world knowledge (*e.g.*, the relationship between Harry Potter and Dudley's mom, as shown in the right example).

Zero-shot VQA: We conduct a quantitative assessment for zero-shot VQA on OK-VQA (Marino et al., 2019), GQA (Hudson and Manning, 2019), and VQAv2 (Goyal et al., 2017) benchmarks. Table 3 demonstrates that our method achieves comparable performance with BLIP-2, suggesting that our approach preserves the global image understanding ability.

Table 4: Comparison of BLIP-2 (Li et al., 2023b) combined with various region modeling methods for zero-shot referring image segmentation on the RefCOCO, RefCOCO+, and G-Ref validation sets.

	RefCOCO	RefCOCO+	GRef
Shtedritski et al. (2023)	14.85	15.76	15.86
Wang et al. (2023a)	27.91	31.45	31.38
RegionVLM (Ours)	**38.74**	**31.47**	**33.94**

Table 5: Comparison with BLIP-2 (Li et al., 2023b) for zero-shot image captioning on the NoCaps dataset.

	BLEU@4	SPICE	CIDEr
BLIP-2	43.4	14.0	105.8
RegionVLM (Ours)	**47.7**	**15.5**	**119.1**

Zero-shot Captioning: Table 5 presents the zero-shot captioning performance on the NoCaps (Agrawal et al., 2019) benchmark. Compared to BLIP-2, our model achieved improved performance across all three evaluation metrics: BLEU@4, SPICE, and CIDEr. We believe that our regional modeling contributes to the model's ability to capture fine-grained information, resulting in descriptive and detailed captions.

Comparison with other region modeling methods: We compare our method with two recent techniques that can inject regional information into BLIP-2 on RIS. We used the same evaluation settings, including the mask proposals from SAM (Kirillov et al., 2023) and the matching process between the generated captions and give descriptions. For Shtedritski et al. (2023), the image with a red circle drawn on each proposal area was inserted into BLIP-2 to generate a caption. For Wang et al. (2023a), the cropped box corresponding to each proposal is resized to the original image size, and BLIP-2 generates the caption based it. Table 4 demonstrates that our method achieves significantly better performance compared to those

Table 6: Robustness of our model against the noisy input scribbles.

Dilation	0	3	7	15
mIoU	37.73	37.64	36.39	35.77

two methods, and comparable performance with Wang et al. (2023a) on RefCOCO+. The language descriptions from RefCOCO+ tend to depict the object itself with less focus on its surrounding contexts or location. However, RefCOCO depicts the surrounding context of the object such as its location, so global information should be considered together (see Figure 6). Wang et al. (2023a) inject only region-of-interest into the model by cropping the proposal region, thereby losing the proposal's contexts and location. Our method can consider both local and global information for the proposal, yielding satisfactory results across all benchmarks for RIS.

4.3 Discussion

Robustness against the noisy input scribbles: Our interactive system expects user scribble inputs. However, in practice, scribbles obtained from users can be fall outside the intended object. We argue that our model is robust against noisy user input because the scribbles in the Localized Narratives dataset are collected through the free-form mouse movements of human annotators, which are inherently noisy. We support this argument with an additional quantitative analysis on RIS. As described in Section 3.4, the object proposal masks are obtained by SAM (Kirillov et al., 2023). Instead of utilizing the SAM-generated mask directly, we introduce some noise to simulate a noisy scribble environment. More precisely, we enlarge each SAM-generated mask by varying dilation ratios. This simulation represents a scenario in which a user provides a coarse mouse scribble that is not perfectly aligned with the target object but could contain the outside of the object. Table 6 demonstrates the robustness of our model against the noisy input scribbles.

Sensitivity to K: We analyze the sensitivity of the RIS performance to the value of K. Table 7 presents the RIS performance, varying K at test time, using the model trained with $K = 10$. This demonstrates that our model operates successfully even when the number of provided points differs between training and testing.

Table 7: Comparison of referring image segmentation performance by varying K.

	RefCOCO	RefCOCO+	GRef
$K = 5$	37.56	30.95	32.75
$K = 10$	**38.74**	**31.47**	**33.94**
$K = 15$	37.98	31.14	33.11

Table 8: Effectiveness of point representation design using Localized Narratives (LN) on zero-shot VQA.

	OK-VQA	GQA	VQAv2
LN w/o points	39.56	42.90	61.74
LN w/ points	**41.88**	**43.50**	**63.22**

Leveraging Localized Narratives without Point Representation: We can study the effectiveness of our proposed method by training the model trained with LN without the point representation. However, without the point representation, the model cannot perform tasks that require explicit region indication. Therefore, we present zero-shot performance on VQA for the model fine-tuned by using LN + LAION without the point representation. Table 8 shows that the point representation brings better VQA performance. We believe this is because LN contains less diverse and descriptive captions compared to the existing image-text pair datasets. However, our proposed model can preserve global understanding ability by separating the learning of global understanding and local understanding through the point representation.

5 Conclusion

In this study, we have addressed the limited region understanding ability of existing vision-language pre-training models. We proposed a model that can input the indication of the region, which isseamlessly integrated into the existing model. In addition, we utilized Localized Narratives to learn the general knowledge of image regions. Our experiments showcase the superior performance of our generalist model across a diverse set of zero-shot region understanding tasks, without compromising its ability for global image comprehension tasks. As a generalist model, we foresee significant potential for further enhancement through instruction tuning (Liu et al., 2023b; Dai et al., 2023), establishing a promising direction for future research.

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee. 2019a. What is wrong with scene text recognition model comparisons? dataset and model analysis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4715–4723.

Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. 2019b. Character region awareness for text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9365–9374.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR.

Sanghyuk Chun. 2024. Improved probabilistic image-text representations. In *International Conference on Learning Representations (ICLR)*.

Sanghyuk Chun, Wonjae Kim, Song Park, Minsuk Chang Chang, and Seong Joon Oh. 2022. ECCV Caption: Correcting false negatives by collecting machine-and-human-verified image-caption associations for MS-COCO. In *European Conference on Computer Vision (ECCV)*.

Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. 2021. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8415–8424.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

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

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. 2023. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*.

Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, Rongrong Ji, and Chunhua Shen. 2022. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *Advances in neural information processing systems*, 35:35959–35970.

Shijie Geng, Jianbo Yuan, Yu Tian, Yuxiao Chen, and Yongfeng Zhang. 2023. Hiclip: Contrastive language-image pretraining with hierarchy-aware attention. *arXiv preprint arXiv:2303.02995*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. 2021. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*.

Drew A Hudson and Christopher D Manning. 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*, pages 6700–6709.

Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. 2021. A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models. *arXiv preprint arXiv:2110.08484*.

Woojeong Jin, Subhabrata Mukherjee, Yu Cheng, Yelong Shen, Weizhu Chen, Ahmed Hassan Awadallah, Damien Jose, and Xiang Ren. 2023. Grill: Grounded vision-language pre-training via aligning text and image regions. *arXiv preprint arXiv:2305.14676*.

Dongwon Kim, Namyup Kim, Cuiling Lan, and Suha Kwak. 2023. Shatter and gather: Learning referring image segmentation with text supervision. *arXiv preprint arXiv:2308.15512*.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023. Grounding language models to images for multimodal inputs and outputs. *ICML*.

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

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

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2023. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*.

Jungbeom Lee, Jooyoung Choi, Jisoo Mok, and Sungroh Yoon. 2021a. Reducing information bottleneck for weakly supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 34:27408–27421.

Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. 2019. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5267–5276.

Jungbeom Lee, Eunji Kim, Jisoo Mok, and Sungroh Yoon. 2022a. Anti-adversarially manipulated attributions for weakly supervised semantic segmentation and object localization. *IEEE transactions on pattern analysis and machine intelligence*.

Jungbeom Lee, Eunji Kim, and Sungroh Yoon. 2021b. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4071–4080.

Jungbeom Lee, Sungjin Lee, Jinseok Nam, Seunghak Yu, Jaeyoung Do, and Tara Taghavi. 2023. Weakly supervised referring image segmentation with intra-chunk and inter-chunk consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21870–21881.

Jungbeom Lee, Seong Joon Oh, Sangdoo Yun, Junsuk Choe, Eunji Kim, and Sungroh Yoon. 2022b. Weakly supervised semantic segmentation using out-of-distribution data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16897–16906.

Jungbeom Lee, Jihun Yi, Chaehun Shin, and Sungroh Yoon. 2021c. Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2643–2652.

Jingyao Li, Pengguang Chen, Shengju Qian, and Jiaya Jia. 2023a. Tagclip: Improving discrimination ability of open-vocabulary semantic segmentation. *arXiv preprint arXiv:2304.07547*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021a. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022b. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. 2021b. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Fang Liu, Yuhao Liu, Yuqiu Kong, Ke Xu, Lihe Zhang, Baocai Yin, Gerhard Hancke, and Rynson Lau. 2023a. Referring image segmentation using text supervision. *arXiv preprint arXiv:2308.14575*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. 2022. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, pages 529–544. Springer.

OpenAI. 2023. Gpt-4v(ision) system card.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting vision and language with localized narratives. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 647–664. Springer.

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

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. 2023. What does clip know about a red circle? visual prompt engineering for vlms. *arXiv preprint arXiv:2304.06712*.

Robin Strudel, Ivan Laptev, and Cordelia Schmid. 2022. Weakly-supervised segmentation of referring expressions. *arXiv preprint arXiv:2205.04725*.

Rui Sun, Zhecan Wang, Haoxuan You, Noel Codella, Kai-Wei Chang, and Shih-Fu Chang. 2023. Unifine: A unified and fine-grained approach for zero-shot vision-language understanding. *arXiv preprint arXiv:2307.00862*.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

Paul Voigtlaender, Soravit Changpinyo, Jordi Pont-Tuset, Radu Soricut, and Vittorio Ferrari. 2023. Connecting vision and language with video localized narratives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2461–2471.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR.

Teng Wang, Jinrui Zhang, Junjie Fei, Yixiao Ge, Hao Zheng, Yunlong Tang, Zhe Li, Mingqi Gao, Shanshan Zhao, Ying Shan, et al. 2023a. Caption anything: Interactive image description with diverse multimodal controls. *arXiv preprint arXiv:2305.02677*.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. 2023b. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2023c. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19175–19186.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9:1.

Yuan Yao, Qianyu Chen, Ao Zhang, Wei Ji, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2022. Pevl: Position-enhanced pre-training and prompt tuning for vision-language models. *arXiv preprint arXiv:2205.11169*.

Seungryong Yoo, Eunji Kim, Dahuin Jung, Jungbeom Lee, and Sungroh Yoon. 2023. Improving visual prompt tuning for self-supervised vision transformers. In *International Conference on Machine Learning*, pages 40075–40092. PMLR.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 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*, pages 69–85. Springer.

Seonghoon Yu, Paul Hongsuck Seo, and Jeany Son. 2023. Zero-shot referring image segmentation with global-local context features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19456–19465.

Sukmin Yun, Seong Hyeon Park, Paul Hongsuck Seo, and Jinwoo Shin. 2023. Ifseg: Image-free semantic segmentation via vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2967–2977.

Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. 2023. Contextual object detection with multimodal large language models. *arXiv preprint arXiv:2305.18279*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731.

Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. 2022. Glipv2: Unifying localization and vision-language understanding. *Advances in Neural Information Processing Systems*, 35:36067–36080.

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

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. 2022. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803.

Qiang Zhou, Chaohui Yu, Shaofeng Zhang, Sitong Wu, Zhibing Wang, and Fan Wang. 2023a. Regionblip: A unified multi-modal pre-training framework for holistic and regional comprehension. *arXiv preprint arXiv:2308.02299*.

Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. 2023b. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11175–11185.

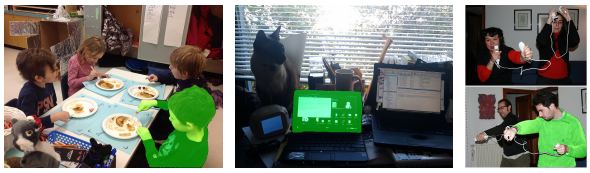

Figure 8: Examples of our failure cases on zero-shot referring image segmentation.

Table 9: Number of images and number of region-caption pairs for each dataset.

	# of images	# of region-caption pairs
Image LN	306K	445K
Video LN	125K	149K
Visual Genome	77K	1.7M
LAION	115M	115M

A Appendix

Dataset Details. As mentioned in Section 4.1, we use a combination of image Localized Narratives (Pont-Tuset et al., 2020), video Localized Narratives (Voigtlaender et al., 2023), Visual Genome (Krishna et al., 2017), and LAION-400M filtered by Li et al. (2023b). We present statistics of each dataset in Table 9.

Failure Case Analysis for RIS. As demonstrated in Table 1 and Figure 6, our method successfully performs zero-shot RIS, but may occasionally yield unsatisfactory results. Figure 8 shows some failure cases. Figure 8(*left*) shows that our model tends to struggle to recognize the character in the image. We believe that we can further utilize Optical Character Recognition (OCR) (Baek et al., 2019b,a; Kim et al., 2022) datasets, which provide characters in an image together with their location. Figure 8(*middle*) shows that our method correctly identifies the target object, but tends to focus on small partial regions of the target object. This limitation is also explored in recent weakly supervised segmentation studies (Lee et al., 2021a,b). We conjecture that the reason for this is that only small regions of the target object can provide sufficient information to generate captions that align with the given language descriptions. Figure 8(*right*) shows that our method produces lower accuracy of RIS although our method identified the referred object successfully. The language descriptions from RIS datasets tend to describe a person by using only a portion of the individual, such as "grey sweater" in the example. Therefore, our method successfully identifies the "grey sweater" only, but since the actual ground truth includes all regions of a man wearing the grey sweater, these cases impact the overall performance.

Limitations. In contrast to global image-text pair datasets, which can be automatically collected from the web, our dataset may have limitations in terms of scalability. To address this, we can generate pseudo region captions using our trained model, as BLIP-2 utilizes the pseudo captions produced by BLIP (Li et al., 2022a) captioning model. Additionally, our current evaluations focus mainly on zero-shot downstream tasks. It is also worth to explore the possibility of our method for transfer learning (Yoo et al., 2023), semi-supervised learning (Lee et al., 2019, 2022a,b), few-shot learning (Jin et al., 2023; Alayrac et al., 2022), and weakly supervised learning (Lee et al., 2023, 2021c).

Potential Risks. Since our model is based on frozen LLMs, it shares similar potential risks to LLMs, such as generating offensive output, vulnerability to attacks, and leaking personal sensitive data. To address this, we may need an additional filtering module to prevent such output from being conveyed to users.