Grounded Image Text Matching with Mismatched Relation Reasoning

Yu Wu[1,*] Yana Wei[1,*] Haozhe Wang[1] Yongfei Liu[1] Sibei Yang[1,2] Xuming He[1,2]
[1]ShanghaiTech University, [2]Shanghai Engineering Research Center of Intelligent Vision and Imaging
{wuyu1, weiyn1, yangsb, hexm}@shanghaitech.edu.cn jasper.whz@outlook.com liuyongfei314@gmail.com

Abstract

This paper introduces Grounded Image Text Matching with Mismatched Relation (GITM-MR), a novel visual-linguistic joint task that evaluates the relation understanding capabilities of transformer-based pre-trained models. GITM-MR requires a model to first determine if an expression describes an image, then localize referred objects or ground the mismatched parts of the text. We provide a benchmark for evaluating vision-language (VL) models on this task, with a focus on the challenging settings of limited training data and out-of-distribution sentence lengths. Our evaluation demonstrates that pre-trained VL models often lack data efficiency and length generalization ability. To address this, we propose the Relation-sensitive Correspondence Reasoning Network (RCRN), which incorporates relation-aware reasoning via bi-directional message propagation guided by language structure. Our RCRN can be interpreted as a modular program and delivers strong performance in terms of both length generalization and data efficiency. The code and data are available on https://github.com/SHTUPLUS/GITM-MR.

1. Introduction

Recently, transformer-based vision-language (VL) pre-trained models have made significant progress on various VL tasks by fine-tuning on downstream tasks [8, 64, 58, 29, 16]. Despite their success, the representation capacity of such VL pre-trained models remains poorly understood. As a result, an increasing number of studies start to probe the limitations of those learned models from several aspects. The first aspect focuses on the lack of fine-grained understanding of multi-modal data, which is indicated by the weak VL correspondence learned on relations among the entities [36]. In particular, the researchers constructed subtle relation phrase differences in text to reveal the poor matching capability between visual relations and linguistic relations [46, 21, 38], or required the model to output fine-grained predictions to thoroughly examine the model's understanding of VL correspondence on different uni-modal component representations [52]. Additionally, while large models generally deliver superior performance, they tend to overfit on small datasets, emphasizing the need to enhance the data-efficiency on fine-tuning [15]. Furthermore, when the text is composed of multiple relations, inferring VL correspondence typically involves combining those semantic relations. This combination requires the model to reason the global semantic correspondence, where the length generalization is shown to be a critical weakness in reasoning capability of pre-trained transformers [3, 54]. However, most of those studies only focus on a single aspect of the VL models and hence produce a partial view on their limitations.

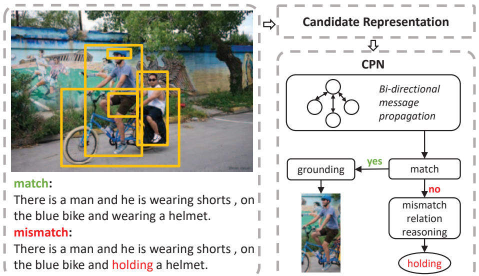

match:
There is a man and he is wearing shorts , on the blue bike and wearing a helmet.
mismatch:
There is a man and he is wearing shorts , on the blue bike and holding a helmet.
Candidate Representation
CPN
Bi-directional message propagation
grounding
yes
match
no
mismatch relation reasoning
holding

Figure 1. Illustration of the GITM-MR task and the proposed model RCRN. The task is to first identify whether the expression describes the image, and then localize the referred object if the pair are matched, or otherwise ground the mismatched language part. Both data efficiency and length generalization are required.

To better understand the VL pre-trained models, we introduce a novel VL joint task named *Grounded Image Text Matching with Mismatched Relation (GITM-MR)* which focuses on distinguishing nuance in relations within text and image, and requires a model to show explicit comprehension on relations in both modalities. As shown in Fig. 1, the task first requires the model to identify whether an expression describes a given image, and then grounds the referred

[*]Both authors contributed equally to this work, which was supported by Shanghai Science and Technology Program 21010502700, Shanghai Frontiers Science Center of Human-centered Artificial Intelligence and MoE Key Lab of Intelligent Perception and Human-Machine Collaboration (ShanghaiTech University).

object if the image-text pair are matched or localizes the mismatched relation in the expression otherwise. The two fine-grained sub-tasks verify the specific understanding on relations both in match and mismatch scenarios. Moreover, the proposed task also allows us to evaluate the data efficiency of model's fine-tuning and its generalization ability. In particular, we design a learning setting with limited annotations, where only the data in a limited size and domain have annotations (i.e., a mix of in- and out-of-distribution (OOD) test data). Given the task design, we then build a new benchmark based on the Ref-Reasoning [61] dataset, which is more general and challenging than previous probing benchmarks on relations and with low linguistic bias.

Leveraging our proposed benchmark, we evaluate several representative state-of-the-art models on the task, assessing their performance under limited data sizes and varying domains of input length. Our results reveal their limitations in the both settings. To alleviate these drawbacks, we propose a unified modular framework that jointly solves subtasks in the GITM-MR problem. By explicitly capturing relation correspondence and utilizing it to propagate messages, our approach is capable of obtaining a comprehensive understanding of global correspondence. Additionally, its lightweight and compositional module design allows for excellent data efficiency and length generalization, accompanied by a more transparent reasoning process.

Specifically, we develop a graph neural network on a parsed language scene graph (LSG), named *Relation-sensitive Correspondence Reasoning Network (RCRN)*, which exploits the language structure to perform relation-aware reasoning in a belief space of visual-linguistic alignment. Our RCRN uses a set of primitive reasoning modules to operate on the pre-trained vision-language features under the guidance of LSG, and predict the fine-level grounding in both match and mismatch scenarios. As demonstrated by the experiments, our method outperforms in both data efficiency and length generalization settings, demonstrating its effectiveness in learning relation correspondences with data efficiency and OOD generalization capability.

The contribution of our work is summarized as follows:

- We propose a new VL task GITM-MR that challenges relation correspondence learning in VL pre-trained model, and focusing on both data efficiency and out-of-distribution text length setting, with a new benchmark for GITM-MR.
- We identify the shortcomings of current VL models through the benchmark, and develop a modular graph network RCRN to address these issues.
- Our method achieves superior results on both settings, validating its relation learning efficiency and generalization ability, which hopefully inspires the following exploration of these challenges.

2. Related Works

Probing Tasks for VL Models Despite the rapid development on VL pre-trained models [8, 41, 19, 64, 58, 29], emergent works reveal the limitations and gain some deep understanding on pre-trained representations, including implausible scenes understanding [10], language and vision priors exploration [20], and modality ablation [18]. A considerable number of them have revealed the weakness of VL models in relation understanding. Early work proposed a benchmark FOIL [46, 45] by substituting a word in caption to test non-pretrained VL models on identifying minor differences. Subsequent works [21, 38, 36, 53, 63] further evaluate pre-trained models on detecting subtle differences in relations or predicate-noun dependencies in zero-shot scenarios, and reach a consensus on the poor quality of pre-trained VL representation for relations. FGVE [52] extends this evaluation to the task of fine-grained entailment with model fine-tuning, but lacks evaluation in vision modality. Recently, a new benchmark [42] is proposed to simultaneously assess the concept understanding of representations learned from both modalities. In contrast, our task focuses on generating fine-grained multi-modal output under different fine-tuning settings in order to systematically investigate the relation understanding problem in VL pre-training.

OOD Generalization The general problem of OOD generalization has attracted much attention recently [33, 6, 47, 30, 48]. Among them, several studies have considered the tasks with generalization to processing or generating longer sequences. Some works [3, 54] have studied length generalization in different uni-modal seq-to-seq tasks, and find that fine-tuning pre-trained transformer models can lead to poor out-of-distribution (OOD) performance. To address this issue, various solutions have been proposed, including modifications to attention mechanism [39, 11, 17] and the design of recurrent networks [44, 4], but their application in VL tasks is nontrivial. Other recent works have also revealed the challenge of OOD generalization for the VL tasks, but most of them focuses on the answer bias of Visual Question Answering (VQA) [50, 1, 2]. The attempted solutions include the causal learning [37] and stable learning [51] methods. Perhaps most related to us is the modular network approaches [28, 49, 65], which use program-based design to achieve good generalization for certain OOD scenario such as question domains in VQA. We follow the modular design but focus on the OOD problem induced by the expression complexity in the grounded image text matching task.

Vision-language Matching Image-Text Matching (ITM) and Referring Expression Grounding (REG) are common cross-modal matching tasks in vision-language domain. The problem of ITM aims to predict a global similarity between a text and an image. Classical methods for ITM fol-

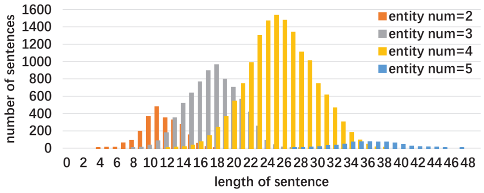


Figure 2. The length distribution of sentences containing different numbers of entities in the validation set. Different colors represent sentences with different number of entities.

low the core idea of learning a common space where linguistic and visual features are compatible [27, 31, 7, 14, 9]. Recent approach [41] concentrates on learning the relation-aware global similarity by a regularization strategy on training. Our model derives a global similarity measure by propagating messages on local correspondences, capable of capturing subtle changes in relations.

The goal of REG is to locate an object in an image based on a language expression. Most methods focus on learning a holistic visual and linguistic representation in a joint space for aligning the object and text, based on CNN+RNN architectures [35, 66, 25] or the Transformer [12, 13, 32, 40, 59]. Another line of research aims to exploit the linguistic structure in the text for better context modeling [62, 34, 60, 56]. In particular, SGMN [61] parses the text into language components and performs reasoning under the guidance of the linguistic structure. Our method introduces a flexible contextual representation and a bi-directional propagation to extend correspondence reasoning beyond the task of REG.

3. Problem Setup and Benchmark

We aim to study the representation capacity of the VL pretrained models for fine-grained downstream tasks. To this end, we introduce the *Grounded Image Text Matching with Mismatched Relation (GITM-MR)* task, which is a joint matching and grounding task focusing on relations in vision and language domain. In this section, we present the problem setting of GITM-MR, followed by building a new benchmark for the task.

GITM-MR Task Given an image and a referring expression, our goal is to determine if the expression describes an object in the image or not (termed as *Image-Text Matching (ITM)*), and then to localize the referred object if it is matched (termed as *Referring Expression Grounding (REG)*), or otherwise ground the mismatched relation phrase (termed as *Mismatched Relation Reasoning (MRR)*). Notably, the only difference between match and mismatch is a relation phrase in the expression, making our task particularly challenging.

Formally, given an image I and a text description L, we denote their matching state as y, indicating whether the image matches the description ($y = 1$) or not ($y = 0$). For the matched case, the referred object has a corresponding location in the image, represented by its bounding box $z \in \mathbb{R}^4$. For the mismatch case, we denote the set of all possible relation phrases (i.e. predicates) for mismatch reasoning as $\mathcal{R} = \{r_i\}_{i=1}^{N_e}$, and there is a target mismatched relation r_i^* in the relation phrase set $\mathcal{R}$. Given the input I and L, our goal is to first predict y for the input pair, and then predict z for the referred object in L when $y = 1$, or otherwise predict r_i^* from $\mathcal{R}$ to explain why $y = 0$.

For model training, we follow the convention of ITM and REG tasks, and assume the annotations for those two subtasks are provided. However, it is usually costly to collect annotations for the subtle MRR task. Consequently, we assume a weakly-supervised learning setting where only the matching state label y is given for mismatched data.

Our Benchmark To conduct systematical study, we construct a benchmark for the GITM-MR task, including a new dataset and two challenging evaluation settings. We first build a dataset of image-text pairs from the public REC dataset, Ref-Reasoning [61], by substituting the relation phrases to generate mismatched referring expressions for images. From the provided relation phrases in the LSG offered by the dataset, we manually select a subset of 27 commonly-occurred relations, and assign some contextual close but semantically different ones for each relation in the subset as their mismatched candidates. We control the linguistic bias by several measures, including keeping the relation distribution and checking by language-only model, the details of which are stated in the Suppl.

However, replacing the relation phrases may introduce falsely mismatched cases. To ensure the quality of our data, we filter out those false mismatch cases in the test set with the help of annotators. After filtering, the dataset contains more than 1M referring expressions in 60K images. It has 1M, 50K and 10K expression-referent pairs for training, validation, testing, respectively. We only select the original image text pairs which have generated mismatched pairs to keep the matching label balanced. Fig. 2 shows the length distribution of sentences in the validation set of the benchmark. We find it is more general and challenging than previous benchmarks on relation understanding, and refer the reader to see more details of the dataset in Suppl.

Evaluation Settings Given the dataset, we develop two evaluation settings to test pretrained VL models on their relation understanding. The first setting is designed to investigate data efficiency during the fine-tuning process. We train the models on small-scale datasets and tested on in-distribution test sets. The second setting is designed to evaluate the generalization performance of models with regards to different text input lengths. Here the test set contains

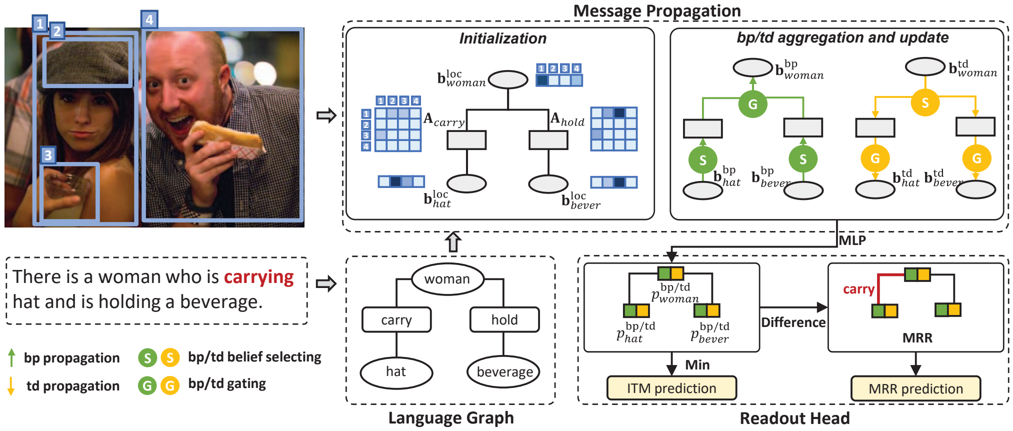

Message Propagation
Initialization
bp/td aggregation and update
$\mathbf{b}^{loc}_{woman}$
$\mathbf{A}_{carry}$
$\mathbf{A}_{hold}$
$\mathbf{b}^{loc}_{hat}$
$\mathbf{b}^{loc}_{bever}$
$\mathbf{b}^{bp}_{woman}$
$\mathbf{b}^{td}_{woman}$
$\mathbf{b}^{bp}_{hat}$ $\mathbf{b}^{bp}_{bever}$
$\mathbf{b}^{td}_{hat}$ $\mathbf{b}^{td}_{bever}$
1 2 3 4
There is a woman who is carrying hat and is holding a beverage.
woman
carry
hold
hat
beverage
Language Graph
bp propagation
td propagation
bp/td belief selecting
bp/td gating
MLP
$p^{bp/td}_{woman}$
$p^{bp/td}_{hat}$ $p^{bp/td}_{bever}$
Difference
carry
MRR
Min
ITM prediction
MRR prediction
Readout Head

Figure 3. Model overview with a mismatched case. Given an image and an expression, we first generate visual and linguistic candidates by a detector and a language parser, and compute their representations. Then we use the Context-sensitive Propagation Network to inter alignments between visual-linguistic candidates, which conducts bi-directional message propagation based on the language graph. The propagation initializes the messages by computing local beliefs, selectively aggregates the context information and updates the belief with a context-sensitive gating function. Predictions for this case are obtained by exploiting the beliefs from the propagation. In the language graph, the ellipses represent entity phrases and the rectangles stand for relation phrases. Further elaboration can be found in the main text.

longer and more complex samples on relation combinations compared to the training set.

Specifically, according to the sentence length distribution in Fig. 2, we construct two training sets, named Train-Len16 and Train-Len11, which includes sentences containing less than 16 and 11 words respectively. In the test phase, the model needs to perform the reasoning task on sentences with lengths less than 11 and 16 (in-distribution), as well as sentences longer than these lengths (out-of-distribution). As two training sets accounts for 5% and 10% of the overall training data, those two settings allows us to evaluate the model's data efficiency in fine-tuning and OOD generalization ability simultaneously. Detailed data partition rules are explained in Suppl.

4. Our Approach

In this section, we present our model design for the GITM-MR task and the learning objectives for training with weakly-annotated data. To tackle the problem, we develop a unified and generalizable framework as outlined in Fig. 1, which consists of two main components: a representation network that computes an initial representation of the linguistic and visual components, and a context-aware correspondence reasoning network for final match predictions. Below we first introduce the two main model components in Sec. 4.1 and 4.2, and then the model learning in Sec. 4.3.

4.1. Candidate Generation and Representation

4.1.1 Candidate Generation

Our method first generates a set of visual objects from the image and phrase candidates from the corresponding text. Specifically, for the visual object candidates, we adopt a pre-trained object detector [64] to generate N_o box proposals $\mathcal{O} = \{o_i\}_{i=1}^{N_o}$ from the image I. The detector also provides an initial visual feature (i.e., pooled ROI feature) for each object proposal. For the language phrase candidates, we utilize an off-the-shelf parser [43] to produce a language scene graph $\mathcal{G} = (\mathcal{E}, \mathcal{R})$ of the text expression L, which consists of a set of entity phrases $\mathcal{E} = \{e_i\}_{i=1}^{N_e}$ as its nodes and relation phrases as its edges $\mathcal{R} = \{r_{ij}\}$, where r_{ij} connects to subject e_i and object e_j. We denote the number of relation phrases as $|\mathcal{R}| = N_r$.

4.1.2 Candidate Representation

Given the visual and linguistic candidates, we compute their representations in two steps, as detailed below.

Token-level Representation The first step generates an initial representation for the word tokens and visual objects based on a pre-trained vision-language model. Here we adopt a shallow version of the UNITER model [8], aiming to compute a feature representation for the input tokens in two modalities with a generic vision-language alignment from the pre-trained model.*

Candidate Features In the second step, we then compile the token-level representations into a set of features for

*We note that such a generic representation module has a limited capacity and is typically insufficient for our downstream tasks via a simple fine-tuning procedure as in [8, 58]. Nevertheless, it provides rich semantic features for the subsequent alignment reasoning network (c.f. Sec. 4.2) after a minimal fine-tuning (described in Sec. 5.1), which complements our explicit reasoning process.

our visual and linguistic components used in the subsequent reasoning network. Here we adopt a similar strategy as in SGMN [61] for computing the features.

Concretely, for visual objects o_i, we augment its token-level feature $\mathbf{o}_i$ by its spatial location $\mathbf{l}_i^o$ (i.e., bounding box parameters). In addition, for every object pair (o_i, o_j) in the image, we represent its relational feature as $\mathbf{r}_{ij}^o$, which is computed from the features of its two objects. For the linguistic entities in $\mathcal{E}$ and relation phrases in $\mathcal{R}$, we feed the corresponding word tokens into a Bi-LSTM [22] with self-attention [24] to generate two sets of features. In the first set, each entities e_i is encoded by a tuple of three features $(\mathbf{e}_i, \mathbf{l}_i^e, \mathbf{h}_i^e)$, where $\mathbf{e}_i$ is an embedding into the space of object feature $\mathbf{o}_i$, $\mathbf{l}_i^e$ is an embedding into the space of object location $\mathbf{l}_i^o$, and $\mathbf{h}_i^e$ is a linguistic feature from the LSTM states. For the second feature set, we generate a representation $\mathbf{r}_{ij}^e$ for each relation phrase $r_{ij} \in \mathcal{R}$. We refer the reader to Suppl. for the details of our feature design.

4.2. Context-sensitive Propagation Network

We now introduce our second model component, the Context-sensitive Propagation Network (CPN), which aims to infer the alignment between the visual-linguistic candidates and predict the match label of input image-text pair. To this end, we build a modular graph neural network on the language graph $\mathcal{G}$, which takes the candidate features as input and performs context-aware reasoning in a belief space of visual-linguistic alignment. Such a structured reasoning process enables us to exploit the regularity in language, i.e. the connections of entities by relations, leading to a relation-aware and compositional alignment model with potentially better generalization.

Specifically, as illustrated in the language graph in Fig. 3, the i-th node of the graph network corresponds to an entity phrase e_i and the edge between the node i and j corresponds to a relation phrase r_{ij}. We follow the common assumption that the corresponding $\mathcal{G}$ has a tree structure [61, 34] and the referred entity is regarded as the root node. To perform reasoning, the graph network uses a bi-directional message propagation procedure to infer a relation-aware correspondence belief (w.r.t. the visual candidates) for all the entity nodes (described in Sec. 4.2.1), followed by a readout head module to predict the fine-grained and global match scores (described in Sec. 4.2.2). Finally, we formulate each key operation in the graph reasoning as an easy-to-interpret primitive module, and convert our reasoning process into an explainable program in Sec. 4.2.3.

4.2.1 Bi-directional Message Propagation

Given the graph network, we aim to compute a message for each node, representing a belief on the correspondence between the associated entity phrase and all the visual objects in the image. In order to resolve the ambiguity in the local matching of entities, we introduce a bi-directional message propagation procedure to aggregate all the local beliefs on phrase-visual region correspondence throughout the graph.

Concretely, inspired by belief propagation [5], we perform two parallel passes of message propagation on the tree graph $\mathcal{G}$ along two directions: 1) bottom-up direction (denoted as $\mathbf{bp}$) first computes messages from the leaf nodes and then updates their parents recursively until the root is reached; 2) top-down direction (denoted as $\mathbf{td}$) starts from the root node and updates the children nodes recursively until reaching all the leaves. For each direction, the message propagation aggregates the context information on cross-modal matching and update the belief on phrase alignment at each node in the following three steps:

(1) ***Message initialization:*** We initialize a local belief $\mathbf{b}_i^{\text{loc}}$ at node i based on a similarity measure between the entity e_i and all the visual objects. Here we define the similarity by a weighted combination of feature similarities in appearance and spatial location space. Specifically, for each pair of entity e_i and object o_j, we compute two similarities as follows,

$$b_{ij}^{\text{app}} = \mathrm{F}_{\text{sim}}(\mathbf{e}_i, \mathbf{o}_j), \quad b_{ij}^{\text{pos}} = \mathrm{F}_{\text{sim}}(\mathbf{l}_i^e, \mathbf{l}_j^o) \tag{1}$$

where F_{sim} is an MLP network computing the similarity between features from two modalities. The local belief $\mathbf{b}_i^{\text{loc}}$ then integrates two similarities as follows:

$$\mathbf{b}_i^{\text{loc}} = \mathrm{F}_{\text{norm}}\left(\beta_i^{\text{app}}\mathbf{b}_i^{\text{app}} + \beta_i^{\text{pos}}\mathbf{b}_i^{\text{pos}}\right) \tag{2}$$

where $\mathbf{b}_i^t = [b_{ij}^t]_{j=1}^{N_o}$, $t \in \{\text{app}, \text{pos}\}$. Here β_i^t is the weight for combining similarities and is computed as $\beta_i^t = \text{sigmoid}\left(\mathbf{W}_t^\top [\mathbf{h}_i^e, \mathbf{b}_i^{\text{app}}, \mathbf{b}_i^{\text{pos}}]\right)$, where F_{norm} is a normalization function to rescale its input elements to $[0, 1]$.

(2) ***Message aggregation:*** According to the direction $\mathbf{td}$ or $\mathbf{bp}$, message aggregation collects beliefs from the upstream neighbors in a context-sensitive manner. Formally, we first prune the local beliefs by selecting top K maximum values, which significantly reduces the computation complexity for large N_o and is written as,

$$\mathbf{b}_j^{\text{sel}} = \mathrm{F}_{\text{topk}}\left(\mathbf{b}_j^{\text{loc}}\right) \tag{3}$$

In contrast to [61], our selection is belief-based and more informative for propagation. Given the sparsified beliefs, we then aggregate the messages from the neigbhoring nodes by taking into account the alignment of the relation phrases on the edges as follows,

$$\mathbf{b}_i^{\text{agg}} = \prod_{j \in \text{Ctx}(i)} \mathbf{A}_{ij}\mathbf{b}_j^{\text{sel}}, \quad [\mathbf{A}_{ij}]_{kl} = \mathrm{F}_{\text{sim}}\left(\mathbf{r}_{ij}^e, \mathbf{r}_{kl}^o\right) \tag{4}$$

where $\text{Ctx}(i)$ denotes the upstream neighbor set of node v_i according to the propagation direction. We model the messages provided by relations explicitly through alignment on edges, which forces the model to capture the nuances of different relations. For the aggregation operator, we take the element-wise product of beliefs, which can be viewed as a soft version of logical AND. Unlike the min pooling used in [49, 23], our design leads to a more robust and smooth learning procedure.

(3) ***Message update:*** Finally, we update the belief $\mathbf{b}_i$ at node i by a context-sensitive integration of the aggregated neighboring belief and the local belief. To achieve this, we design a gated product as follows:

$$\beta_i^{\text{rel}} = \text{sigmoid}\left(\mathbf{W}_{\text{rel}}^\top [\mathbf{h}_i^e, \mathbf{b}_i^{\text{app}}, \mathbf{b}_i^{\text{pos}}]\right) \tag{5}$$

$$\mathbf{b}_i = \mathrm{F}_{\text{norm}}\left(\mathbf{b}_i^{\text{loc}} \circ (\mathbf{b}_i^{\text{agg}})^{\beta_i^{\text{rel}}}\right) \tag{6}$$

where $\circ$ denotes the element-wise product. Instead of using the language features as in [61, 55], our gating function also takes the local beliefs as input, making the network more sensitive to the image context.

Given the updated beliefs, we compute a confidence score for each node matching to any object in the image by applying an MLP network, denoted as $p_i = \text{sigmoid}(\text{MLP}_{\text{match}}(\mathbf{b}_i))$. In the end, the message propagation along two directions generates two belief vectors ($\mathbf{b}_i^{\text{bp}}$ and $\mathbf{b}_i^{\text{td}}$) and two confidence scores (p_i^{bp} and p_i^{td}) for each node i.

4.2.2 Readout Head

After the graph-based reasoning, we now introduce a readout head network to generate the predictions for three subtasks, i.e., ITM, REG and MRR, by exploiting the beliefs and confidence scores from the bi-directional propagation.

Image-Text Matching To predict the global matching label y, we take a minimum pooling of the matching confidence scores across the entire graph as below,

$$P(y=1|I,L) = \min_i \left(\min(p_i^{\text{bp}}, p_i^{\text{td}})\right) \tag{7}$$

REG Task For the referent grounding, we take the belief vector of the root node $\mathbf{b}_{\text{root}}^{\text{bp}}$ from the bottom-up direction and locate the matching visual object o_{i^*} by choosing the largest belief value: $i^* = \arg\max_i \mathbf{b}_{\text{root}}^{\text{bp}}$. A standard regression head is also employed for more precise localization and its details are left to the Suppl.

MRR Task For mismatched relation prediction, we exploits the inconsistency of the match confidence scores along the graph edges. We found the mismatched relation typically leads to a confidence gap between its two incident nodes. As such, we choose the edge with the largest match confidence difference between its two nodes:

$$\hat{r} = \arg\min_{r_{ij}\in\mathcal{R}} (\{(p_i^{\text{bp}} - p_j^{\text{bp}}) + (p_j^{\text{td}} - p_i^{\text{td}})\}), \tag{8}$$

where we sum the differences from two propagation directions to make the decision more robust.

Table 1. The full list of the reasoning modules and their implementations. We use bold lowercase variables to represent vectors and bold uppercase variables to represent matrices. Regular variables are scalars. $\mathbf{W}$ is trainable parameter matrix.

Modules	**In**	**Out**	**Implementation**
Local Correspondences:			
`Sim`	$\mathbf{v}^{\text{txt}}, \mathbf{v}^{\text{vis}}$	b	$\mathrm{F}_{\text{sim}}(\mathbf{v}^{\text{txt}}, \mathbf{v}^{\text{vis}})$
`GateSum`	$\{\mathbf{v}_i\}, \{\mathbf{b}_i\}$	$\mathbf{b}$	$\mathrm{F}_{\text{norm}}\left(\sum_i \text{sigmoid}\left(\mathbf{W}^\top[\mathbf{v}_i, \mathbf{b}_i]\right)\mathbf{b}_i\right)$
Correspondence Reasoning:			
`Select`	$\mathbf{b}$	$\mathbf{b}'$	$\mathrm{F}_{\text{topk}}(\mathbf{b})$
`Aggregate`	$\{\mathbf{A}_i\}, \{\mathbf{b}_i\}$	$\mathbf{b}'$	$\prod_i(\mathbf{A}_i\mathbf{b}_i)$
`GateProd`	$\mathbf{b}, \mathbf{b}', \beta_i$	$\mathbf{b}''$	$\mathrm{F}_{\text{norm}}\left(\mathbf{b} \circ (\mathbf{b}')^{\beta_i}\right)$
`Classify`	$\mathbf{b}$	p	$\text{MLP}_{\text{match}}(\mathbf{b})$
Readout Head:			
`Locate`	$\mathbf{p}$	i	$\arg\min(\mathbf{p})$
`Compare`	p_i, p_j	d	$p_i - p_j$
`And`	$\{p_i\}$	p	$\min_i p_i$

4.2.3 Modular Network Interpretation

We now introduce an interpretation of our using the modular network framework. As shown in Tab. 1, we summarize the key operations in the RCRN into a set of primitive modules with semantic meanings. This allows us to build a explainable reasoning program: (1) In the message initialization step, local beliefs are computed using `Sim`, and are combined by `GateSum` to generate the initial belief on each node. (2) For message aggregation, we first use `Select` to prune each node and then apply `Aggregate` to collect the messages from all neighbors, considering the alignment of relations particularly. (3) For message update, `GateProd` is used to combine the aggregated message with the local belief to generate the updated belief on each node. The `Classify` module provides a confidence score indicating the likelihood that the phrase entity matches an visual object. (4) In the readout head, we use two `Compare` modules in two directions respectively for MRR and one `And` operation for the ITM task, and finally generate predictions for all three subtasks via `Locate` module. For instance, the grounding prediction is achieved by `Locate`$(-\mathbf{b}_{\text{ref}})$. Several example are provided in Sec. 5.4 and Suppl. to illustrate the interpretable reasoning process, along with further detailed implementations of the modules.

We note that such a program formulation indicates that the model can enjoy a good out-of-domain generalization due to the compositionality of language graphs. Even facing with more complex text structures, we can perform this reasoning process by reusing those modules on new graphs.

4.3. Model Learning

We train the network on a multitask loss with end-to-end back propagation. Cross-Entropy (CE) loss and Binary Cross-Entropy (BCE) loss are used for the REG and ITM tasks, respectively:

$$\mathcal{L}_{\text{grd}} = \text{CE}\left(\mathbf{p}_{\text{grd}}, \mathbf{p}^*_{\text{grd}}\right) \tag{9}$$

$$\mathcal{L}_{\text{match}} = \text{BCE}\left(p_{\text{match}}, y^*\right) \tag{10}$$

where $\mathbf{p}_{\text{grd}} = \text{softmax}(\mathbf{b}^{\text{bp}}_{\text{root}})$ stands for the predicted grounding distribution, $\mathbf{p}^*_{\text{grd}}$ denotes the ground-truth one-hot grounding label, and $p_{\text{match}} = P(y = 1|I, L)$ denotes the predicted match probability. When the object boxes are from a detector, the box with the maximum IoU with the ground-truth box of the referent is labeled as the localization output. The overall loss is defined as :

$$\mathcal{L} = \mathcal{L}_{\text{grd}} + \mu\mathcal{L}_{\text{match}} + \omega\mathcal{L}_{\text{reg}} \tag{11}$$

where μ and ω are hyper-parameters to balance the three terms. ω would be zero if the IoU of the ground-truth box and the selected proposal is less than 0.5. The regression loss is standard and can be found in Suppl.

5. Experiments

In this section, we conducted experiments on the proposed benchmark to test several baselines of VL pre-trained models on inferring relation correspondences, with a specific focus on their data efficiency and length generalization capabilities. We then compare our model with those strong baselines to show its effectiveness in these settings. Finally we illustrate the visualizations and interpretations of the reasoning process of our model.

5.1. Experiment Setup

Metrics The evaluation metrics include classification accuracy for three subtasks. The grounding result for a matched case is considered as correct when it is identified as matching and the predicted box has at least 0.5 IoU with its ground-truth location. The grounding accuracy (i.e. Recall@1) is the ratio of correctly grounded cases. For mismatch reasoning, a mismatched case needs to be correctly classified and the mismatched relation should be accurately selected from the candidate set. The MRR accuracy is the top-1 accuracy among the candidates.

VL Pre-trained Models We benchmark five VL models on the proposed dataset. We first choose TCL [58], UNITER [8] and FIBER [16], where TCL is the state-of-the-art pre-trained model for image-text retrieval, UNITER is a vanilla transformer-based VL pre-trained model for matching and grounding tasks, and FIBER uses a specific fine-grained pre-training design. Additionally, we adapt the latest fine-grained visual entailment (FGVE) model [52] to our setting. We compare its potential fine-grained multi-task capability with ours, which is designed based on the VL pre-trained model Oscar+[64]. Because none of the pretrained VL models can fulfill three subtasks jointly, we also apply the multiple instance learning strategy [57] on UNITER to cope with the weakly-supervised MRR task. As shown in Suppl, we modify those models by adding subtask heads to perform three subtasks. We note that TCL and UNITER use the same pre-training datasets and the pre-training corpus of FIBER and FGVE is larger while all the models are fine-tuned on the GITM-MR, which leads to a relatively fair comparison.

Implementation Details For generating object candidates, we take the off-the-shelf detector, VinVL [64], to extract at most 100 proposals with object score great than 0.1 for each image. For model training, the reasoning component of RCRN is trained by the Adam optimizer with the learning rate set to 5e-4 for the proposed dataset. The learning rate of candidate representation module stated in Sec. 4.1 is set to 5e-7. We train 80K iterations with batch-size 64 totally. The hyper-parameter K, μ and ω are set as 5, 3 and 1 based on validation, respectively. The representation module uses the first 6 transformer encoder layers in the pre-trained UNITER. The ablation study on layers and more implementation details are included in the Suppl.

5.2. Results and Comparisons

Relation Understanding with Limited Data Tab. 2 (**In-Distribution** part) shows the performances of pre-trained models on the relation understanding under the condition of limited data. The results indicate that TCL struggles to achieve a match accuracy of more than chance. This could be attributed to its emphasis on global image-text matching during pre-training, which may hinder its ability to perform fine-grained understanding tasks. Though FGVE is designed for fine-grained entailment prediction, it still lacks capability on mismatch relation reasoning in our scenario. This is likely due to its reliance on the pre-trained VL model for the alignment between vision and language relations. UNITER and FIBER perform relatively better on the match task than TCL because they both use the World-Region Alignment (WRA) as a pre-training task, and FIBER is also pre-trained with phrase grounding, which helps to establish fine-grained correspondence between image and text. The overall conclusion drawn from these findings is that pre-trained models exhibit unsatisfying performance on the challenging benchmark under the limited data condition, indicating the need for further development in establishing effective visual-linguistic correspondence for relation understanding.

In comparison, our RCRN outperforms the SOTA pre-trained models on match and grounding tasks in the in-distribution scenarios, even with fewer parameters. The model's performance with limited training data indicates the effectiveness of its lightweight and structured design. The design incorporates appropriate inductive biases based

Table 2. Experiment results of testing models' relation understanding with limited data and ability of length generalization. UNITER+MIL represents the UNITER with multiple instance learning strategy applied. FGVE+MAX represents the fine-grained visual entailment model with max pooling on knowledge elements to predict mismatched relation. '-' means the model can't handle the corresponding task.

Training Set	Method	#Pretrain Images	#Param	Full Test			In-Distribution			Out-of-Distribution		
				Match%	Grounding%	MRR%	Match%	Grounding%	MRR%	Match%	Grounding%	MRR%
Train-Len16	TCL [58]	4M	210M	52.77	-	-	53.11	-	-	52.70	-	-
	UNITER [8]	4M	111M	61.93	24.16	-	71.29	40.63	-	59.94	20.67	-
	UNITER+MIL [57]	4M	111M	59.13	18.46	47.5	68.83	41.53	**61.38**	57.07	13.56	44.53
	FIBER [16]	4.8M	254M	52.31	9.01	-	59.04	37.15	-	50.88	3.03	-
	FGVE+MAX [52]	5.7M	113M	54.30	-	33.23	58.37	-	51.79	53.43	-	29.26
	RCRN (Ours)	4M	90M	**66.59**	**29.10**	**55.72**	**71.85**	**46.24**	59.04	**65.47**	**25.46**	**55.01**
Train-Len11	TCL [58]	4M	210M	51.71	-	-	52.08	-	-	51.70	-	-
	UNITER [8]	4M	111M	55.51	10.24	-	68.22	30.05	-	54.97	9.42	-
	UNITER+MIL [57]	4M	111M	55.43	11.44	42.73	65.04	29.56	**61.65**	55.02	10.69	41.93
	FIBER [16]	4.8M	254M	53.52	11.74	-	57.21	29.56	-	53.36	11.00	-
	FGVE+MAX [52]	5.7M	113M	53.40	-	38.05	55.26	-	40.78	53.31	-	37.93
	RCRN (Ours)	4M	90M	**63.41**	**21.45**	**52.67**	**68.70**	**37.93**	57.77	**63.19**	**20.77**	**52.46**

on sentence structure to establish both relation and entity correspondence in a visual-linguistic context, and learns the reasoning strategy effectively.

Length Generalization Results about length generalization ability are shown in the **Out-of-Distribution** part of Tab. 2. Overall, the OOD performance of pre-trained models on both test settings was generally poor compared to their in-distribution performance, especially on the grounding and MRR subtasks, which explicitly reflect the models' understanding of relations. In particular, in the OOD setting, FIBER's match accuracy is similar to TCL, with both models only slightly exceeding 50%. The OOD results suggest that these models tend to overfit to the training data biases, without truly establishing the universal VL correspondence of the relations, thus preventing them from generalizing to longer sentences.

In contrast, RCRN achieved notably better performance than the other models on all three subtasks in both OOD scenarios. Specifically, RCRN trained with Train-Len11 outperformed the best performance of other models by 8.17%, 10.08%, and 10.53% on all three tasks. This success can be attributed to the modular design, which helps the model learn to establish local visual-linguistic correspondence and the universal reasoning strategy. During test, the model can exploit the regularity in sentence structures to compose the modules and deal with longer sentences.

Discussion These experiments not only help us investigate the two challenges that the pre-trained models faced, but also lead to other interesting findings. First, our observation that fine-grained pre-training tasks, such as WRA, can help models learn local correspondences on downstream tasks, is consistent with the findings of [36]. Second, our findings in the OOD scenario are consistent with previous works [3, 54], which suggest that pre-trained models are prone to overfitting to the training data length biases.

Table 3. Ablation study on the benchmark's OOD validation set.

Langβ	Ctxβ	Bi-MP	VLP	Mat	Grd	MRR
-	-	-	✓(w/o MP)	54.79	13.99	41.47
-	-	✓	✓	59.65	23.80	45.58
✓	-	✓	✓	60.88	24.18	49.57
✓	✓	-	✓	60.07	23.88	48.62
✓	✓	✓	-	60.70	**29.25**	36.65
✓	✓	✓	✓	**61.47**	23.37	**53.23**

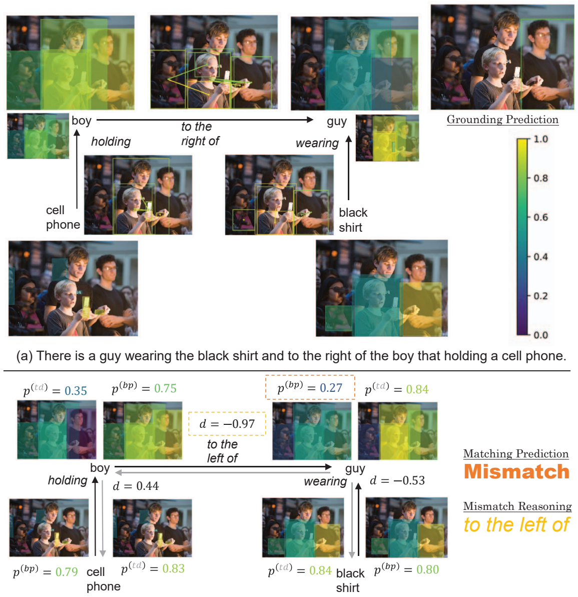
Moreover, to independently investigate the capability on two subtasks, namely grounding and MRR, we also evaluate on an oracle setting, where we suppose that the models always predict correctly on the matching task. The experiment demonstrates our RCRN has strong performance on both subtasks, and the details of this experiment are provided in the Suppl.

Overall, our experiments provide insights into the challenges and potential solutions for the relation understanding in VL models. The fine-grained pre-training tasks can help models learn local correspondences, while the modular design can facilitate the composition of modules to deal with longer sentences. We believe that our findings offer a promising direction to effectively address these challenges.

5.3. Ablation study

In this section, we present the effectiveness of each component in the proposed RCRN through ablative experiments trained on the Train-Len16 set, with the results in Tab. 3.

We conduct ablation study by removing some specific designs. **Lang**β stands for using the language features as one input for gate functions in Eq. 2 and Eq. 6. Removing it indicates constant β's in the propagation. **Ctx**β means feeding initial similarities into the gate functions. **Bi-MP** represents propagating twice individually using different β^{rel}'s for matching task reasoning. Without this part, only the propagation from the root to leaves is conducted. **VLP**



(a) There is a guy wearing the black shirt and to the right of the boy that holding a cell phone.

(b) There is a guy wearing the black shirt and to the left of the boy that holding a cell phone.

Figure 4. The visualization of the reasoning process of RCRN. The entity and relation correspondence maps are shown by color maps on boxes and box pairs respectively, where only those with top 5 correspondence scores for each phrase are included. Small color maps in (a) are the local correspondence maps and the large ones are the propagated contextualized maps. The boxes of parents in box pairs are dotted on their centers. Dashed boxes indicate the pooling results.

stands for introducing vision-language pre-trained transformer layers to encode features. Particularly, the first line with VLP only stands for the model which consists of 6 transformer layers from UNITER and the UNITER task heads stated in Sec. 5.1 and Sec. 5.2.

The results show that all of these designs improve the matching performance. VLP model (first line in Tab. 3) with the same setting as the VLP layers in RCRN performs poorly on OOD scenario, which demonstrates our footnote statement in Sec. 4.1. Nevertheless, VLP component of RCRN is significant on MRR performance, probably because it provides accurate preliminary vision-language alignments on relations.

5.4. Visualization and Interpretability

As shown in Fig. 4, we visualize a pair of matched and mismatched samples of our RCRN. Fig. 4(a) shows the bottom-up propagation process followed by grounding prediction on a matched image-text pair, and Fig. 4(b) presents the bi-directional propagation results with readout process for MRR on a mismatched pair. In Fig. 4(a), the propagation eliminates the ambiguities on entities under the guidance of explicit relation correspondences. Fig. 4(b) illustrates that the intermediate reasons of the predictions can be traced. The sum of match confidence differences along the mismatched relation, denoted as d in the dotted box, are notably lower than other d's.

Table 4. Intermediate diagnosis results of RCRN. E.G.R. means Entity Grounding Recall. R.G.R. means Referent Grounding Recall. R.C.S. means Relation Correspondence Scores.

Inference	E.G.R.			R.G.R.			R.C.S.		
Method	R@1	R@3	R@5	R@1	R@3	R@5	GT	Mean	Max
By chance	6.37	18.25	28.78	5.74	16.94	27.22	-	-	-
w/o MP	50.50	80.59	89.79	38.88	75.84	88.11	0.61	0.17	0.69
w/ MP	60.82	85.58	92.16	69.46	90.92	95.59	0.61	0.17	0.69

Tab. 4 demonstrates some interpretable intermediate diagnosis results of RCRN on a subset of the validation set, with ground-truth box proposals. We analyze from the aspect of phrase-level visual grounding recall and statistics on relation correspondence scores. To obtain the true correspondence between the language scene graph and the bounding boxes, we search the parsed language scene graph in the original image scene graph in GQA dataset [26] with a noisy graph matching algorithm. See Suppl. for the detail of the algorithm.

The results verify that reasonable local correspondences are built, since the mean of initial recall (i.e. inference without MP) on entity phrase grounding is much higher than the random guess. As for relation correspondences, the average values on the ground-truth vision-language relation correspondences are far above the average mean value on all the pairs and very close to the average maximum. Moreover, message propagation notably helps refine the correspondence maps, especially on the referent entities.

6. Conclusion

In this paper, we have introduced a novel VL joint task, Grounded Image Text Matching with Mismatched Relation (GITM-MR), which provides a challenging benchmark for evaluating pre-trained models on the relation reasoning. Our experiments have shown that some state-of-the-art pre-trained models struggle with the task under the setting of limited data and out-of-distribution sentence lengths. To address this problem, we develop a *Relation-sensitive Correspondence Reasoning Network (RCRN)* to compute contextualized cross-modal alignment of both entities and relations, and ground the fine-grained result in the image or expression. The proposed method outperforms prior SOTA pre-trained models, which demonstrates its strong generalization and data efficiency. Our work sheds light on the limitations of current pre-trained models and the importance of establishing fine-grained vision-language correspondence for relation understanding.

References

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4971–4980, 2018. 2

[2] Arjun R Akula. Question generation for evaluating cross-dataset shifts in multi-modal grounding. *arXiv preprint arXiv:2201.09639*, 2022. 2

[3] Cem Anil, Yuhuai Wu, Anders Andreassen, Aitor Lewkowycz, Vedant Misra, Vinay Ramasesh, Ambrose Slone, Guy Gur-Ari, Ethan Dyer, and Behnam Neyshabur. Exploring length generalization in large language models. *arXiv preprint arXiv:2207.04901*, 2022. 1, 2, 8

[4] Arpit Bansal, Avi Schwarzschild, Eitan Borgnia, Zeyad Emam, Furong Huang, Micah Goldblum, and Tom Goldstein. End-to-end algorithm synthesis with recurrent networks: Logical extrapolation without overthinking. *arXiv preprint arXiv:2202.05826*, 2022. 2

[5] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006. 5

[6] Peter Bühlmann. Invariance, causality and robustness. *Statistical Science*, 35(3):404–426, 2020. 2

[7] Tianlang Chen and Jiebo Luo. Expressing objects just like words: Recurrent visual embedding for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10583–10590, 2020. 3

[8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 1, 2, 4, 7, 8

[9] Mengjun Cheng, Yipeng Sun, Longchao Wang, Xiongwei Zhu, Kun Yao, Jie Chen, Guoli Song, Junyu Han, Jingtuo Liu, Errui Ding, et al. Vista: vision and scene text aggregation for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5184–5193, 2022. 3

[10] Myung Jin Choi, Antonio Torralba, and Alan S Willsky. Context models and out-of-context objects. *Pattern Recognition Letters*, 33(7):853–862, 2012. 2

[11] JU Da, Stephen Roller, Sainbayar Sukhbaatar, and Jason E Weston. Staircase attention for recurrent processing of sequences. In *Advances in Neural Information Processing Systems*. 2

[12] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779, 2021. 3

[13] Jiajun Deng, Zhengyuan Yang, Daqing Liu, Tianlang Chen, Wengang Zhou, Yanyong Zhang, Houqiang Li, and Wanli Ouyang. Transvg++: End-to-end visual grounding with language conditioned vision transformer. *arXiv preprint arXiv:2206.06619*, 2022. 3

[14] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1218–1226, 2021. 3

[15] Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*, 2020. 1

[16] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. *arXiv preprint arXiv:2206.07643*, 2022. 1, 7, 8

[17] Yann Dubois, Gautier Dagan, Dieuwke Hupkes, and Elia Bruni. Location attention for extrapolation to longer sequences. *arXiv preprint arXiv:1911.03872*, 2019. 2

[18] Stella Frank, Emanuele Bugliarello, and Desmond Elliott. Vision-and-language or vision-for-language. *On Cross-Modal Influence in Multimodal Transformers.(2021). DOI: https://doi. org/10.18653/v1/2021. emnlp-main*, 775, 2021. 2

[19] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628, 2020. 2

[20] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 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, 2017. 2

[21] Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644, 2021. 1, 2

[22] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 5

[23] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 804–813, 2017. 6

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

[25] Binbin Huang, Dongze Lian, Weixin Luo, and Shenghua Gao. Look before you leap: Learning landmark features for one-stage visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16888–16897, 2021. 3

[26] Drew A Hudson and Christopher D Manning. 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, 2019. 9

[27] Zhong Ji, Haoran Wang, Jungong Han, and Yanwei Pang. Saliency-guided attention network for image-sentence matching. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5754–5763, 2019. 3

[28] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE international conference on computer vision*, pages 2989–2998, 2017. 2

[29] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 1, 2

[30] Kun Kuang, Ruoxuan Xiong, Peng Cui, Susan Athey, and Bo Li. Stable prediction with model misspecification and agnostic distribution shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4485–4492, 2020. 2

[31] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 4654–4662, 2019. 3

[32] Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *Advances in Neural Information Processing Systems*, 34:19652–19664, 2021. 3

[33] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 624–639, 2018. 2

[34] Yongfei Liu, Bo Wan, Xiaodan Zhu, and Xuming He. Learning cross-modal context graph for visual grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11645–11652, 2020. 3, 5

[35] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *European Conference on Computer Vision*, pages 792–807. Springer, 2016. 3

[36] Mitja Nikolaus, Emmanuelle Salin, Stephane Ayache, Abdellah Fourtassi, and Benoit Favre. Do vision-and-language transformers learn grounded predicate-noun dependencies? *arXiv preprint arXiv:2210.12079*, 2022. 1, 2, 8

[37] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710, 2021. 2

[38] Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, 2022. 1, 2

[39] Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021. 2

[40] Mengxue Qu, Yu Wu, Wu Liu, Qiqi Gong, Xiaodan Liang, Olga Russakovsky, Yao Zhao, and Yunchao Wei. Siri: A simple selective retraining mechanism for transformer-based visual grounding. In *European Conference on Computer Vision*, pages 546–562. Springer, 2022. 3

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

[42] Emmanuelle Salin, Badreddine Farah, Stéphane Ayache, and Benoit Favre. Are vision-language transformers learning multimodal representations? a probing perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11248–11257, 2022. 2

[43] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pages 70–80, 2015. 4

[44] Avi Schwarzschild, Eitan Borgnia, Arjun Gupta, Furong Huang, Uzi Vishkin, Micah Goldblum, and Tom Goldstein. Can you learn an algorithm? generalizing from easy to hard problems with recurrent networks. *Advances in Neural Information Processing Systems*, 34:6695–6706, 2021. 2

[45] Ravi Shekhar, Sandro Pezzelle, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. Vision and language integration: Moving beyond objects. In *IWCS 2017—12th International Conference on Computational Semantics—Short papers*, 2017. 2

[46] Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. Foil it! find one mismatch between image and language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265, 2017. 1, 2

[47] Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Disentangled generative causal representation learning. *arXiv preprint arXiv:2010.02637*, 2020. 2

[48] Zheyan Shen, Peng Cui, Tong Zhang, and Kun Kunag. Stable learning via sample reweighting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5692–5699, 2020. 2

[49] Jiaxin Shi, Hanwang Zhang, and Juanzi Li. Explainable and explicit visual reasoning over scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8376–8384, 2019. 2, 6

[50] Damien Teney, Ehsan Abbasnejad, Kushal Kafle, Robik Shrestha, Christopher Kanan, and Anton Van Den Hengel. On the value of out-of-distribution testing: An example of goodhart's law. *Advances in Neural Information Processing Systems*, 33:407–417, 2020. 2

[51] Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. Unshuffling data for improved generalization in visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1417–1427, 2021. 2

[52] Christopher Thomas, Yipeng Zhang, and Shih-Fu Chang. Fine-grained visual entailment. *arXiv preprint arXiv:2203.15704*, 2022. 1, 2, 7, 8

[53] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 2

[54] Dušan Variš and Ondřej Bojar. Sequence length is a domain: Length-based overfitting in transformer models. *arXiv preprint arXiv:2109.07276*, 2021. 1, 2, 8

[55] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. *International Conference on Learning Representations*, 2018. accepted as poster. 6

[56] Jia Wang, Jingcheng Ke, Hong-Han Shuai, Yung-Hui Li, and Wen-Huang Cheng. Referring expression comprehension via enhanced cross-modal graph attention networks. *ACM Journal of the ACM (JACM)*, 2022. 3

[57] Yun Wang, Juncheng Li, and Florian Metze. A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35. IEEE, 2019. 7, 8

[58] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. 2022. 1, 2, 4, 7, 8

[59] Li Yang, Yan Xu, Chunfeng Yuan, Wei Liu, Bing Li, and Weiming Hu. Improving visual grounding with visual-linguistic verification and iterative reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9499–9508, 2022. 3

[60] Sibei Yang, Guanbin Li, and Yizhou Yu. Dynamic graph attention for referring expression comprehension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4644–4653, 2019. 3

[61] Sibei Yang, Guanbin Li, and Yizhou Yu. Graph-structured referring expression reasoning in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9952–9961, 2020. 2, 3, 5, 6

[62] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[63] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 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*, 2022. 2

[64] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 1, 2, 4, 7

[65] Zelin Zhao, Karan Samel, Binghong Chen, et al. Proto: Program-guided transformer for program-guided tasks. *Advances in Neural Information Processing Systems*, 34:17021–17036, 2021. 2

[66] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, and Anton Van Den Hengel. Parallel attention: A unified framework for visual object discovery through dialogs and queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4252–4261, 2018. 3