Divide and Conquer: Answering Questions with Object Factorization and Compositional Reasoning

Shi Chen Qi Zhao
Department of Computer Science and Engineering,
University of Minnesota
{chen4595, qzhao}@umn.edu

Abstract

Humans have the innate capability to answer diverse questions, which is rooted in the natural ability to correlate different concepts based on their semantic relationships and decompose difficult problems into sub-tasks. On the contrary, existing visual reasoning methods assume training samples that capture every possible object and reasoning problem, and rely on black-boxed models that commonly exploit statistical priors. They have yet to develop the capability to address novel objects or spurious biases in real-world scenarios, and also fall short of interpreting the rationales behind their decisions. Inspired by humans' reasoning of the visual world, we tackle the aforementioned challenges from a compositional perspective, and propose an integral framework consisting of a principled object factorization method and a novel neural module network. Our factorization method decomposes objects based on their key characteristics, and automatically derives prototypes that represent a wide range of objects. With these prototypes encoding important semantics, the proposed network then correlates objects by measuring their similarity on a common semantic space and makes decisions with a compositional reasoning process. It is capable of answering questions with diverse objects regardless of their availability during training, and overcoming the issues of biased question-answer distributions. In addition to the enhanced generalizability, our framework also provides an interpretable interface for understanding the decision-making process of models. Our code is available at https://github.com/szzexpoi/POEM.

1. Introduction

One of the fundamental goals in artificial intelligence is to develop systems that are able to reason with the complexity of real-world data to make decisions. Most existing visual question answering (VQA) methods [2, 13, 28, 29, 33, 34, 38, 49, 57] assume a complete overlap between objects involved in training and testing, and commonly rely on the spurious distributions of questions and answers [39]. As a result, they have limited generalizability toward real-life visual reasoning, and also lack the ability to justify the reasoning process that leads to the answers.

"All mammals are animals. All elephants are mammals. Therefore, all elephants are animals [5]." The wide application of syllogistic logic reflects key characteristics of the ways humans reason about the world. Unlike models [2, 38, 49] that utilize implicit features and heavily exploit statistical priors, humans correlate diverse objects from the compositional perspective based on their shared characteristics [26] and tackle problems with a structured reasoning process, which is both generalizable and interpretable.

To address the complexity of real-world problems, this study aims to develop object factorization and compositional reasoning capabilities in models. As shown in Figure 1, our approach bridges diverse objects by projecting them onto a common space formed by discriminative prototypes (*e.g.*, round shape, stuffed toy), and formulates the reasoning process with atomic steps [48] representing essential reasoning skills (*e.g.*, *Find*, *Relate*). The prototypes are derived with object factorization, and they represent important semantics of objects (*e.g.,* honey jar → <round shape, container ...>, teddy bear → <bear, stuffed toy ...>). With an improved understanding of semantic relationships, our framework correlates objects (*e.g.,* honey jar and container, stuffed toy and teddy bear) based on their commonalities in characteristics, leading to enhanced robustness against the diversity of objects and data biases. It also allows interpretations of the model's reasoning process consistent with the ways humans describe their own thinking [8].

Compared to previous studies [2, 21, 33, 34, 48, 49], our major distinction lies in (1) the composition in two important dimensions of visual reasoning, *i.e.*, objects and the reasoning process, and (2) a tight coupling between them. Instead of using black-boxed features or a direct mapping between question and answer that is vulnerable to object di-

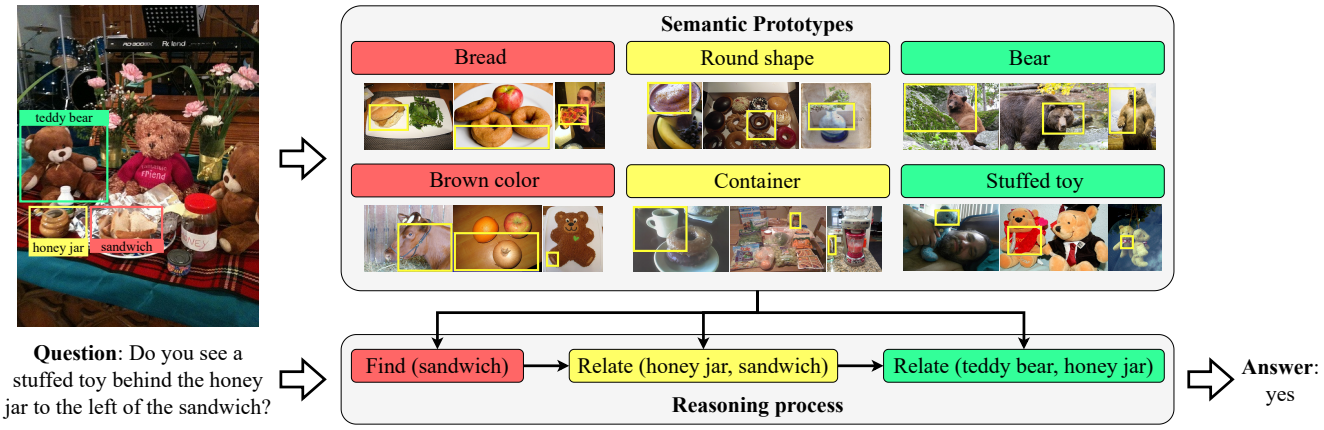

Figure 1. Overview of our method that represents objects with semantically meaningful prototypes and makes decisions via an explicit reasoning process. Honey jar is a novel object unseen during training. Note that our prototypes are not limited to a set of manually defined categories, but learned from factorizing objects to encode broader characteristics (*e.g.,* shapes, colors, object categories).

versity or data biases, our method decomposes objects into bases representing discriminative semantics, and develops a prototypical neural module network to explicitly bridge objects with a compositional reasoning paradigm. The proposed method naturally approaches generalizability with its compositional nature, handling novel objects and variable data distributions. It also provides a transparent interface for interpreting how models parse objects based on their characteristics and incorporate them for visual reasoning.

To summarize, our major contributions are as follows:

1. We identify the significance of tightly coupling the compositionality between objects and the reasoning process, and for the first time investigate its effectiveness in generalizable and interpretable reasoning.
2. We propose a principled method that automatically derives prototypes with object factorization, which plays a key role in encoding key characteristics of objects.
3. We develop a new neural module network that adaptively reasons on the commonalities of different objects along a structured decision-making process.
4. We perform extensive analyses to shed light on the roles of compositionality in reasoning with novel objects and variable data distributions.

2. Related Works

Our study is most related to previous efforts on visual question answering, zero-short learning for VQA, and VQA with out-of-distribution (OOD) questions.

Visual question answering. With the diversity in language and visual semantics, visual question answering has become a popular task for studying models' reasoning capability [16]. A large body of research develops datasets [4,6,18,24,25,40,42,58] and models [2,3,21,22,28,33,34,38,48,49] for VQA. Early datasets typically rely on crowdsourcing [4, 18, 58] to collect human-annotated questions. Several recent studies [24, 25, 53] use functional programs to automatically generate questions based on pre-defined rules and enable more balanced data distributions. There is also an increasing interest in investigating different types of reasoning, *e.g.,* scene text understanding [6], reasoning on context [42], and knowledge-based reasoning [40]. These data efforts lead to the development of methods that advance VQA models from different perspectives, including multi-modal fusion [13,29,57], attention mechanism [2,28], structured inference [21–23, 48], and vision-and-language pretraining [33, 34, 38, 49]. The aforementioned studies assume that every semantic in the test questions is well illustrated during training, and pay little attention to the models' generalizability to real-world scenarios that inevitably involve novel objects and diverse question-answer distributions. Our study aims to fill the gap with an integral framework that develops generalizable decision making capability with object factorization and compositional reasoning.

Zero-shot learning for VQA. Zero-shot VQA aims to answer questions with novel objects. The pioneering work [51] proposes the zero-shot setting for the VQA task, and benchmarks several techniques for improving the models' generalizability, including explicit object detector [2] and pretrained word embeddings [43]. Ramakrishnan *et al.* [45] leverage self-supervised pretraining to learn more generalizable features with external data. Wang *et al.* [54] combine visual models trained on different datasets with an attention mechanism. Whitehead *et al.* [55] decompose VQA into two sub-problems, *i.e.,* concept grounding and skill matching, and propose additional training objectives to address unseen objects in the questions. Besides the above stud-

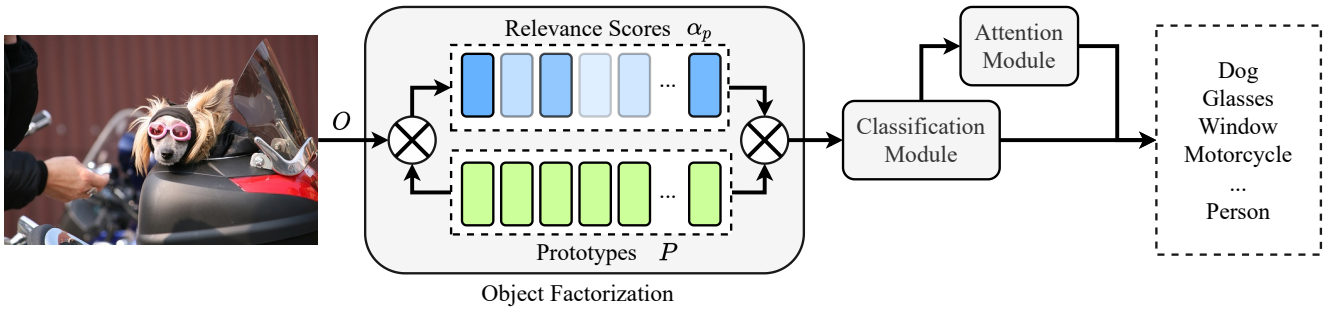

Figure 2. Overview of our prototype learning paradigm based on object factorization.

ies concerning novel concepts in language, there are also attempts [12, 52] that leverage exemplar-based methods to address novel concepts in both visual and textual modalities. While showing usefulness, these studies either leverage external data, which violates zero-shot assumption [32], or require memorizing a vast amount of supporting samples. As a result, they have yet to develop the capability of generalizing toward real-world scenarios. The key differentiators of our method lie in its ability to connect novel and known objects and the use of an explicit reasoning process. By bridging objects based on their commonalities and decomposing the decision-making process into atomic reasoning steps, it improves the generalizability and interpretability without relying on external data or memorizing samples.

VQA with OOD questions. To develop models that can truly reason on the visual-textual inputs instead of relying on statistical priors, VQA with out-of-distribution questions has gained considerable attention. In particular, Agrawal *et al.* [1] present the first VQA dataset with adversarial distributions between the training and validation data. A more recent study [27] analyzes models' robustness against biases by differentiating evaluation questions based on their question-answer distributions. To tackle the issues of harmful biases, a series of studies make progresses by improving visual attention [47, 56], reducing biases toward individual modalities [7, 15, 44], and leveraging ensemble techniques [11, 20]. There are also attempts [9, 17, 35] that use data augmentation to increase accuracy on biased data. However, they alter the distributions of training samples with additional data, and violate the original intent of the VQA with OOD questions task [50]. Our study is complementary to existing efforts, and it differentiates itself by investigating the usefulness of object factorization and compositional reasoning for addressing biases. It does not balance training samples to remove data biases, but instead focuses on enhancing models' own reasoning capabilities.

3. Methodology

Visual reasoning would benefit from capabilities of correlating objects based on their characteristics and decomposing problems into atomic steps [19]. This section presents a new framework for improving the robustness against questions with novel objects and diverse question-answer distributions. It consists of two novel components: (1) a principled method that automatically derives semantically plausible prototypes to represent different objects, and (2) a neural module network that bridges objects by incorporating discriminative prototypes in an explicit reasoning process. Besides the enhanced generalizability, the method also provides an interpretable interface for elaborating on the rationales behind the model's decisions.

3.1. Bridging Diverse Objects with Factorization

Inspired by humans' reasoning process that classifies objects based on their semantics similarity, a primary goal of our study is to derive semantic prototypes that can represent a vast amount of objects. The prototypes encode different aspects of the objects, and augment models with the capability to bridge diverse objects for more generalizable reasoning. Unlike previous studies [14, 41] that construct prototypes based on manually annotations and have difficulties scaling to different scenarios, we propose to automatically learn discriminative prototypes by factorizing various objects. Object factorization plays a key role in parsing the fine-grained characteristics of objects (*e.g.,* shapes, textures, and super-categories), which facilitates understanding of the semantic relationships between objects.

As illustrated in Figure 2, our prototype learning method leverages the multi-label classification task [10, 46, 59, 61] to discover discriminative prototypes in a data-driven manner. Given an input image, we train a deep neural network that predicts all object categories in the visual scene. Different from conventional approaches that recognize objects based on their visual features O, we decompose objects with trainable prototypes P and utilize the combinations of their matched prototypes for classification:

$$\alpha_p = \delta(O_i \cdot P) \tag{1}$$

$$C^i = Cls(\sum_{k=1}^{K} \alpha_p^k P^k) \tag{2}$$

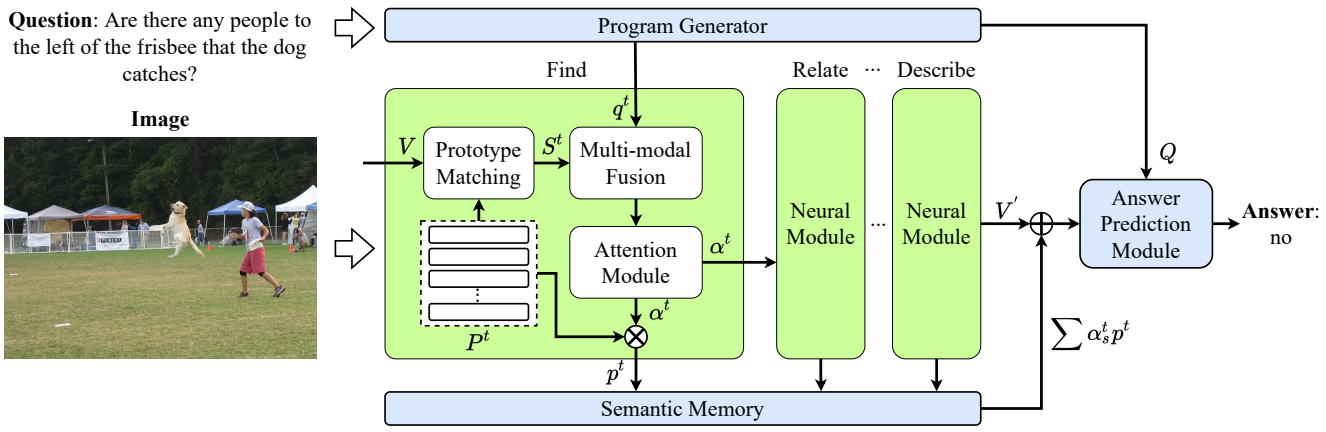

Figure 3. Illustration of the proposed prototypical neural module network. $\otimes$ and $\oplus$ denote dot product and concatenation, respectively.

where $\alpha_p \in \mathbb{R}^{1\times K}$ denotes relevance scores between the i^{th} object and K prototypes. $O_i \cdot P$ is the dot product between features and prototypes, which corresponds to their cosine similarity. δ is the sigmoid activation function for normalization. C^i is the prediction for the current object and Cls is the classification module. To facilitate optimization without expensive instance-level annotations, we leverage an attention module A_{cls} to dynamically aggregate instance-level predictions into the image-wise prediction C:

$$\alpha_c^i = A_{cls}(\sum_{k=1}^{K} \alpha_p^k P^k) \tag{3}$$

$$C = \sum_{i=1}^{N} \alpha_c^i C^i \tag{4}$$

where N is the number of objects. We train the network with a standard binary cross-entropy loss, and select prototypes P that provide the highest validation performance.

The aforementioned paradigm factorizes objects into a set of bases formed by different prototypes. The prototypes encode important semantics of objects (see Section 4.5), and serve as critical components of our reasoning model, as detailed in the next subsection.

3.2. Prototypical Neural Module Network

With our prototypes constructing the common semantic space for bridging diverse objects, we further propose a novel prototypical neural module network to adaptively incorporate the correlation between objects with a compositional reasoning process. Compared to previous VQA methods [2,12,33,45,48,49,52,55], the advantages of our model lie in (1) its capability to generalize to both known and novel objects in the visual-textual domains, (2) the enhanced robustness against the spurious data biases, and (3) the interpretability of the decision-making procedure.

Figure 3 provides an overview of the proposed neural module network with prototype matching and a semantic memory module. The principal idea behind our model is to take advantage of semantic relationships encoded in our learned prototypes (Section 3.1), and reason with a structured decision-making procedure. Neural module networks [3,21,48] are a body of interpretable reasoning methods that perform visual reasoning with two steps: (1) decomposing the reasoning process into discrete reasoning steps, where each step is associated with an atomic module (*e.g., Find* module for locating regions of interest), and (2) sequentially executing the modules on visual-textual inputs and gathering information to predict the answer. In addition to the use of an explicit reasoning process, our model utilizes prototype matching to project features of diverse objects onto the semantic space formed by prototypes, enabling it to measure their semantic relationships and tightly couple relevant objects. A semantic memory module is also proposed to adaptively combine important semantics captured at different reasoning steps, which facilitates joint reasoning throughout the whole reasoning process.

Specifically, unlike conventional methods that rely on raw visual features and pay little attention to semantic relationships between objects, our model leverages the learned prototypes to parse different objects and correlate them based on their fine-grained characteristics. At each t^{th} reasoning step, our network computes the similarity between visual features and prototypes, explicitly representing objects based on their corresponding characteristics, and uses the similarity scores $S^t \in \mathbb{R}^{N\times K}$ for decision making:

$$S_i^t = \phi(V_i \cdot P^t) \tag{5}$$

where $V_i \in \mathbb{R}^{1\times D}$ is D dimensional visual features for the i^{th} visual regions (N regions in total), and $P^t \in \mathbb{R}^{D\times K}$ represents the K prototypes. $V_i \cdot P^t$ is the dot product between features and prototypes. ϕ is the hyperbolic tangent

activation function for normalizing the scores. Upon obtaining the similarity scores, we then locate the regions of interest for the current reasoning step:

$$\alpha^t = A^t(F^t(S^t, q^t)) \tag{6}$$

where $\alpha^t \in \mathbb{R}^{1\times N}$ is the attention map highlighting the important regions, F^t and A^t are the multi-modal fusion and attention module, respectively. q^t is the query information derived from the question.

Another key differentiator between our proposed model and existing neural module networks [21, 48] is the incorporation of semantic memory. Instead of determining the answer solely based on the question Q and attended visual features $V^{'}$, we further take into account the prototypes attended over time $p^t = \alpha^t \cdot S^t$, and bridge objects of interest at different reasoning steps. Our semantic memory module uses an attention mechanism to adaptively incorporate key prototypes at different steps:

$$\alpha_s = A_s(Q) \tag{7}$$

$$\hat{y} = Ans([V^{'}; \sum_{t=1}^{T} \alpha_s^t p^t], Q) \tag{8}$$

where $\alpha_s \in \mathbb{R}^{1\times T}$ represents attention weights for T reasoning steps, and A_s is the module for attention computation. $\hat{y}$ is the predicted answer, and Ans is the answer prediction module. $[;]$ denotes the concatenation of features.

Our proposed model associates objects based on their relationships with distinct prototypes, and adaptively combines important prototypes captured throughout the reasoning process. It takes advantage of the compositionality in both objects involved during visual reasoning (*i.e.,* from objects to their characteristics) and the reasoning process (*i.e.,* from questions to reasoning steps), which plays an essential role in addressing the diversity of objects and the spurious data biases. The compositional nature of the model also allows better interpretation of the underlying decision-making procedure (Section 4.6).

4. Experiments

This section presents implementation details (Section 4.1) and experiments to analyze the proposed method. We experiment with two different settings of VQA, including zero-shot VQA (Section 4.2) and VQA with out-of-distribution questions (Section 4.3), to validate the robustness of our method. We also provide an ablation study with different prototypes (Section 4.4) to demonstrate the advantages of object factorization. Besides examining the effectiveness of our method, we perform extensive analyses to shed light on the following questions: (1) What do prototypes learn? Do they encode common characteristics among objects? (Section 4.5); and (2) How do models reason to answer diverse questions? (Section 4.6)

4.1. Implementations

Datasets. The primary goal of our experiments is to study models' generalizability to tackle real-world problems. We experiment with two different settings, each representing a common type of generalization: (1) **Zero-shot VQA** estimates models' generalizability toward both known and novel objects. Three popular datasets are used in our experiments, including VQA [18], GQA [24] and the recently introduced Novel-VQA [55]. Following [45, 55], for the VQA and GQA datasets, we reconstruct their training and validation sets to have a disjoint set of objects. For each dataset, we randomly select ten objects from the object pools and use them as novel objects unavailable during training. Similar to [12, 45, 51], we exclude all training questions that either contain words related to the selected objects or use images with the objects, and divide the original validation sets into Known and Novel splits based on objects required for reasoning. For the Novel-VQA dataset, we adopt the original training and test sets [55], which only considers novel objects in the questions (Novel-Q); (2) **VQA with OOD questions** focuses on evaluating models' generalizability to data with diverse question-answer distributions. We experiment with two popular datasets, including VQA-CP [1] with adversarial distributions between training and evaluation, and GQA-OOD [27] that utilizes balanced training questions but differentiates evaluation questions based on their question-answer distributions.

Evaluation. We evaluate models with common metrics for visual question answering. For the VQA, Novel-VQA, and VQA-CP datasets, we adopt VQA accuracy [4] as the evaluation metric, which considers multiple candidate answers. For the GQA and GQA-OOD datasets, we use the standard accuracy since each question has an unique answer. We also follow [27] and consider differences in accuracy when answering in-distribution and out-of-distribution questions (*i.e.,* $\triangle$) on the GQA-OOD dataset.

Model specification. We use the state-of-the-art neural module network XNM [48] as our baseline, which provides competitive performance on multiple datasets without loss of interpretability. We follow [60] and replace the *Transform* module with the *Relate* module to improve attention propagation. Following [34, 49, 55], we adopt the UpDown features [2] that capture 36 semantically meaningful regions (*i.e.,* N=36) as the visual inputs. To enable understanding of unseen vocabulary, we follow [51, 55] and use Glove vectors [43] to initialize the word embeddings. Other settings, *e.g.,* network specification and training configuration, are the same as those defined in the original papers [48, 60] without tuning. The aforementioned baseline is incorporated with our method discussed in Section 3.2 to enable object factorization and compositional reasoning.

Prototype learning. We derive our prototypes with the multi-label classification task [59, 61], which aims to pre-

Table 1. Comparative results on zero-shot VQA.

	VQA		GQA		Novel-VQA
	Known	Novel	Known	Novel	Novel-Q
UpDown [2]	55.65	48.53	52.73	51.35	51.40
VisualBert [33]	-	-	59.85	58.80	-
Skill-Concept [55]	-	-	-	-	59.80
XNM [48]	62.05	52.81	59.39	57.54	57.54
XNM+POEM	63.80	54.82	60.60	59.71	60.73

Table 2. Comparative results on VQA with OOD questions.

	VQA-CP	GQA-OOD	
	acc ↑	acc-tail ↑	△ ↓
Bias Product [11]	39.93	30.8	12.0
AdvReg [44]	41.17	-	-
Hint [47]	46.73	-	-
RUBi [7]	47.11	35.7	14.3
SCR [56]	49.45	-	-
LMH [11]	52.05	32.2	11.5
LMH+Entropies [15]	54.55	-	-
XNM [48]	51.54	46.14	14.4
XNM+POEM	53.99	46.89	10.7

dict all object categories that exist in an image. The image-wise ground truth is constructed with object detection labels (VQA, VQA-CP and Novel-VQA) or scene graphs (GQA and GQA-OOD). UpDown features are used as inputs to the classification network, which encode semantic information of different objects. For zero-shot VQA, we train the model on each dataset with training images that do not contain the selected novel objects, and evaluate it on validation images with only known objects. For VQA with OOD questions, we use the original training and validation sets without excluding samples. We set the number of prototypes in our experiments to 1000 (*i.e.,* $K = 1000$), which is comparable to the number of fine-grained objects in GQA (*i.e.,* ~ 1000 object categories). The network is trained with Adam [30] optimizer for 60 epochs, the learning rate and batch size are set to 4×10^{-4} and 128, respectively. Prototypes with the best validation performance are used in our VQA model.

4.2. Results on Zero-shot VQA

We first demonstrate the effectiveness of our prototypical neural module (POEM) network on answering questions with both known and novel objects. We compare it with four approaches, including (1) UpDown [2] that does not explicitly model the reasoning process, (2) Our baseline XNM [48] with an explicitly reasoning process, (3) VisualBert [33] that utilizes vision-and-language pretraining on external data containing the novel objects (*i.e.,* MSCOCO [36]), and (4) Skill-Concept [55] that is the current state-of-the-art on Novel-VQA dataset, which explicitly exploits novel objects in images with around 97% of them covered in a handcrafted reference training set. All of the models are trained with UpDown visual features and initialized with pretrained word embeddings.

We draw three key observations from the results in Table 1: (1) While Updown and XNM show similar accuracy under the standard VQA setting [2, 48], the latter provides stronger performance on zero-shot VQA. The results indicate that the compositional reasoning process is not only helpful in interpretability, but also important for model performance and generalizability; (2) By leveraging external data, VisualBert provides better performance than both aforementioned models. However, it violates the original intent of zero-shot VQA with novel objects actually covered in the external data, making them impractical for real-world scenarios with unseen objects; and (3) Differently, with our prototypes bridging different objects and the reasoning process, the proposed method improves the performance of the XNM baseline without relying on external data, and achieves overall the best results in answering questions with both known and novel objects in the visual-textual data. It also outperforms Skill-Concept on the Novel-VQA dataset that only considers novel objects in the questions. Note that Skill-Concept assumes the availability of novel objects in training images and is not applicable to zero-shot setting on VQA and GQA datasets, while our method has the capability to generalize toward broader scenarios.

4.3. Results on VQA with OOD Questions

Next, we investigate the robustness of our method against spurious data biases. We compare our method with eight approaches that do not exploit additional data, including Bias Product [11], AdvReg [44], RUBi [7], and LMH+Entropies [15] for reducing single-modal biases, Hint [47] and SCR [56] for boosting visual attention, LMH [11] for learning the residual of biases, and our baseline XNM. Following [11, 15, 20], when experimenting on the VQA-CP dataset [1], both XNM and our model incorporate the learned mixed-in module [11] to address known biases.

Results in Table 2 show that our method is able to improve the XNM baseline by a considerable margin. With object factorization and compositional reasoning, it not only improves the robustness against adversarial distributions (*i.e.,* VQA-CP), but also reduces the performance gap between answering in-distribution and out-of-distribution questions (*i.e.,* △ for GQA-OOD). Compared to existing methods that introduce additional regularization and have difficulties generalizing to different datasets [7, 11, 15], our method augments models' reasoning capability without im-

Table 3. Comparative results for different prototypes.

	VQA		GQA		Novel-VQA	VQA-CP
	Known	Novel	Known	Novel	Novel-Q	OOD
Scratch	62.66	53.66	58.87	57.69	59.48	50.39
Object	62.56	53.23	60.48	59.46	56.67	52.23
Textual	60.61	52.33	55.31	53.22	59.17	51.09
Ours	63.80	54.82	60.60	59.71	60.73	53.99

posing data-specific constrains and thus enjoys better generalizability. The aforementioned observations suggest the significance of explicitly bridging objects with their fine-grained characteristics for overcoming data biases.

4.4. Ablation Study on Object Factorization

A key component of our framework is the proposed method for learning discriminative prototypes with object factorization. In this section, we perform an ablation study to investigate the usefulness of different prototypes (see our supplementary materials for additional ablation studies on model design). Specifically, we consider three types of alternative prototypes: (1) Prototypes that are randomly initialized and learned from scratch on the VQA task (Scratch); (2) Prototypes specific to manually defined objects (Object), which are learned with the same multi-label classification task as our approach but without object factorization; and (3) Prototypes derived from Glove [43] word embeddings of concepts covered in the Visual Genome [31] dataset (Textual), including objects, attributes, and relationships.

We made three observations from results in Table 3: (1) Randomly initialized prototypes lead to inferior performance among all datasets, indicating the significance of explicitly learning semantically plausible prototypes; (2) While object-based prototypes show reasonable performance on the GQA dataset with detailed annotations (*i.e.,* $\sim$ 1000 object categories), they have negligible improvements on the VQA, Novel-VQA, and VQA-CP datasets with abstract-level annotations (*i.e.,* 80 object categories). The large gap in performance gain across datasets demonstrates the advantages of fine-grained categorization, and more importantly, highlights the need to learn discriminative prototypes without relying on extensive annotations. For this, our method utilizes object factorization to automatically decompose objects into more elaborated semantics, and brings considerable improvements among datasets with both abstract and detailed object annotations; (3) Textual prototypes result in a visible drop in accuracy, despite the consideration of various concepts. This is likely caused by the difficulty of correlating objects across the visual and textual domains. Differently, our method directly captures diverse characteristics of objects from visual data and does not suffer from the discrepancies between modalities.

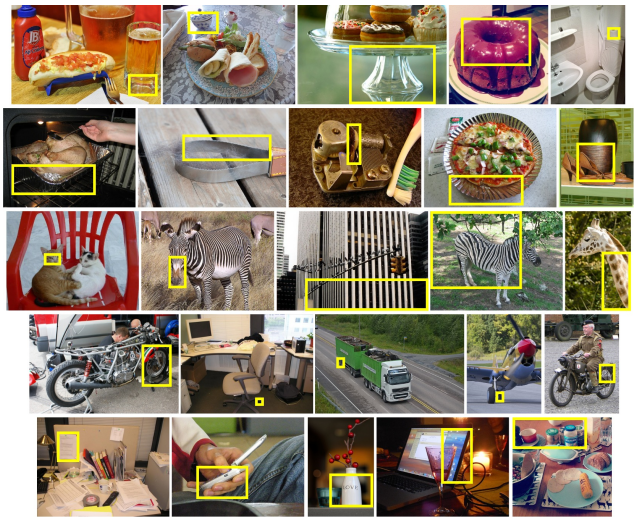

Figure 4. Examples of semantics encoded in diverse prototypes.

4.5. What Do Prototypes Learn? Do They Encode Common Characteristics among Objects?

Results in the previous sections demonstrate that our method learns discriminative prototypes to represent a variety of objects and bring enhanced generalizability across various settings. This section further demonstrates its effectiveness by investigating how our prototypes correlate different objects.

We first study what each individual prototype learns. In Figure 4, we visualize instances (regions inside bounding boxes) most relevant to the prototypes (measured with their relevance scores on different prototypes, *i.e.,* α_p^k in Equation 2, k denotes the indices of prototypes). The results show that our prototypes learn to represent a diverse pool of semantics. They not only capture low-level visual cues, such as shapes (*e.g.,* round objects in the 1^{st} row), textures (*e.g.,* objects with jagged texture in the 2^{nd} row), and patterns (*e.g.,* objects with stripes in the 3^{rd} row), but also encode high-level semantics including object categories (*e.g.,* wheels in the 4^{th} row) and commonalities in semantics (*e.g.,* all objects in the 5^{th} row are displaying text).

With our prototypes encoding abundant semantics, we further analyze their effectiveness in correlating relevant objects based on their characteristics. Specifically, we calculate the average relevance scores α_p (see Equation 1) between objects in the GQA dataset and all prototypes, and then apply k-means algorithm [37] (k=30) to cluster objects using the scores. As shown in Table 4, by representing objects based on the prototypes, we can correlate objects that belong to similar categories (*e.g.,* drinks and utensils in the 1^{st} and 2^{nd} rows), commonly appear in the same scenarios (*e.g.,* baseball games and bedrooms in the 3^{rd} and 4^{th} rows)

Table 4. Different groups of objects that are clustered based on their relevance to prototypes. Please refer to our supplementary materials for the complete results with 30 groups.

Group	Objects
1	cup, saucer, glass, beer, mug, juice, beverage, liquid, smoothie, coffee
2	fork, spoon, knife, silverware, utensil, ladle, chopstick, tongs, spatula, butter knife
3	spectator, umpire, batter, catcher, crowd, net, player, baseball, stadium, bleachers
4	bed, sofa, pillow, bedspread, headboard, comforter, couch, sheet, ottoman, mattress
5	sticker, newspaper, paper, sign, book, tape, drawing, CD, letter, label
6	water, sand, sea, rock, ocean, boulders, lake, beach, shore, river

or share similar characteristics (*e.g.,* objects related to text and landscape in the 5^{th} and 6^{th} rows). The results demonstrate the effectiveness of our prototypes in parsing objects based on their commonalities in the semantic space.

4.6. How Do Models Reason to Answer Questions?

Besides improving the generalizability, our method also enables interpretation of the decision-making process by visualizing the regions of interest (ROIs) at each reasoning step and prototypes matched with the observations. In this section, we provide qualitative examples to study the underlying rationales behind the derivation of answers.

Figure 5 shows the reasoning process of our method. For each question, we visualize the reasoning steps represented by neural modules (top), attention maps highlighting the ROIs (middle, α^t in Equation 6), and images associated to prototypes matched with the observations (bottom). It shows that our method can correlate objects based on various characteristics and locate those important ones. In the 1^{st} example, while the apples are not explicitly mentioned in the question, our model correlates them with prototypes for different fruits ("fruit" is a keyword in the question) and pays focused attention in the *Find* step. Besides capturing semantic relationships about object categories, prototypes also help identify the cat based on its attribute (*i.e.,* brown color) within the *Filter* step, and enable the model to reason on the relative position between objects (*i.e., Relate* step). In the 2^{nd} example, our model not only highlights the correct objects in the first two *Find* steps (*i.e.,* the banana and the mat), but also identifies the key characteristics that contribute to reasoning (*i.e.,* matching observations with prototypes related to colors instead of object categories or other attributes). In the 3^{rd} example, woman in the question is an unseen object during training. With object factorization,

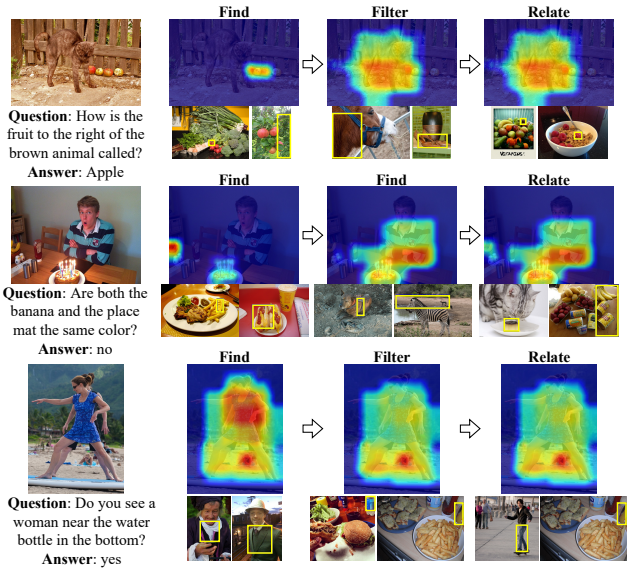

Figure 5. Illustrations of the reasoning process. From left to right are the input images together with questions and predicted answers, and sequences of reasoning steps.

our model successfully associates it to known objects (*i.e.*, men) based on the similarity in clothing, and progressively shifts the attention from both women (*i.e., Find* step) to the one lying next to the bottle (*i.e., Filter* and *Relate* steps).

5. Conclusion

This study is an effort toward generalizable and interpretable AI systems for real-world applications. It draws inspiration from the ways humans reason with the visual world, and investigates the effectiveness of integrating the compositionality of objects and the reasoning process. Our work distinguishes itself with a principled method for automatically factorizing objects into fine-grained semantics, bridging novel and known objects, and a new neural module network with a compositional decision-making process. The compositionality in both dimensions addresses object diversity and spurious data biases, enhancing model generalizability toward a broad range of scenarios. It also enables interpretation of the rationales behind the model's decisions. Experimental results demonstrate the advantages of our method under diverse settings, and provide insights on how our model reasons with the visual-textual inputs. We hope that this study can be useful for future developments of trustworthy visual reasoning models with more human-like intelligence and generalizability.

Acknowledgements

This work is supported by NSF Grants 2143197 and 2227450.

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 *CVPR*, pages 4971–4980, 2018. 3, 5, 6

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086, 2018. 1, 2, 4, 5, 6

[3] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *CVPR*, pages 39–48, 2016. 2, 4

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *ICCV*, pages 2425–2433, 2015. 2, 5

[5] Aristotle. *Aristotle's Prior Analytics*. Oxford University Press, 1989. 1

[6] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marcal Rusinol, Ernest Valveny, C.V. Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *ICCV*, pages 4290–4300, 2019. 2

[7] Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. *NeurIPS*, pages 841–852, 2019. 3, 6

[8] Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. This looks like that: Deep learning for interpretable image recognition. In *NeurIPS*, page 8930–8941, 2019. 1

[9] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *CVPR*, pages 10797–10806, 2020. 3

[10] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *CVPR*, pages 5172–5181, 2019. 3

[11] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *EMNLP*, pages 4069–4082, 2019. 3, 6

[12] Moshiur R. Farazi, Salman H. Khan, and Nick Barnes. From known to the unknown: Transferring knowledge to answer questions about novel visual and semantic concepts. In *Image and Vision Computing*, volume 103, page 103985, 2020. 3, 4, 5

[13] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*, pages 457–468, 2016. 1, 2

[14] Tianyu Gao, Xu Han, Zhiyuan Liu, and Maosong Sun. Hybrid attention-based prototypical networks for noisy few-shot relation classification. In *AAAI*, page 6407–6414, 2019. 3

[15] Itai Gat, Idan Schwartz, Alexander Schwing, and Tamir Hazan. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. In *NeurIPS*, pages 3197–3208, 2020. 3, 6

[16] Donald Geman, Stuart Geman, Neil Hallonquist, and Laurent Younes. Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences*, 112(12):3618–3623, 2015. 2

[17] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. MUTANT: A training paradigm for out-of-distribution generalization in visual question answering. In *EMNLP*, pages 878–892, 2020. 3

[18] 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 *CVPR*, pages 6325–6334, 2017. 2, 5

[19] Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *Arxiv*, 2020. 3

[20] Xinzhe Han, Shuhui Wang, Chi Su, Qingming Huang, and Qi Tian. Greedy gradient ensemble for robust visual question answering. In *ICCV*, pages 1564–1573, 2021. 3, 6

[21] Ronghang Hu, Jacob Andreas, Trevor Darrell, and Kate Saenko. Explainable neural computation via stack neural module networks. In *ECCV*, pages 55–71, 2018. 1, 2, 4, 5

[22] Drew Hudson and Christopher D Manning. Learning by abstraction: The neural state machine. In *NeurIPS*, volume 32, 2019. 2

[23] Drew Arad Hudson and Christopher D. Manning. Compositional attention networks for machine reasoning. In *ICLR*, 2018. 2

[24] Drew A. Hudson and Christopher D. Manning. GQA: a new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6693–6702, 2019. 2, 5

[25] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, pages 1988–1997, 2017. 2

[26] P. N. Johnson-Laird. Deductive reasoning. *Annual Review of Psychology*, 50(1):109–135, 1999. 1

[27] Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. Roses are red, violets are blue... but should vqa expect them to? In *CVPR*, pages 2776–2785, 2021. 3, 5

[28] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *NeurIPS*, pages 1571–1581, 2018. 1, 2

[29] Jin-Hwa Kim, Kyoung Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. In *ICLR*, 2017. 1, 2

[30] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[31] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalanditis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *IJCV*, volume 123, pages 32–73, 2017. 7

[32] Hugo Larochelle, Dumitru Erhan, and Y. Bengio. Zero-data learning of new tasks. In *AAAI*, volume 2, pages 646–651, 2008. 3

[33] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. What does BERT with vision look at? In *ACL*, pages 5265–5275, 2020. 1, 2, 4, 6

[34] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, pages 121–137, 2020. 1, 2, 5

[35] Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. Learning to contrast the counterfactual samples for robust visual question answering. In *EMNLP*, pages 3285–3292, 2020. 3

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 6

[37] S. Lloyd. Least squares quantization in pcm. In *IEEE Transactions on Information Theory*, volume 28, pages 129–137, 1982. 7

[38] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, volume 32, pages 13–23, 2019. 1, 2

[39] Varun Manjunatha, Nirat Saini, and Larry S. Davis. Explicit bias discovery in visual question answering models. In *CVPR*, pages 9554–9563, 2019. 1

[40] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, pages 3190–3199, 2019. 2

[41] Yingwei Pan, Ting Yao, Yehao Li, Yu Wang, Chong-Wah Ngo, and Tao Mei. Transferrable prototypical networks for unsupervised domain adaptation. In *CVPR*, pages 2234–2242, 2019. 3

[42] Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. Visualcomet: Reasoning about the dynamic context of a still image. In *ECCV*, pages 508–524, 2020. 2

[43] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014. 2, 5, 7

[44] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. In *NeurIPS*, pages 1548–1558, 2018. 3, 6

[45] Santhosh K. Ramakrishnan, Ambar Pal, Gaurav Sharma, and Anurag Mittal. An empirical evaluation of visual question answering for novel objects. In *CVPR*, pages 7312–7321, 2017. 2, 4, 5

[46] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *ICCV*, pages 82–91, 2021. 3

[47] Ramprasaath R. Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. Taking a HINT: Leveraging explanations to make vision and language models more grounded. In *ICCV*, pages 2591–2600, 2019. 3, 6

[48] Jiaxin Shi, Hanwang Zhang, and Juanzi Li. Explainable and explicit visual reasoning over scene graphs. In *CVPR*, pages 8368–8376, 2019. 1, 2, 4, 5, 6

[49] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *EMNLP*, pages 5100–5111, 2019. 1, 2, 4, 5

[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. In *NeurIPS*, pages 407–417, 2020. 3

[51] Damien Teney and Anton van den Hengel. Zero-shot visual question answering. *Arxiv*, 2016. 2, 5

[52] Damien Teney and Anton van den Hengel. Visual question answering as a meta learning task. In *ECCV*, pages 229–245, 2018. 3, 4

[53] Harm De Vries, Dzmitry Bahdanau, Shikhar Murty, Aaron C. Courville, and Philippe Beaudoin. CLOSURE: assessing systematic generalization of CLEVR models. In *NeurIPS Workshop*, 2019. 2

[54] Peng Wang, Qi Wu, Chunhua Shen, and Anton van den Hengel. The vqa-machine: Learning how to use existing vision algorithms to answer new questions. In *CVPR*, pages 3909–3918, 2017. 2

[55] Spencer Whitehead, Hui Wu, Heng Ji, Rogerio Feris, and Kate Saenko. Separating skills and concepts for novel visual question answering. In *CVPR*, pages 5628–5637, 2021. 2, 4, 5, 6

[56] Jialin Wu and Raymond J. Mooney. Self-critical reasoning for robust visual question answering. In *NeurIPS*, page 8604–8614, 2019. 3, 6

[57] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *ICCV*, pages 1839–1848, 2017. 1, 2

[58] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, pages 6713–6724, 2019. 2

[59] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2014. 3, 5

[60] Yifeng Zhang, Ming Jiang, and Qi Zhao. Explicit knowledge incorporation for visual reasoning. In *CVPR*, pages 1356–1365, 2021. 5

[61] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. Learning spatial regularization with image-level supervisions for multi-label image classification. In *CVPR*, pages 2027–2036, 2017. 3, 5