

VCD: A Dataset for Visual Commonsense Discovery in Images

Xiangqing Shen, Fanfan Wang, Siwei Wu, and Rui Xia*

School of Computer Science and Engineering,
Nanjing University of Science and Technology, China
{xiangqing.shen, ffwang, wusiwei, rxia}@njjust.edu.cn

Abstract

Visual commonsense plays a vital role in understanding and reasoning about the visual world. While commonsense knowledge bases like ConceptNet provide structured collections of general facts, they lack visually grounded representations. Scene graph datasets like Visual Genome, though rich in object-level descriptions, primarily focus on directly observable information and lack systematic categorization of commonsense knowledge. We present Visual Commonsense Dataset (VCD), a large-scale dataset containing over 100,000 images and 14 million object-commonsense pairs that bridges this gap. VCD introduces a novel three-level taxonomy for visual commonsense, integrating both Seen (directly observable) and Unseen (inferred) commonsense across Property, Action, and Space aspects. Each commonsense is represented as a triple where the head entity is grounded to object bounding boxes in images, enabling scene-dependent and object-specific visual commonsense representation. To demonstrate VCD's utility, we develop VCM, a generative model that combines a vision-language model with instruction tuning to discover diverse visual commonsense from images. Extensive evaluations demonstrate both the high quality of VCD and its value as a resource for advancing visually grounded commonsense understanding and reasoning. Our dataset and code will be released on <https://github.com/NUSTM/VCD>.

1 Introduction

Commonsense, comprising facts and principles humans rely on in daily life, is essential for decision-making and behavior. Integrating it into AI systems enhances human-like reasoning, improves interpretability, and has become a growing area of research. Visual commonsense, a portion of commonsense, refers to general knowledge about the

visual world. While existing commonsense knowledge bases, *e.g.*, ConceptNet (Speer et al., 2017), include some visual commonsense represented in textual form, they lack visually grounded commonsense—specific, contextually rich knowledge tied to actual visual scenes. Such limitation results in restricted coverage and insufficient detail for effectively bridging vision and language understanding. Cognitive science research (Kahneman et al., 1992) indicates that humans perceive the world by focusing on objects in a scene, noting their attributes, spatial relationships, and actions to gather multidimensional information. This information forms the basis of visual commonsense, which is inherently scene-dependent and object-specific.

On the other hand, scene graph datasets in the field of computer vision, *e.g.*, Visual Genome (VG) (Krishna et al., 2017), although provide rich object-level descriptions of attributes, actions, and relationships, typically lack a systematic categorization of commonsense. Moreover, they predominantly focus on commonsense directly observable in images (referred to as *seen commonsense* in this paper), while neglecting commonsense not visually apparent but still relevant to the image and can be inferred by general world knowledge (referred to as *unseen commonsense*). For example, in Fig. 1, given a scene depicting “a man skateboarding on a busy street”, humans can naturally infer unseen commonsense like “the man might be hit by a car”. Such unseen commonsense is crucial for deep visual understanding and reasoning, but has received insufficient attention in current research.

To address these challenges, we present VCD, a large-scale Visual Commonsense Dataset by integrating and linking Visual Genome and ConceptNet. VCD includes over 100,000 images with more than 14 million object-commonsense pairs, where each image is annotated with objects it contains, and each object is further annotated with its related visual commonsense triples. Similar to Concept-

*Corresponding author.

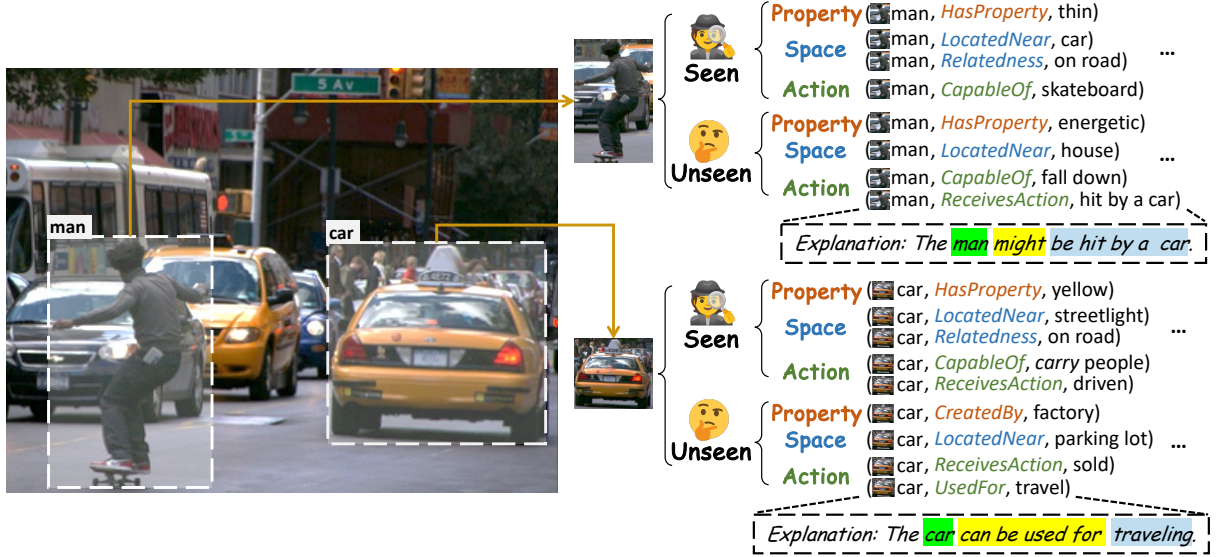


Figure 1: An example from VCD. Given the left image, two objects (a man and a car) are annotated along with their associated 11 visual commonsense triples. These triples are organized within a hierarchical taxonomy. For example, (man, LocatedNear, car) is a Seen commonsense under the Space aspect, (man, ReceivesAction, hit by a car) is an Unseen commonsense under the Action aspect.

Net, each commonsense is represented as a (head, relation, tail) triple; but the head here is a language-vision pair, consisting of an entity and its corresponding bounding box in the current image. By grounding commonsense entities to the bounding boxes in the image, VCD bridges the gap between linguistic knowledge and visual information.

We introduce a three-layer taxonomy to categorizing these visual commonsense triples. First, we identify visually relevant categories from the 34 basic knowledge types defined in ConceptNet as the foundational layer. These categories are then grouped into three fundamental aspects widely studied in computer vision (*i.e.*, Property, Action, and Space), constituting the second layer. For the top layer, we distinguish commonsense knowledge based on its visual observability in the given image (*i.e.*, Seen commonsense versus Unseen commonsense). This hierarchical taxonomy provides a comprehensive framework for organizing visual commonsense knowledge, bridging NLP and CV domains while enabling analysis of both observable and inferential visual relationships.

VCD captures rich patterns and relationships that reflect the visual world, enabling the discovery of scene-dependent and object-specific visual commonsense. In this regard, we train a generative model, VCM, that integrates a vision-language model with instruction tuning, to generate visual commonsense from images. The instructions cover

diverse types of commonsense within the taxonomy, enabling VCM to generate different categories of commonsense triples according to the provided instruction, spanning both Seen and Unseen visual commonsense across the Property, Action, and Space aspects.

Extensive evaluations, including both automatic and human evaluations, demonstrate 1) the high quality of the VCD dataset, 2) the strong performance in visual commonsense discovery, particularly surpassing GPT-4o in identifying unseen commonsense, and 3) the enhancement of downstream vision-language tasks through the discovered visual commonsense knowledge. These comprehensive evaluations demonstrate VCD’s value as a foundational resource for discovering and leveraging visual commonsense, advancing visually-grounded commonsense AI.

2 Related Work

Commonsense in text has been a longstanding research focus, with early studies primarily dedicated to constructing commonsense knowledge bases. ConceptNet (Speer et al., 2017) integrates multiple knowledge bases. ASER (Zhang et al., 2020b) captures selectional preference knowledge extracted from over 11 billion tokens of unstructured text. TransOMCS (Zhang et al., 2020a) employs linguistic graphs to align ASER with ConceptNet. DISCOS (Fang et al., 2021) enhances commonsense of

ASER by aggregating information from neighboring concepts. ATOMIC (Sap et al., 2019; Hwang et al., 2021; Shen et al., 2023) is a collection of if-then knowledge triplets centered on daily events.

Research on visual commonsense has evolved from focusing on the specific commonsense category to broader, more generalized commonsense categories. Early work focused on different specific dimensions of images, including taxonomy (Chen et al., 2013), unary affordance (Chao et al., 2015), physical properties (Zellers et al., 2021; Tang et al., 2023), and spatial relationships (Yatskar et al., 2016; Xu et al., 2018; Collell et al., 2018; Diomataris et al., 2021). More recent studies have expanded beyond these individual dimensions to explore generalized visual commonsense knowledge (Vedantam et al., 2015; Chen et al., 2022; Liu et al., 2022; Zellers et al., 2019; Zhang et al., 2022; Li et al., 2023; Singh et al., 2023; Xia et al., 2023). A crucial aspect of visual commonsense is the generation of scene graphs, which models object interactions within an image to support high-level reasoning (Krishna et al., 2017; Yu et al., 2017). Another important research direction involves multimodal knowledge graphs (Oñoro-Rubio et al., 2017; Ferrada et al., 2017; Liu et al., 2019; Alberts et al., 2020; Wang et al., 2020), which extend traditional knowledge graphs by associating entities with non-textual data, such as images. However, no existing multimodal knowledge graphs are explicitly designed to capture visual commonsense. We distinguish visual recognition from visual commonsense, aligning the latter’s “seen” aspects with ViCor’s (Zhou et al., 2024) Visual Commonsense Understanding (VCU). While visual recognition identifies objects and attributes (e.g., a “man,” “thin”), VCU, or “seen” commonsense, provides an explicit, structured understanding of this literal visual content, such as “(man, HasProperty, thin)” or “Person washing dishes”. Therefore, visual commonsense leverages visual recognition to build a structured, queryable layer of knowledge about directly observable elements and their explicit relationships within a scene, forming a foundational step towards deeper reasoning.

3 Visual Commonsense Dataset Construction



3.1 Preliminary Resources

ConceptNet (Speer et al., 2017) is a multilingual commonsense knowledge base that comprises a

vast collection of manually curated triples, each representing words or phrases and their commonsense relationships. It systematically defines 34 categories of commonsense and encompasses more than 4 million English triples. However, its textual representation limits its ability to effectively capture scene-dependent and object-specific visual commonsense, which is crucial for understanding real-world contexts.

Visual Genome (VG) (Krishna et al., 2017) is a large-scale scene graph dataset containing 108,077 images with dense annotations of 5.4 million region descriptions, 3.8 million object instances, 2.8 million attributes, and 2.3 million relations. Despite its extensive coverage, VG lacks a structured categorization of visual commonsense and does not include unseen commonsense.

3.2 Visual Commonsense Taxonomy

We introduce a three-layer hierarchical taxonomy to organize the visual commonsense. At first, we identify visually relevant categories from 34 basic knowledge types defined in ConceptNet, establishing the foundational layer. These categories are then grouped into three fundamental aspects commonly employed in computer vision (*i.e.*, Property (Tang et al., 2023), Action (Chao et al., 2015), and Space (Collell et al., 2018)), forming the second layer. At the top layer, we introduce a visibility dimension that classifies knowledge as seen or unseen commonsense. Seen commonsense contains directly observable commonsense from images, While unseen commonsense involves inferred commonsense requiring contextual reasoning or life experiences. This results in a hierarchical visual commonsense taxonomy. Taking the image in Fig. 1 as an example,  (man, *LocatedNear*, car) is a Seen commonsense under the Space aspect,  (man, *ReceivesAction*, hit by a car) is an Unseen commonsense under the Action aspect. More details of the visual commonsense taxonomy can be found in App. A.1.

3.3 Seen Commonsense Annotation

VG encompasses a diverse range of real-world scenes, enriched with detailed annotations including object-level triples and region-level phrases, each accompanied by a bounding box. These annotations make VG a valuable resource for capturing a broad spectrum of seen commonsense about various entities in an image by processing its existing object-level triples and region-level phrases.

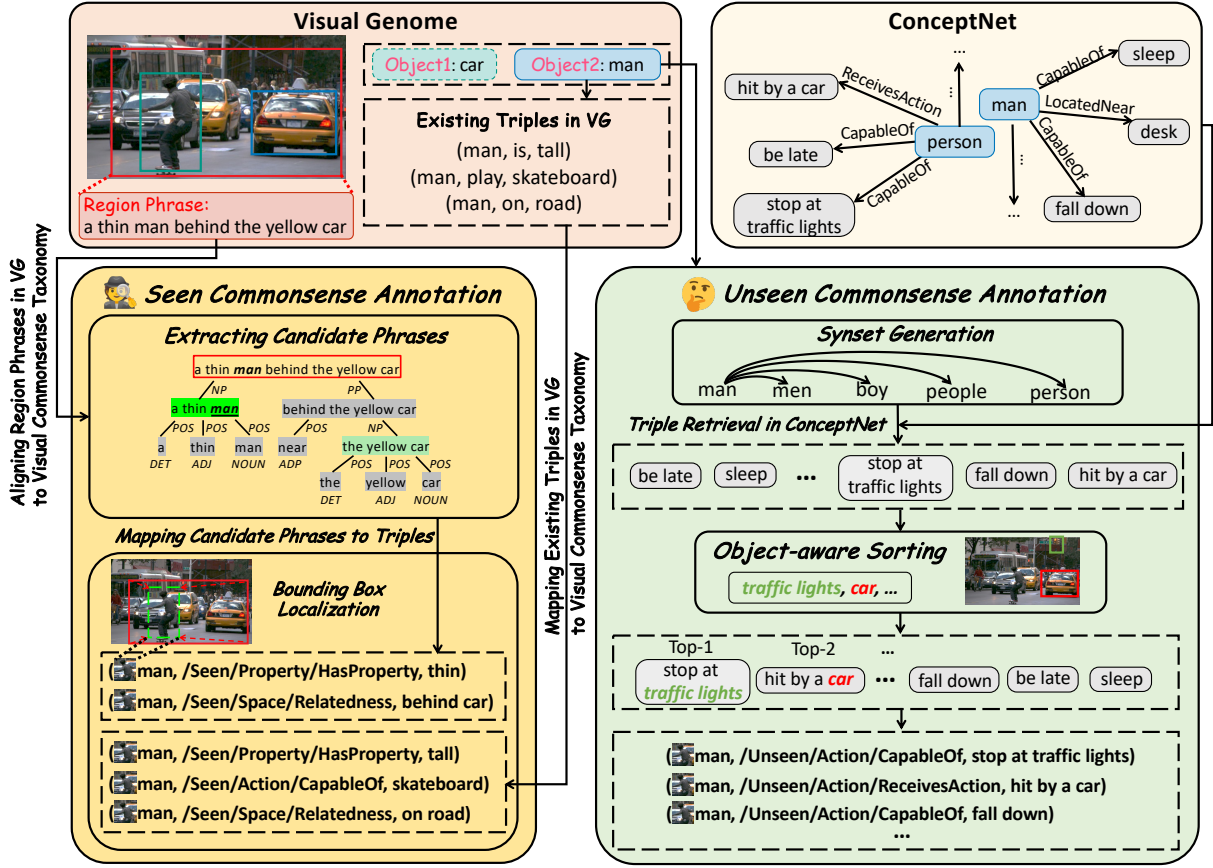


Figure 2: The construction process of VCD.

Therefore, our approach to annotate seen commonsense is to map existing triples in VG to our visual commonsense taxonomy, and align region phrases in VG with the visual commonsense taxonomy.

3.3.1 Mapping Existing Triples in VG to Visual Commonsense Taxonomy

Fig. 2 shows that VG includes objects marked with bounding boxes and annotated with descriptive triples. These triples encapsulate visible attributes, actions, and spatial relationships. For instance, (man, is, tall) represents an attribute, while (man, play, skateboard) and (man, on, road) represent actions and spatial relationships, respectively.

To map these triples to seen commonsense categories, we establish part-of-speech based mapping rules. Verbs are mapped to /Seen/Action, adjectives to /Seen/Property, and prepositions to /Seen/Space. For example, (man, is, tall), as shown in Fig. 2, is mapped to (man, /Seen/Property/HasProperty, tall) with the recognition of *tall* as an *adjective*.

Furthermore, /Seen/Space/LocatedNear captures co-occurrence relationships, implying that two entities often appear together within the same visual

scenario without a specific spatial relation. For instance, since “man” and “car” co-occur in Fig. 2, one could infer a seen commonsense triple (man, /Seen/Space/LocatedNear, car).

3.3.2 Aligning Region Phrases in VG to Visual Commonsense Taxonomy

Only mapping existing triples in VG may lead to omissions. However, region phrases in VG can supplement these triples. For example, consider a region phrase “a thin man behind the yellow car” from VG in Fig. 2. This phrase implicitly contains several seen commonsense triples that are missing from the existing triples, such as (man, /Seen/Space/Relatedness, behind car), (man, /Seen/Property/HasProperty, thin). Consequently, we extract additional triples from region phrases using an automatic process. A manual review is then conducted to ensure the reliability and accuracy of the extracted triples.

Extracting Candidate Phrases We begin by applying constituency parsing to region phrases to extract candidate phrases, including preposition, verb, and noun phrases. For example, given the

Table 1: Comparison with other visual commonsense datasets. # Categories represents the number of visual commonsense categories included in each dataset.

	Seen	Unseen	# Categories	# Images	# BBox	# Commonsense
ConceptNet (Speer et al., 2017)	✓	✓	34	✗	✗	≈4M
Visual Genome (Krishna et al., 2017)	✓	✗	✗	≈106K	≈4.1M	≈5M
SpatialCS (Liu et al., 2022)	✓	✗	1	✗	✗	1224
ViComTe (Zhang et al., 2022)	✓	✗	1	✗	✗	11114
VEC (Li et al., 2023)	✓	✗	2	✗	✗	4090
VIPHY (Singh et al., 2023)	✓	✗	2	✗	✗	≈30K
ImageNetVC (Xia et al., 2023)	✓	✗	2	✗	✗	4976
VCD	✓	✓	11	≈106K	≈2.4M	≈14M

region phrase “*a thin man behind the yellow car*” in Fig. 2, constituency parsing classifies the entire phrase as a prepositional phrase, as well as “*a thin man*” and “*the yellow car*” as noun phrases.

Mapping Candidate Phrases to Triples Upon candidate phrases, dependency parsing is used to determine their syntactic structure. Then, for each type of candidate phrase, we use a carefully-defined set of mapping rules to map syntactic structures to commonsense triples. To illustrate, for the noun phrase “*a thin man*” where “man” is the *root*, we apply the mapping rule “*adjective + noun* → (noun, /Seen/Property/HasProperty, adjective)” to yield the triple (man, /Seen/Property/HasProperty, thin). Full set of mapping rules is in App. A.2.

Bounding Box Localization To determine the bounding boxes for objects in triples extracted from regional phrases, we first compute the overlap ratio between the given region and the annotated bounding boxes. We then filter out boxes with an overlap ratio below a predefined threshold, retaining only those that meet or exceed this criterion to form a candidate set. Next, we match object names from the triples to the candidate set, preserving only those triples that have a unique correspondence, while discarding those with multiple matches or no valid match. The resulting triples are then used to enhance seen commonsense triples of objects.

3.4 Unseen Commonsense Annotation

Unseen commonsense is essential for visual commonsense reasoning beyond direct visual perception. While Visual Genome (VG) focuses on seen commonsense directly observable in images, it lacks annotations for unseen commonsense that is not visually present. ConceptNet complements this by providing a rich source of unseen common-

sense.

Therefore, our approach to annotating unseen commonsense is to extract relevant knowledge triples from ConceptNet that correspond to objects in the image, serving as unseen commonsense knowledge for the given image. The process consists of the following steps:

Synset Generation To enhance the coverage of unseen commonsense retrieved, we first lemmatize the names of objects in VG for synset generation, as shown in Fig. 2.

Triple Retrieval in ConceptNet Using the generated synset, we retrieve ConceptNet for unseen commonsense for each object, ensuring a comprehensive collection of unseen commonsense associated with each identified object in an image.

Object-aware Sorting Humans naturally consider all objects within a scene when making associations. As shown in Fig. 2, an image of a man skateboarding alongside many cars may evoke unseen commonsense that the man is at risk of being hit by a car. This connection between “man” and “car” arises from their co-occurrence in the image. Building on this cognitive process, we prioritize unseen commonsense that involves objects present in the image, as they are more intuitively derived from the visual context. This object-aware sorting strategy ensures that retrieved commonsense aligns more closely with human reasoning process.

3.5 Dataset Statistics

Tab. 1 provides a comprehensive comparison of our dataset (VCD) with existing visual commonsense datasets. Unlike most datasets, VCD integrates both seen and unseen commonsense, exhibiting distinct advantages in coverage, diversity, and scale.

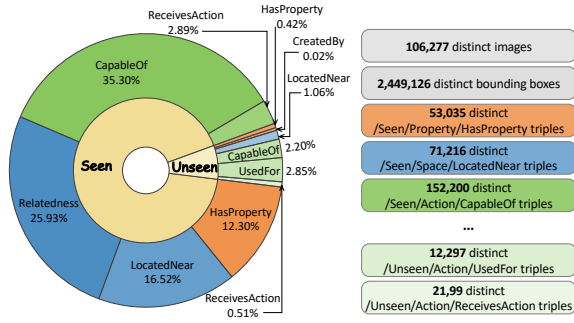


Figure 3: The statistics of VCD.

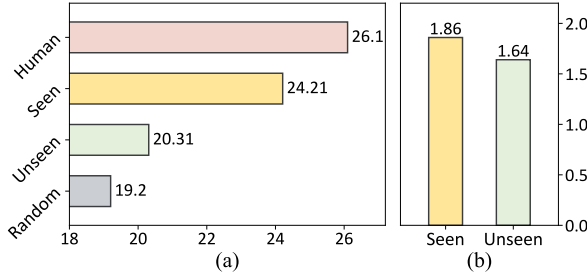


Figure 4: (a) Automatic evaluation using CLIP similarity scores; (b) Human evaluation with Likert scale ratings.

Fig. 3 illustrates that VCD consists of 106,277 images and 2,449,126 bounding boxes, encompassing 18,136 unique object names. Furthermore, Fig. 3 presents the distribution of distinct commonsense triples across various categories, along with their respective proportions within VCD. Examples of VCD are provided in App. A.3.

3.6 Dataset Quality Control

VCD is built upon ConceptNet and VG, both of which are high-quality, manually annotated resources. While these foundations provide reliable base data, our work focuses on linking and aligning these resources. Therefore, we first evaluate the performance of our linking and alignment processes, followed by both automatic and human evaluations of the final annotations.

Evaluation of the Off-the-shelf Annotation Tools

We utilize spaCy (Honnibal et al., 2020) for part-of-speech tagging and dependency parsing on region phrases in VG. A human evaluation of 200 samples confirms a 99% accuracy rate. Similarly, for constituency parsing, AllenNLP (Gardner et al., 2018) achieves 97.5% accuracy based on human evaluation of 200 samples.¹

¹The high accuracy of both spaCy and AllenNLP is largely due to simple linguistic structure of region phrases in VG.

Evaluation of the Iterative Annotation Process

To establish the rule set in Sec. 3.3.2, we follow an iterative annotation process. In each iteration, 200 samples are examined. If the error rate exceeds 5%, the rules are refined and reassessed. This process repeats until the error rate falls below 5%, ensuring that VCD adheres to rigorous quality standards.

Automatic Evaluation of the Annotated Commonsense

Following Gadre et al. (2023); Schuhmann et al. (2022), we evaluate the quality of VCD by calculating image-commonsense matching scores using the CLIP model. As shown in Fig. 4(a), we derive a lower bound by randomly sampling commonsense triples for an image and computing the match score. For the upper bound, we manually annotate ground truth seen commonsense for 300 images and calculate their matching scores. Fig. 4(a) shows that the seen commonsense matching score is slightly below the upper bound, indicating the high quality of VCD. The unseen commonsense matching score is only slightly higher than the lower bound, as these scores pertain to objects within the image but do not match the image semantically.

Human Evaluation of the Annotated Commonsense

In addition to automatic evaluation, we perform human evaluation on 4,000 randomly selected images using a 0-2 Likert scale (higher is better), with ratings provided by undergraduate students working on vision-language learning. Following (Ouyang et al., 2022), we first assess their agreement with researcher-labeled examples and select the 10 evaluators with the highest agreement scores. Fleiss’s Kappa of 0.804 indicates a good agreement. As shown in Fig. 4(b), seen commonsense receives high preference, aligning with CLIP scores. While unseen commonsense has lower CLIP scores, evaluators still favor it, indicating that it effectively reflects commonsense not depicted in the image.

4 Visual Commonsense Discovery and Its Evaluation

VCD provides a comprehensive foundation for discovering visual commonsense from images, capturing rich patterns and relationships that enable the application of such knowledge to enhance downstream VL tasks. Leveraging this dataset, we train VCM, a **V**isual **C**ommonsense **D**iscovery **M**odel that combines a generative VL architecture with instruction tuning. This allows VCM to generate visual

commonsense from novel images. The generated visual commonsense can be used to improve the performance of downstream VL applications.²

4.1 Training a Visual Commonsense Discovery Model

Given an image I , an object o_i with a bounding box, and a commonsense category r_k , VCM aims to generate a set of m commonsense triples T_i^k :

$$T_i^k = \{t_1, t_2, \dots, t_m\} = \text{VCM}(I, o_i, r_k), \quad (1)$$

where $o_i \in \{o_1, \dots, o_j\}$ represents the set of objects identified within I (each annotated with a bounding box), $r_k \in \{r_1, \dots, r_l\}$ represents the set of all types of visual commonsense, $t_m = (o_i, r_k, c)$ is a commonsense triple, and c is in the form of nouns, adjectives, or phrases.

Additionally, we iterate over each object $o_i \in \{o_1, \dots, o_j\}$ and each type of visual commonsense $r_k \in \{r_1, \dots, r_l\}$, in order to discover a comprehensive set of commonsense triples \mathcal{T} :

$$\mathcal{T} = \bigcup_{i=1}^j \bigcup_{k=1}^l T_i^k = \{T_1^1, \dots, T_1^l, T_2^1, \dots, T_j^l\}. \quad (2)$$

The input of VCM comprises an image, the name of an object with a bounding box, and a category of visual commonsense to discover. As shown in Fig. 5, these versatile elements are integrated into one instruction template using instruction tuning methods (Dai et al., 2023; Xu et al., 2023). The output of VCM is a series of commonsense triples generated in an autoregressive manner. VCM aims to minimize the following loss function:

$$\mathcal{L} = - \sum_{i=1}^{|y|} \log P_{\theta}(y_i | y_{<i}, x), \quad (3)$$

where θ denotes model parameters, x represents the instruction and y denotes the commonsense. The training details are provided in App. B

4.2 Evaluation of Visual Commonsense Discovery

4.2.1 Evaluation Protocol

Automatic Evaluation Metrics For the automatic evaluation of VCM’s visual commonsense discovery capabilities, we employ metrics for natural language generation: BLEU-1 (B-1) (Papineni

²As this paper primarily focuses on VCD construction, we present a concise version of this part here due to space limitations, with more detailed descriptions in the Appendix.

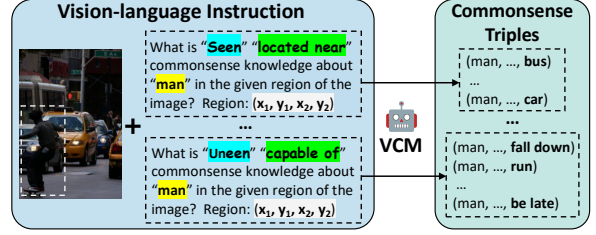


Figure 5: The input and output in VCM.

Table 2: Automatic evaluation results of VCM.

Model	B-1↑	B-2↑	R↑	M↑	W↓
SPHINX	6.4	2.0	7.6	6.2	170.5
Qwen-VL	9.4	3.0	11.5	10.2	120.5
GPT-4o	15.1	6.3	19.5	16.0	100.7
OFA _{large}	1.1	0.3	2.0	1.2	100.8
VCM _{tiny}	41.3	31.5	44.8	37.8	78.3
VCM _{medium}	48.8	37.6	52.4	45.2	73.5
VCM _{base}	53.9	42.4	56.8	50.1	71.0
VCM _{large}	56.6	45.6	59.9	53.3	67.1
w/o image	51.3	43.7	56.4	51.6	72.5
w/o region	51.5	42.6	55.4	47.7	69.4
w/o name	39.9	28.9	44.3	36.2	91.5

et al., 2002), BLEU-2 (B-2), Rouge-L (R) (Lin, 2004), METEOR (M) (Banerjee and Lavie, 2005), and Word Error Rate (W) (Su et al., 1992).

Human Evaluation Metrics While automatic evaluation metrics provide preliminary insights, they may not fully capture the diversity and quality perceived by humans. Therefore, we conduct a human evaluation, focusing on the correctness and completeness of commonsense generated by VCM. For each type of visual commonsense, 10 images are randomly selected. Their associated outputs are compared against those produced by multimodal large language models. Two independent evaluators analyze the results, and any disagreements are resolved by a third. The evaluations are structured as win/draw/lose comparisons. Fleiss’ kappa scores reveal moderate agreement among evaluators. Evaluators receive salary at a rate of \$8 per hour, which is above the local average wage.

4.2.2 Automatic Evaluation

Tab. 2 summarizes the results of the automatic evaluation. Upon analyzing the results in Tab. 2, we can find that VCM exhibits improvements in all automatic evaluation metrics as the model scale increases, which is consistent with our expectations.

Table 3: Comparison with MLLMs by human evaluation. W.R., D.R. and L.R. represent rate of win, draw and lose, respectively.

VCM_{large}	Seen			Unseen		
	W.R.	D.R.	L.R.	W.R.	D.R.	L.R.
vs. OFA_{large}	100%	0%	0%	100%	0%	0%
vs. VCM_{base}	34%	56%	10%	36%	62%	2%
vs. SPHINX	68%	14%	18%	91%	7%	2%
vs. Qwen-VL	59%	29%	12%	86%	11%	3%
vs. GPT-4o	28%	42%	30%	41%	29%	30%

Particularly, VCM_{tiny} shows a significant decrease in performance, highlighting the importance of the model scale for VCD.

4.2.3 Human Evaluation

The results of the human evaluation are summarized in Tab. 3. VCM_{large} consistently outperforms VCM_{base} across both seen and unseen commonsense discovery, aligning with the findings of automatic evaluation and reinforcing the reliability of automated metrics. Furthermore, VCM_{large} significantly outperforms SPHINX, which struggles to adhere to instructions, particularly in identifying and generating unseen commonsense.

While GPT-4o demonstrates strong performance in generating diverse examples of seen commonsense, it underperforms in unseen commonsense discovery, often confusing seen and unseen commonsense. This suggests that GPT-4o faces challenges in distinguishing and generating them as distinct categories. In contrast, VCM_{large} exhibits a stronger ability to associate unseen commonsense with image objects, leading to more precise unseen commonsense discovery.

4.2.4 Ablation Study

To evaluate the impact of images, bounding boxes, and object names on VCM’s performance, we conduct ablation studies. Results in Tab. 2 show that removing any of these elements degrades performance, with the removal of object names causing the most significant drop, even below VCM_{base} . This highlights the critical role of textual information from object names.

4.3 Evaluation on Vision-language Tasks

We further evaluate the effectiveness of visual commonsense discovered by VCM on two downstream VL tasks. The first is a dedicated evaluation of a model’s visual commonsense capabilities,

Table 4: Visual commonsense capacity on ImageNetVC.

	COL.	SHA.	MAT	COM.	OTH.	AVG
OFA	47.2	72.6	66.7	100.0	85.1	80.7
VCM	56.6	69.3	73.5	99.7	88.1	83.5

Table 5: Significance of commonsense on VQA tasks.

	VQAv2	OK-VQA
OFA	75.3	33.8
w/ commonsense	75.8	34.6
Qwen-VL-7B	79.5	58.6
w/ commonsense	79.9	60.0

while the second involves visual question answering (VQA) datasets that require both image understanding and external knowledge. Intuitively, the discovered visual commonsense can enhance performance on both VL tasks. Due to space limitations, we provide more details in the Appendix.

Visual Commonsense Evaluation We assess whether VCD could enhance VCM’s visual commonsense capabilities by comparing the backbone and VCM on IMAGENETVC, where we would expect VCM, fine-tuned on VCD upon the backbone, to demonstrate superior performance. The experimental results are reported in Tab. 4. It is observed that VCM_{large} shows improvements in the categories of Color, Material, Component, and Others, indicating an overall improvement in commonsense knowledge. However, there is a minor decrease in recognizing Shapes, likely due to a deficiency in VCD’s shape-related commonsense.

Visual Question Answering This section evaluates on VQAv2 (Goyal et al., 2017) and OK-VQA (Marino et al., 2019) to underscore the significance of the commonsense discovered by VCM for VQA. We compare the results from the backbone against those from the backbone incorporating commonsense discovered by VCM as complement information for answering the question. The experimental results are reported in Tab. 5. We can find that integrating the commonsense discovered by VCM indeed enhances the performance of both VQA datasets, which is evidence of the significance of VCD for downstream VL tasks.

5 Conclusion

We introduced VCD, a large-scale Visual Commonsense Dataset that bridges the gap between linguistic and visual commonsense. By combining structured relations from ConceptNet with object-level annotations from Visual Genome, VCD provides both seen and unseen commonsense across three aspects: Property, Action, and Space. This hierarchical taxonomy enables fine-grained, scene-dependent, and object-specific commonsense representation. To demonstrate its utility, we developed VCM, a generative model trained with instruction tuning. The model exhibits a strong ability in discovering various types of visual commonsense, and improves performance on vision-language tasks like visual question answering. VCD provides a foundation for enhancing visually grounded commonsense understanding and reasoning, enabling AI systems to better capture visual commonsense knowledge for real-world applications.

Limitations

One limitation of our study is that, due to space constraints, this paper primarily focuses on introducing VCD and the corresponding task of visual commonsense discovery. Our evaluation of the significance of discovered visual commonsense for vision-language tasks is relatively simple. In future work, we plan to conduct a more systematic and comprehensive evaluation across a broader range of downstream vision-language tasks.

Furthermore, this study considers only visual commonsense in static images, leaving out dynamic, temporal, and causal visual commonsense as reflected in videos. Exploring these aspects presents a promising research direction which we aim to pursue in future work.

Acknowledgments

This work was supported by the Natural Science Foundation of China (No. 62476134).

References

Houda Alberts, Teresa Huang, Yash Deshpande, Yibo Liu, Kyunghyun Cho, Clara Vania, and Iacer Calixto. 2020. Visualsem: a high-quality knowledge graph for vision and language. *arXiv preprint arXiv:2008.09150*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of*

the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, pages 65–72.

- Yu-Wei Chao, Zhan Wang, Rada Mihalcea, and Jia Deng. 2015. *Mining semantic affordances of visual object categories*. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4259–4267.
- Xiang Chen, Ningyu Zhang, Lei Li, Shumin Deng, Chuanqi Tan, Changliang Xu, Fei Huang, Luo Si, and Huajun Chen. 2022. *Hybrid transformer with multi-level fusion for multimodal knowledge graph completion*. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 904–915. ACM.
- Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. 2013. *Neil: Extracting visual knowledge from web data*. In *2013 IEEE International Conference on Computer Vision*, pages 1409–1416.
- Guillem Collell, Luc Van Gool, and Marie-Francine Moens. 2018. *Acquiring common sense spatial knowledge through implicit spatial templates*. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. *Instructblip: Towards general-purpose vision-language models with instruction tuning*. *CoRR*, abs/2305.06500.
- Markos Diomataris, Nikolaos Gkanatsios, Vassilis Pitsikalis, and Petros Maragos. 2021. Grounding consistency: Distilling spatial common sense for precise visual relationship detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15911–15920.
- Tianqing Fang, Hongming Zhang, Weiqi Wang, Yangqiu Song, and Bin He. 2021. *DISCOS: bridging the gap between discourse knowledge and commonsense knowledge*. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 2648–2659. ACM / IW3C2.
- Sebastián Ferrada, Benjamin Bustos, and Aidan Hogan. 2017. *Imgpedia: a linked dataset with content-based analysis of wikimedia images*. In *The Semantic Web- ISWC 2017: 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part II 16*, pages 84–93. Springer.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. 2023. *Mme: A comprehensive evaluation benchmark for multimodal large language models*. *arXiv preprint arXiv:2306.13394*.
- Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen,

- Ryan Marten, Mitchell Wortsman, Dhruva Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah M. Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander J. Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. 2023. [Datacomp: In search of the next generation of multimodal datasets](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. [AllenNLP: A deep semantic natural language processing platform](#). In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. [spaCy: Industrial-strength Natural Language Processing in Python](#).
- Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. [\(comet-\) atomic 2020: On symbolic and neural commonsense knowledge graphs](#). In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 6384–6392. AAAI Press.
- Daniel Kahneman, Anne Treisman, and Brian J Gibbs. 1992. [The reviewing of object files: Object-specific integration of information](#). *Cognitive Psychology*, 24(2):175–219.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. [Visual genome: Connecting language and vision using crowdsourced dense image annotations](#). *International Journal of Computer Vision*, 123(1):32–73.
- Lei Li, Jingjing Xu, Qingxiu Dong, Ce Zheng, Xu Sun, Lingpeng Kong, and Qi Liu. 2023. [Can language models understand physical concepts?](#) In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11843–11861, Singapore. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Xiao Liu, Da Yin, Yansong Feng, and Dongyan Zhao. 2022. [Things not written in text: Exploring spatial commonsense from visual signals](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2365–2376, Dublin, Ireland. Association for Computational Linguistics.
- Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-Rubio, and David S Rosenblum. 2019. Mmkg: multi-modal knowledge graphs. In *The Semantic Web: 16th International Conference, ESWC 2019, Portorož, Slovenia, June 2–6, 2019, Proceedings 16*, pages 459–474. Springer.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. [OK-VQA: A visual question answering benchmark requiring external knowledge](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3195–3204. Computer Vision Foundation / IEEE.
- Daniel Onoro-Rubio, Mathias Niepert, Alberto García-Durán, Roberto González, and Roberto J López-Sastre. 2017. Answering visual-relational queries in web-extracted knowledge graphs. *arXiv preprint arXiv:1709.02314*.
- OpenAI. 2023. [GPT-4 technical report](#). *CoRR*, abs/2303.08774.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [BLEU: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

- Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. Visual-comet: Reasoning about the dynamic context of a still image. In *Computer Vision – ECCV 2020*, pages 508–524, Cham. Springer International Publishing.
- Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 01, pages 3027–3035.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. [LAION-5B: an open large-scale dataset for training next generation image-text models](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Xiangqing Shen, Siwei Wu, and Rui Xia. 2023. [Dense-atomic: Towards densely-connected ATOMIC with high knowledge coverage and massive multi-hop paths](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2023, Toronto, Canada, July 9-14, 2023, pages 13292–13305. Association for Computational Linguistics.
- Shikhar Singh, Ehsan Qasemi, and Muhao Chen. 2023. [VIPHY: Probing “visible” physical commonsense knowledge](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7113–7128, Singapore. Association for Computational Linguistics.
- Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. [Conceptnet 5.5: An open multilingual graph of general knowledge](#). In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.
- Keh-Yih Su, Ming-Wen Wu, and Jing-Shin Chang. 1992. [A new quantitative quality measure for machine translation systems](#). In *COLING 1992 Volume 2: The 15th International Conference on Computational Linguistics*.
- Qu Tang, Xiangyu Zhu, Zhen Lei, and Zhaoxiang Zhang. 2023. Intrinsic physical concepts discovery with object-centric predictive models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23252–23261.
- Ramakrishna Vedantam, Xiao Lin, Tanmay Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. [Learning common sense through visual abstraction](#). In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2542–2550.
- Meng Wang, Haofen Wang, Guilin Qi, and Qiushuo Zheng. 2020. Richpedia: a large-scale, comprehensive multi-modal knowledge graph. *Big Data Research*, 22:100159.
- Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. [OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework](#). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR.
- Heming Xia, Qingxiu Dong, Lei Li, Jingjing Xu, Tianyu Liu, Ziwei Qin, and Zhifang Sui. 2023. [Imagenetvc: Zero- and few-shot visual commonsense evaluation on 1000 imagenet categories](#). *Preprint*, arXiv:2305.15028.
- Frank F. Xu, Bill Yuchen Lin, and Kenny Zhu. 2018. [Automatic extraction of commonsense LocatedNear knowledge](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 96–101, Melbourne, Australia. Association for Computational Linguistics.
- Zhiyang Xu, Ying Shen, and Lifu Huang. 2023. [Multi-Instruct: Improving multi-modal zero-shot learning via instruction tuning](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11445–11465, Toronto, Canada. Association for Computational Linguistics.
- Mark Yatskar, Vicente Ordonez, and Ali Farhadi. 2016. [Stating the obvious: Extracting visual common sense knowledge](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 193–198, San Diego, California. Association for Computational Linguistics.
- Ruichi Yu, Ang Li, Vlad I. Morariu, and Larry S. Davis. 2017. Visual relationship detection with internal and external linguistic knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Rowan Zellers, Ari Holtzman, Matthew Peters, Roozbeh Mottaghi, Aniruddha Kembhavi, Ali Farhadi, and Yejin Choi. 2021. [PIGLeT: Language grounding through neuro-symbolic interaction in a 3D world](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2040–2050, Online. Association for Computational Linguistics.

Chenyu Zhang, Benjamin Van Durme, Zhuowan Li, and Elias Stengel-Eskin. 2022. [Visual commonsense in pretrained unimodal and multimodal models](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5321–5335, Seattle, United States. Association for Computational Linguistics.

Hongming Zhang, Daniel Khashabi, Yangqiu Song, and Dan Roth. 2020a. [Transomcs: From linguistic graphs to commonsense knowledge](#). In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 4004–4010. ijcai.org.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020b. [ASER: A large-scale eventuality knowledge graph](#). In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 201–211. ACM / IW3C2.

Kaiwen Zhou, Kwonjoon Lee, Teruhisa Misu, and Xin Wang. 2024. [ViCor: Bridging visual understanding and commonsense reasoning with large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10783–10795, Bangkok, Thailand. Association for Computational Linguistics.

A Details of VCD Construction

A.1 Definition and Examples for Visual Commonsense Taxonomy

First, we present the definition of categories of visual commonsense in Tab. 6. Then, based on Fig. 1, we furthermore provide examples for each category of visual commonsense as an illustration. To maintain conciseness, not all bounding boxes in the image depicted in Fig. 1 are annotated.

A.2 Full Set of Mapping Rules

This section presents the full set of mapping rules, detailed in Tab. 7. Tab 7’s first column displays the syntactic structures identified through syntactic parsing of noun phrases, which are the result of constituent syntactic analysis. Depending on these syntactic structures, a subsequent mapping to a variety of types of visual commonsense is performed, as indicated in Tab. 7’s third column, and is based on the parts of speech (POS) outlined in the second column. To facilitate comprehension, Tab. 7 also includes corresponding examples and explanations for each distinct mapping rule.

A.3 Examples in VCD

We provide examples from VCD as illustrated in Fig. 6. VCD includes detailed commonsense corresponding to each object within the image, delineated by bounding boxes. This commonsense is expressed in the form of triples. For the sake of conciseness, we do not provide the complete set of commonsense for every object annotated with bounding boxes in the image.

A.4 Word Cloud for VCD

In Fig. 8 and Fig. 9, we provide word clouds for seen and unseen commonsense, respectively. It can be observed that seen commonsense frequently contains words describing positions and colors, *e.g.*, “white” and “on”. While in the word cloud for unseen commonsense, the word cloud contains words like “time” and “place”, which are abstract concepts. The differences well align to the definition of seen and unseen commonsense.

B Details of VCM Training and Evaluation

B.1 Implementation Details of VCM

VCM is based on OFA (Wang et al., 2022), an encoder-decoder architecture. While early vision-language models primarily focused on image-text

Table 6: Definitions and examples of visual commonsense taxonomy.

	Category	Definition	Example
Seen	Property	HasProperty The object has a <i>currently seen</i> property, such as shape, color, or material.	(car, /Seen/Property/HasProperty, yellow)
	Space	LocatedNear The object co-occurs with another object in a <i>currently seen</i> manner, without a specific spatial relationship.	(car, /Seen/Space/LocatedNear, streetlight)
		Relatedness The object has a <i>currently seen</i> spatial relationship with another object.	(car, /Seen/Space/Relatedness, after a car)
	Action	CapableOf The object performs a <i>currently seen</i> active action.	(car, /Seen/Action/CapableOf, drive on road)
		ReceivesAction The object undergoes a <i>currently seen</i> passive action.	(skateboard, /Seen/Action/ReceivesAction, played by man)
Unseen	Property	HasProperty The object has a <i>currently unseen</i> property, such as shape, color, or material.	(iron, /Unseen/Property/HasProperty, hard)
	Space	CreatedBy The object has a <i>currently unseen</i> method of creation.	(car, /Unseen/Property/CreatedBy, factory)
		LocatedNear The object co-occurs with another object in a <i>currently unseen</i> manner, without a specific spatial relationship.	(man, /Unseen/Space/LocatedNear, sofa)
	Action	CapableOf The object performs a <i>currently unseen</i> active action.	(man, /Unseen/Action/CapableOf, grow up)
		UsedFor The object has a <i>currently unseen</i> function or purpose.	(car, /Unseen/Action/UsedFor, drive to work)
		ReceivesAction The object undergoes a <i>currently unseen</i> passive action.	(car, /Unseen/Action/ReceivesAction, hit)

Table 7: Full set of mapping rules.

POS	Category	Example	Explanation
PP	- /Seen/Space/Relatedness	man before the yellow car \rightarrow (man, /Seen/Space/Relatedness, before car)	“ Man ” is the root noun. Regarding the prepositional phrase “before the yellow car”, we simplify it to “before car” to obtain the basic triple form.
VP	VBN /Seen/Action/ReceivesAction	man hit by a yellow car \rightarrow (man, /Seen/Action/ReceivesAction, hit by a car)	“ Man ” is the root noun. Regarding the verbal phrase “hit by a yellow car”, since POS of “hit” is “VBN”, we map it to a passive action.
	VBG /Seen/Action/CapableOf	car driving on the road \rightarrow (car, /Seen/Action/CapableOf, driving on road)	“ Man ” is the root noun. Regarding the verbal phrase “driving on the road”, since POS of “driving” is “VBG”, we map it to a active action.
NP	ADJ /Seen/Property/HasProperty	a small car \rightarrow (car, /Seen/Property/HasProperty, small)	“ Car ” is the root noun. Regarding the noun phrase “a small car”, since POS of “small” is “ADJ”, we map it to “/Seen/Property/HasProperty”.
	VBG /Seen/Action/CapableOf	a running man \rightarrow (man, /Seen/Action/CapableOf, run)	“ Man ” is the root noun. Regarding the noun phrase “a running man”, since POS of “running” is “VBG”, we map it to “/Seen/Action/CapableOf”.

alignment, more recent VL models have enhanced object localization capabilities by extending to region-text alignment. However, few models effectively incorporate coordinate specifications within instructions. We require a model with strong localization abilities that can seamlessly integrate coordinate information into instructions. Consequently, OFA, with its capability to process bounding box coordinates and its moderate model size, serves as a suitable foundation for instruction tuning.

VCD is divided into 8:1:1 for training, validation, and test set. The AdamW optimizer (Loshchilov and Hutter, 2019) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ is utilized for optimization. To avoid overfitting, we apply regularization techniques *i.e.*, dropout with a rate of 0.1, weight decay of 0.01, and label smoothing set at 0.1. We implement a linear decay learning rate scheduler with an initial warmup ratio of 0.06, and VCM is trained on 4 NVIDIA RTX A6000 GPUs for a total of 3 epochs.

B.2 Output with Multiple Triples

We model VCM in a generative manner, allowing the generation of novel commonsense triples that are not contained in the training set. Crucially, VCD provides a rich set of commonsense triples for each object-type pair. we devise a strategy to preserve this diversity in the generated output.

For seen commonsense, given n triples of an

object o and a type r , $\{(o, r, ec_1), \dots, (o, r, ec_n)\}$, we sample m triples and join their tail nodes with $[sep]$, yielding $ec_1[sep] \dots [sep] ec_m$, where $m \leq n$ helps manage the number of commonsense triples desired during generation.

Given the sorted unseen commonsense as described in Sec. 3.3.2, we concatenate the top- k tail nodes into $ic_1[sep] \dots ic_k$ and then add j random nodes sampled from the remaining to form $ic_1[sep] \dots ic_k[sep] ics_{k+1}, \dots, ics_{k+j}$. This strategy ensures the generation of high-priority unseen commonsense while also maintaining a diversity of lower-priority commonsense.

B.3 Evaluation Details of VCM

We select powerful open and closed source MLLMs on MME leaderboard (Fu et al., 2023), SPHINX, Qwen-VL-7B, and GPT-4o (OpenAI, 2023), and then manually compare them with VCM_{large} by carefully crafting the prompts.

The crafted prompts and qualitative results are illustrated in Tab. 9 and Tab. 10. Texts in double quotations are key components that can be adapted based on the object under consideration, the type of visual commonsense, *etc.* Here, we show only the optimally chosen prompt that was used for the experimental results in Tab. 3.

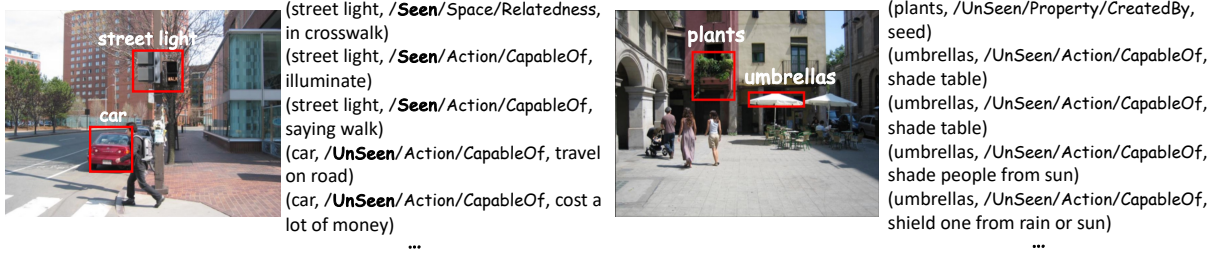


Figure 6: Examples in VCD.

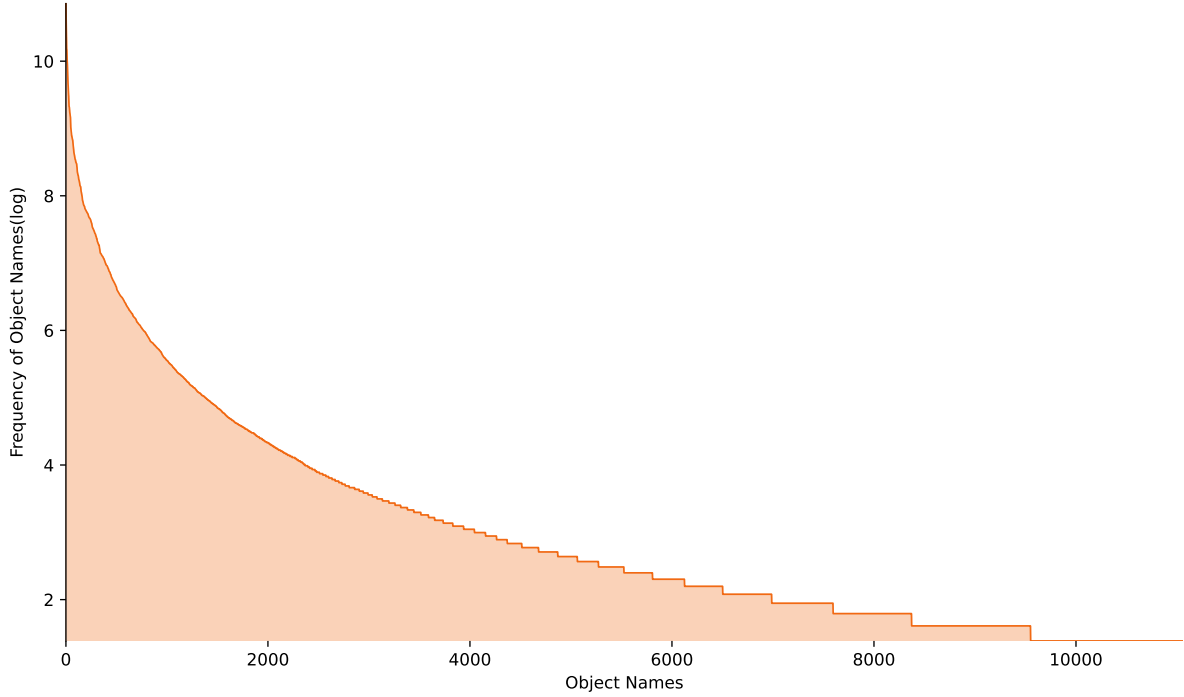


Figure 7: Distribution of object names in VCD.



Figure 8: Word cloud for seen commonsense in VCD.



Figure 9: Word cloud for unseen commonsense in VCD.

B.4 Qualitative Results of VCM

In this section, we showcase a variety of examples produced by VCM_{large} . As observed in Fig. 10, it is evident that VCM_{large} is capable of generating high-quality representations of both seen and unseen commonsense categories associated with a specific object in an image. While the

results are generally accurate, there may be occasional errors. For instance, in the final image of Fig. 10, “/Unseen/Property/HasProperty” of “sidewalk” is correctly discovered as “paved with concrete”; however, this description could be more precisely attributed to the type of “/Seen/Action/ReceivesAction”.

Table 8: Detailed hyperparameters of VCM configuration. We list the configuration for VCM of 4 different sizes.

Model	#Param.	Backbone	Hidden size	Intermediate Size	#Head	#Enc. Layers	#Dec. Layers
VCM _{tiny}	33M	ResNet50	256	1024	4	4	4
VCM _{medium}	93M	ResNet101	512	2048	8	4	4
VCM _{base}	182M	ResNet101	768	3072	12	6	6
VCM _{large}	472M	ResNet152	1024	4096	16	12	12

Table 9: Detailed prompts for evaluating GPT-4o.

Type	Prompt
/E/P/HasProperty	[image] List some “visible” “has property” commonsense about the “field” located in the red bounding box in the image. Output in the following format: (field, visible has property, short phrase).
/E/S/LocatedNear	[image] List some “visible” “located near” commonsense about the “cheese” located in the red bounding box in the image. Output in the following format: (cheese, visible located near, short phrase).
/E/S/Relatedness	[image] List some “visible” “related spatial relation” commonsense about the “hamburger” located in the red bounding box in the image. Output in the following format: (hamburger, related spatial relation, on the desk).
/E/A/CapableOf	[image] List some “visible” “capable of” commonsense about the “wings” located in the red bounding box in the image. Output in the following format: (wings, visible capable of, short phrase).
/E/A/ReceivesAction	[image] List some “visible” “receives passive action” commonsense about the “window” located in the red bounding box in the image. Output in the following format: (window, receives passive action, short phrase).
/I/P/HasProperty	[image] List some “invisible” “has property” commonsense about the “hose” located in the red bounding box in the image. Output in the following format: (hose, invisible has property, short phrase).
/I/P/CreatedBy	[image] List some “invisible” “created by” commonsense about the “tree” located in the red bounding box in the image. Output in the following format: (hose, invisible created by, short phrase).
/I/S/LocatedNear	[image] List some “invisible” “located near” commonsense about the “shadow” located in the red bounding box in the image. Output in the following format: (shadow, invisible located near, short phrase).
/I/A/CapableOf	[image] List some “invisible” “capable of” commonsense about the “tire” located in the red bounding box in the image. Output in the following format: (tire, invisible capable of, short phrase).
/I/A/UsedFor	[image] List some “invisible” “used for” commonsense about the “car” located in the red bounding box in the image. Output in the following format: (car, invisible used for, short phrase).
/I/A/ReceivesAction	[image] List some “invisible” “receives passive action” commonsense about the “picture” located in the red bounding box in the image. Output in the following format: (picture, invisible receives passive action, short phrase).

C Details of Downstream VL Tasks

C.1 Details of Evaluation on IMAGENETVC

IMAGENETVC evaluates commonsense understanding across multiple dimensions, including color, shape, material, components, and other attributes of various objects. For example, the question “What is the color of a koala?” from IMAGENETVC assesses the model’s knowledge of a koala’s typical color, which is brown. To evaluate performance on IMAGENETVC, we compare the backbone model with the backbone trained on VCD by directly testing them on IMAGENETVC, using experimental setting provided by (Xia et al., 2023).

C.2 Details of Evaluation on VQA

We evaluate on VQA tasks to underscore the significance of the commonsense discovered by VCD. While visual commonsense reasoning could be a good option, the datasets (Zellers et al., 2019; Park et al., 2020) mostly focus on human behaviors and states, not offering a broad reflection of commonsense’s role for diverse objects as discovered by

VCD. Therefore, we select VQAv2 (Goyal et al., 2017) OK-VQA (Marino et al., 2019) dataset, a visual question answering task requiring models to utilize visual information from images to answer questions.

It is reasonable to assume that performing VCD on an image can provide additional insights that are instrumental in improving VQA performance, as commonsense serves as additional information for better question answering.

For each question in VQAv2 and OK-VQA, we begin by identifying the entities contained within the questions using dependency parsing. Next, we employ OFA_{large} to determine the bounding boxes corresponding to these entities in the image. Finally, we used VCM_{large} for visual commonsense discovery pertaining to the identified objects in the image.

It is important to note that not all the commonsense discovered by VCM_{large} is equally beneficial for answering the questions. As such, we filter the commonsense and retain only what is most relevant to the question. The reserved commonsense is

Table 10: Detailed prompts for evaluating SPHINX and Qwen-VL-7B.

Type	Prompt
/E/P/HasProperty	[image] Can you tell me what visible properties common sense the “field” has in this image?
/E/S/LocatedNear	[image] Can you tell me what exists near this “cheese” in the bounding box of the image?
/E/S/Relatedness	[image] Can you tell me anything about the location of the “hamburger” in the picture?
/E/A/CapableOf	[image] Please list some “capable of” commonsense you can see about “wings” in the bounding box of the image.
/E/A/ReceivesAction	[image] Can you tell me what are the common sense passively accepted actions of the “window” in the picture?
/I/P/HasProperty	[image] Can you tell me what visible properties commonsense the “car” has in this image?
/I/P/CreatedBy	[image] Please list some “created by” commonsense you can imagine about “tree” in the bounding box in the image.
/I/S/LocatedNear	[image] Please list some “located near” commonsense you can imagine about “shadow” in the bounding box of the image.
/I/A/CapableOf	[image] Please list some “capable of” commonsense you can imagine about “tire” in the bounding box of the image.
/I/A/UsedFor	[image] Please use your imagination to tell me what can “car” in the picture be used for.
/I/A/ReceivesAction	[image] Please list some “receives passive action” commonsense you can imagine about “picture” in the bounding box of the image.

			
<p>(lamp, Unseen/Action/ReceivesAction): placed on table; found in room; found to sit on dresser; found in office building; ...</p> <p>(book, Unseen/Property/CreatedBy): writer; author</p> <p>(blanket, Seen/Space/Relatedness): on bed</p>	<p>(painting, Unseen/Space/LocatedNear): school; rest on easel; museum; gallery opening; attic; ...</p> <p>(church, Seen/Property/HasProperty): large; white</p> <p>(cross, Seen/Space/LocatedNear): table; light; carpet; cross; building; pew; ...</p>	<p>(glass, Unseen/Action/CapableOf): withstand modest forces; shattering; hold water; shattering if left wet; ...</p> <p>(glass, Seen/Action/CapableOf): withstand modest forces; shattering; hold water; shattering if left wet; hold liquid; ...</p> <p>(bread, Seen/Action/ReceivesAction): toasted</p>	<p>(sidewalk, Unseen/Property/HasProperty): sticky; smooth; paved with concrete; flat long and narrow</p> <p>(sign, Unseen/Action/UsedFor): visual understanding; unspoken words; traffic directions; telling to do things; ...</p>

Figure 10: Qualitative results generated by VCM_{large}.

then concatenated with the question, enriching the context for the answer generation.

For fair comparison, both the backbone and the backbone with visual commonsense are finetuned on VQAv2 and OK-VQA from scratch under the identical experimental setting.

To validate the effectiveness of different types of commonsense, we randomly select 100 samples from VQA-v2 and evaluate the accuracy manually with different types of commonsense. Tab. 11 shows the efficacy of seen over unseen commonsense for VQAv2, since VQAv2 is a relatively simple VQA dataset where most of the questions are related to seen commonsense.

Table 11: Accuracy with different type of commonsense on VQAv2.

Category	✗	Explicit	Implicit
Accuracy	0.89	0.92	0.90