A Details of Data Augmentation with External Knowledge Resources

✔ *Enhance Relation Recognition*: We enriched the relationships between objects parsed from the original knowledge descriptions by leveraging the external resource of ConceptNet. ConceptNet comprises commonly observed entities and their connections, where edge weights signify the reliability and frequency of these relationships. The typical value of edge weights in ConceptNet is 1. To prevent the redundancy of common information and to maintain the validity of the enriched relations, we categorized the relationships based on their weights. Relationships with weights less than 1 were deemed "weak" and those with a weight of 1 were labeled "average". We refrained from using these categories for relation enhancement. Instead, only relationships with weights greater than 1, indicative of high reliability, were employed for augmenting the relations.

✔ *Boost Entity Perception*: On the entity side, we augment complement entities and descriptive information with two external knowledge resources. On one hand, for descriptions with a high TF-IDF+ score, we enrich related entities of the object from ConceptNet to create additional knowledge descriptions. The relatedness is based on the between-word relatedness score provided by ConceptNet and we take the threshold as 0.85. On the other hand, we employ the Commonsense Transformers (COMET) [4] model to enrich related new objects and descriptive information. The COMET model is a language model designed to generate commonsense knowledge and understand causal relationships between descriptions. It is pretrained using the atomic dataset, which consists of structured, crowd-sourced knowledge about everyday events and their associated causes and effects. The COMET model can provide neighbor descriptions of the given input of nine different categories of relation. We take the `xAttr` and `oEffect` relation categories and augmented the COMET model by formulating the existing knowledge description texts as the input and choose the corresponding category branch during generation for enriching objects and descriptions respectively.

B Dataset Information

Table 6: Dataset statistics.

split	#image	#descriptor	#relation	#subject & object
Train	75,456	832,351	30,241	302,735
Validation	4,871	64,137	5,164	34,177
Test	4,873	62,579	5,031	32,384

The statistic information of our augmented dataset is summarized in Table 6, where **split** specifies the dataset split, **#image** indicates the number of images in the split, **#descriptor** indicates the total number of relational descriptors of the images, **#relation** is the total number of unique relations in the relational descriptors after deduplication, and **#subject & object** is the total number of subjects and objects contained in the description text.

C Implementation Details

Hyperparameter	Assignment
batch size	4
learning rate optimizer	Adam
Adam epsilon	1e-8
Adam initial learning rate	1e-5
learning rate scheduler	cosine scheduler
Adam decay weight	0.05

Table 7: Hyperparameters for training open relational region detector.

Hyperparameter	Assignment
batch size	4
learning rate optimizer	Adam
Adam epsilon	1e-8
Adam initial learning rate	1e-5
learning rate scheduler	cosine scheduler
Adam decay weight	0.05
α	0.7
ϕ	0.01

Table 8: Hyperparameters for training format-free visual knowledge generator.

Open relational region detector. The visual feature extraction backbone is constructed upon a pre-trained ResNet50-FPN. The detector head incorporates a $\text{BLIP}_{\text{base}}$ equipped with the essential

ViT-B/16 for text supervision, using multiple fully connected layers to derive region features. For each candidate region, we engage a regressor to conduct boundary regression on these features. The detector undergoes fine-tuning for 20 epochs using the relational region bounding box dataset and an Adam optimizer [26]. The hyperparameters for training are detailed in Table 7.

Format-free visual knowledge generator. The format-free visual knowledge generator is initialized from $\text{BLIP}_{\text{base}}$, which incorporates the basic ViT-B/16. We fine-tune the generator model for 20 epochs using the same optimizer as the one employed for the region detector. Detailed hyperparameters for the visual knowledge generator can be found in Table 8.

D Human Evaluation Guidance and Interface

We perform the human evaluation on two of the four in-depth knowledge quality assessment metrics. We build an interface by referring to [48], where raters are presented with a given image and the corresponding knowledge descriptions and are required to choose one from the multiple choice for two questions on whether the knowledge is valid to humans and whether the knowledge description depicts the image. The detailed scoring criteria for *Validity* and *Conformity* are provided below:

- *Validity* (↑): *whether the generated visual knowledge is valid to humans*.
  - 0 (Invalid): The knowledge description does not conform to human cognition, rendering it unreliable or misleading to humans.
  - 1 (Valid): The knowledge description is valid and accurately conforms to human cognition, providing reliable and meaningful knowledge to humans.
- *Conformity* (↑): *whether the generated knowledge faithfully depicts the scenarios in the images*.
  - 0 (Inconsistent): The knowledge description does not faithfully depict the scenarios in the images, showing significant deviations or discrepancies, making it difficult for users to relate the textual information to the visual context.
  - 1 (Partially Conforming): The knowledge description partially conforms to the scenarios in the images, but there might be minor inconsistencies or missing relevant details.
  - 2 (Moderately Conforming): The knowledge description exhibits a moderate level of conformity with the scenarios in the images, capturing the key aspects and providing coherent descriptions.
  - 3 (Highly Conforming): The knowledge description highly conforms to the scenarios in the images, accurately capturing the details and faithfully representing the visual context.

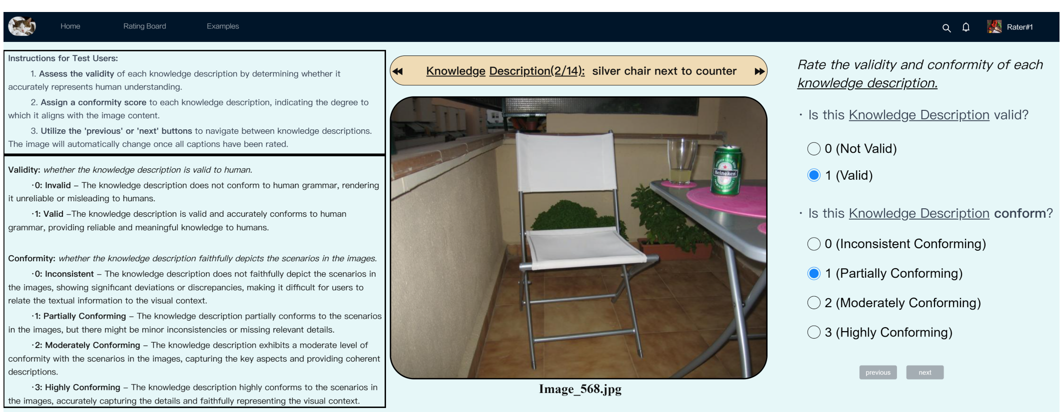


Figure 8: The human evaluation interface for in-depth knowledge quality evaluation.

Agreement/validation We use Cohen's κ as the agreement score to measure potential subjectivity involved in ratings of knowledge quality. Cohen's κ is a statistic that is used to measure inter-rater reliability for qualitative items and is scaled from -1 (perfect systematic disagreement) to 1 (perfect agreement), where values ≤ 0 as indicating *no agreement* and 0.01-0.20 as *none to slight*, 0.21-0.40 as *fair*, 0.41–0.60 as *moderate*, 0.61-0.80 as *substantial*, and 0.81-1.00 as *almost perfect* agreement. Our calculated average pairwise Cohen's κ on human evaluation results from three different raters is 0.76, which indicates a good agreement.

E Parametric Knowledge Prompting Template

Given an image $\mathcal{I}$ and the corresponding extracted visual knowledge from it based on OpenVik, we perform knowledge comparison with parametric knowledge contained in LLM by prompting the gpt-3.5-turbo model with the object information contained in the $\mathcal{I}$. The prompt format is shown in the followings:

```
Suppose you are looking at an image that contains the following subject
and object entities:
Subject list:  [Insert the subject names here]
Object list:  [Insert the object names here]
Please extract 5-10 condensed descriptions that describe the interactions
and/or relations among those entities in the image.  Try to elucidate the
associations and relationships with diverse language formats instead of
being restricted to sub-verb-obj tuples.
```

F More Case Studies of Open Visual Knowledge from OpenVik

Figure 9 shows some other cases on the extracted open visual knowledge from OpenVik. In comparison to VG and Relational Caps, OpenVik exhibits superior performance at capturing novel entities, expanding object interactions through diverse relations, and enriching knowledge representation with nuanced descriptive details. For example for the bottom right image, OpenVik can extract novel entities such as "*tracks*", "*shoe*", diverse relations such as "*sticking out of*", and nuanced descriptive details such as "*cold thick*", "*with man feet on it*", "*brave*". The generated knowledge with a more format-free semantic structure is highlighted in red.

Figure 9: Case studies of open visual knowledge from OpenVik.

G More Qualitative Examples on Applications

G.1 Text-to-Image Retrieval

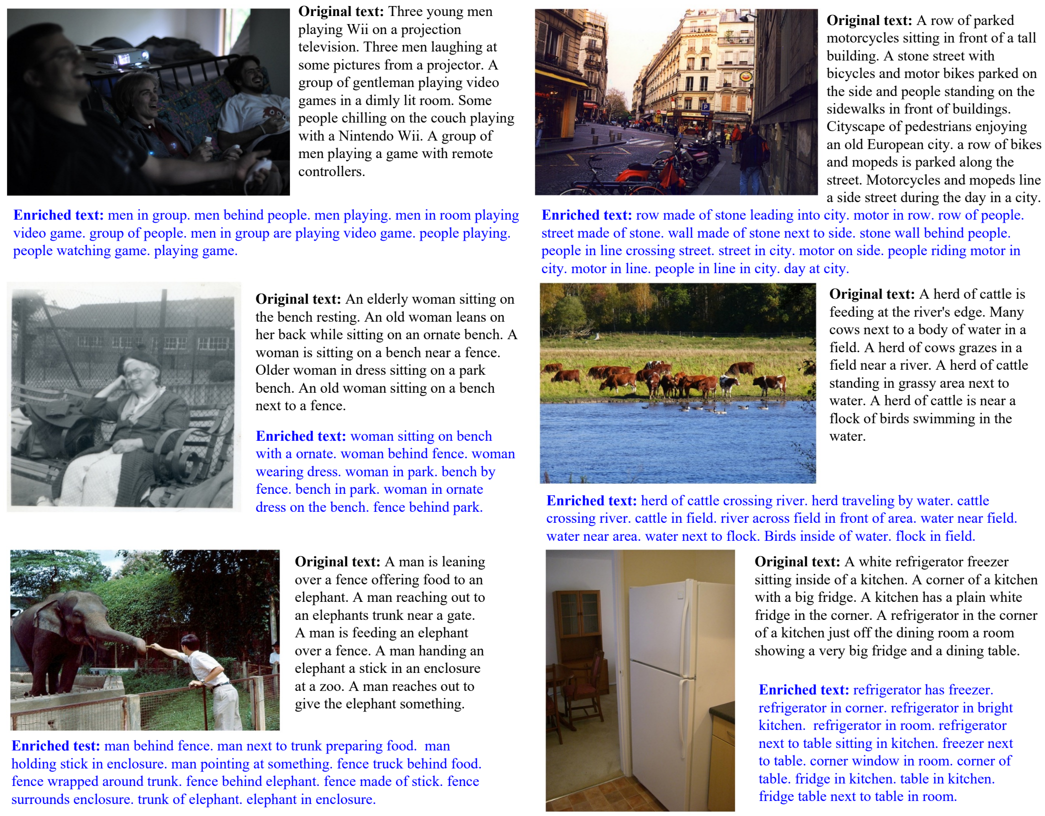


Figure 10: Qualitative examples of `OpenVik` context enrichment on text-to-image retrieval.

Figure 10 presents more qualitative examples of `OpenVik`-based visual knowledge enrichment on captions. The enriched text is based on the objects present in the images themselves, supplemented with additional relationships from our generated visual knowledge in `OpenVik`. It is shown that the introduced relationships often provide new context information that aligns with the visual content of the images. For example, in the image of an old woman sitting on a bench in a park, the enriched context information includes the positional relationship between the "*bench*", "*fence*", and "*park*", which provides a more comprehensive description of the original image.

G.2 Grounded Situation Recognition

Figure 11 presents more qualitative examples of `OpenVik`-based context enrichment in the grounded situation recognition (GSR) task. Our context enrichment setting for the GSR task is to perform enrichment based on verbs like "*shopping*" and "*carrying*". We further restrict the enriched context with the objects contained in the image to avoid noisy enrichment. For example, for the image showing people shopping at a market, the enriched knowledge contexts could be "*the people shopping at market*", "*standing person shopping for fruit*". The idea is to enrich the original description $\mathcal{T}$: "*An image of <verb>*" with relevant actions and relations with the extracted visual knowledge from `OpenVik`, which can potentially help in drawing-in the matched candidates.

G.3 Visual Commonsense Reasoning

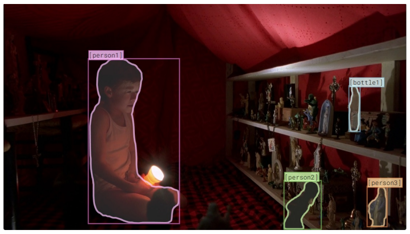
Figure 12 presents more qualitative examples of `OpenVik`-based context enrichment in the visual commonsense reasoning (VCR) task. The context enrichment on VCR is performed at two-level,

Figure 11: Qualitative examples of `OpenVik` context enrichment on task GSR.

incorporating both entities and relations: (1) we parse the question and options to obtain all `(S, O)` pairs and, for each entity pair, apply the same relation augmentation as in the image retrieval task; (2) for the `V` in each option, we enrich the visual context using the same method as illustrated in GSR. It is shown that unrelated answers are usually enriched with contexts that are not relevant to the image, thus enlarging the distance between incorrect answers and the question, e.g., the enriched contexts "*squating person fixing handy bathroom*" for example 3 in Figure 12. At the same time, the knowledge description of the correct answer is enhanced by incorporating information that aligns with the image contents, e.g., the enriched knowledge contexts "*sitting people on red ground*" for example 1 in Figure 12.

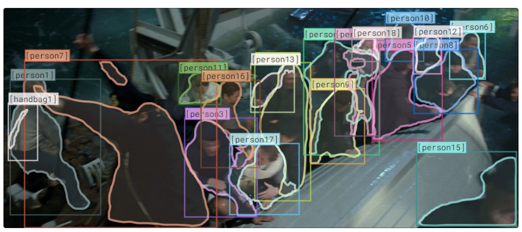
H Full List of Filtered Verbs for GSR

We provide the full list of verbs out of the predefined 504 candidates of GSR [34] that can be accurate-matched or fuzzy-matched to extracted visual knowledge in Table 9, based on which we compose the testing subset for our evaluation on GSR application in Section 5.2.



Question: Where is' Person1 sitting?
A He is in a laboratory.
B He is sitting at a bar. the person sitting behind sneaky barrier.
C In a fort in his house. the person walking by light house.
D He is sitting on the ground. sitting person on red ground.
Answer: D He is sitting on the ground.

Question: Where is Person2 going?
A Person2 is going into the store. the person walking into store.
B Person2 is getting into a carriage. sitting person inside carriage.
C Person1 is going to the bathroom. squating person fixing handy bathroom.
D Person1 is going outside to play after the conversation with Person2 is over.
Answer: A Person2 is going into the store.



Question: Why is Person7 in motion?
A Person14 is running desperately.
B Person7 is climbing over the boat. the person standing inside white boat.
C Person7 is walking fast to the bathroom. squating person fixing handy bathroom.
D Person7 is going to try to protect Person10 from a threat. Person7 is moving forward to challenge what ever could be there.
Answer: B Person7 is climbing over the boat.

Question : What will Person2 do next?
A Person2 will speak angrily at diningtable2, then walk off.
B Person2 will sit down on chair1. painting person near giant chair.
C Person2 will feed bowl1. the person skate boarding in a athletic bowl.
D Person2 will open the box. the person holding a box full of oranges.
Answer: B Person2 will sit down on chair1.

Question: Where are Person1 and Person2?
A Person1 and Person2 are sitting outside of a general store. the person walking by store.
B Person1 and Person2 are standing on top of a train car. jumping person on top board. walking person next to white train. the person walking near active car. yellow train sitting atop track. sliced carrot on top counter red car of old train.
C Person1 and Person2 are in an office. walking person outside office.
D Person1 and Person2 are in the kitchen. the person eating in hungry kitchen.
Answer: C Person1 and Person2 are in an office.

Question: What is Person1 doing here?
A He is in prison serving a prison sentence. person writing sentences.
B He is trying to get information. person gaining information.
C Person1 is a waiter. person talking with waiter in restaurant.
D He is existing a building. walking person near large building.
Answer: C Person1 is a waiter.

Figure 12: Qualitative examples of OpenVik context enrichment on task VCR.

Table 9: The full list of filtered verbs for GSR.

Matching Type	The Word List of Event Types
Accurate	putting, butting, bathing, dusting, rearing, turning, skating, placing, carting, staring, biting, mashing, folding, wetting, sprinkling, branching, drying, standing, flaming, taxiing, performing, circling, molding, parachuting, glowing, fishing, drinking, speaking, pawing, blocking, milking, racing, stripping, potting, spinning, eating, making, kicking, catching, lacing, urinating, sleeping, pressing, buttering, shearing, sliding, hiking, glaring, dipping, swimming, shopping, slicing, shelling, wagging, grilling, crafting, raining, clawing, splashing, rubbing, snowing, breaking, guarding, clipping, sewing, braiding, telephoning, buttoning, waiting, serving, picking, camping, leaning, working, kissing, wrapping, trimming, tripping, pasting, soaring, driving, kneeling, pumping, coloring, lighting, training, ducking, bowing, arching, cooking, checking, pushing, flipping, rocking, cresting, cleaning, reading, nailing, stitching, building, climbing, covering, shelving, attaching, calming, selling, gluing, dyeing, lapping, photographing, peeling, sprouting, licking, displaying, combing, stacking, planting, fastening, buying, mopping, burning, erasing, measuring, dining, tattooing, gardening, decorating, clearing, fixing, weeding, pulling, feeding, watering, crowning, shaking, dripping, emptying, typing, chasing, poking, leaping, pouring, hanging, sniffing, piloting, falling, overflowing, resting, crashing, carving, ballooning, wading, loading, shaving, boarding, pinning, rowing, juggling, shoveling, hugging, throwing, calling, singing, carrying, walking, writing, crouching, floating, painting, opening, tying, riding, strapping, dialing, saying, bubbling, signing, camouflaging, operating, leading, laughing, parading, skiing, drawing, gnawing, celebrating, spreading, filling, giving, running, smelling, plowing, helping, brushing, scooping, adjusting, wrinkling, steering, biking, smiling, spraying, boating, paying, chewing, stuffing, clinging, landing, wheeling, talking, scoring, teaching, jogging, pitching, flapping, tipping, scrubbing, sitting, surfing, stirring, competing, drumming, jumping, filming, dancing, waxing, hitting, recording, baking, waving, washing, signaling, chopping, stretching, rafting, microwaving, phoning, lifting, swinging, releasing, ramming, towing, packing, hauling, frying (*244 words*)
Fuzzy	educating, marching, spanking, descending, smearing, heaving, cramming, inflating, stooping, inserting, squeezing, tugging, tilting, moistening, swarming, subduing, waddling, winking, flexing, punching, attacking, nuzzling, sprinting, sucking, puckering, sketching, rotting, videotaping, complaining, tuning, locking, hurling, pricking, arranging, constructing, slapping, sweeping, restraining, dousing, frisking, twisting, wringing, hoisting, immersing, shredding, blossoming, igniting, spying, offering, pouting, confronting, docking, assembling, prying, grinning, sharpening, pruning, disciplining, nipping, coaching, nagging, storming, handcuffing, apprehending, bouncing, clenching, taping, distributing, striking, studying, plunging, curling, aiming, sowing, grinding, rinsing, punting, mowing, hitchhiking, skipping, leaking, providing, hunching, spoiling, kneading, burying, foraging, lathering, vaulting, ejecting, mending, pinching, deflecting, ascending, peeing, bothering, repairing, pedaling, ailing, fueling, skidding, scraping, soaking, grimacing, scolding, spitting, knocking, crushing, bandaging, saluting, fording, stumbling, discussing, raking, launching, whirling, fetching, brawling, retrieving, snuggling, exercising, colliding, stroking, whipping, tilling, betting, farming, browsing, examining, dropping, barbecuing, ignoring, asking, flinging, perspiring, embracing, slipping, flicking, smashing, arresting, lecturing, tearing, gasping, applying, counting, spilling, dragging, recovering, practicing, scratching, shooting, packaging, hunting, stinging (*154 words*)