Improved Visual Grounding through Self-Consistent Explanations

Supplementary Material

Appendix

We present details about our two-level LLM prompts to obtain high-quality paraphrases for region-centric phrases, provide examples of such extracted paraphrases, conduct additional evaluations on in-the-wild captions on CC3M, justify our selection of ALBEF as our base model, evaluate in more detail the choice of object-centric captions, and provide additional qualitative results.

A. LLM-Prompting Details

This section presents prompting details and generated examples of our proposed two-level self-consistency data augmentation method (refer to Section 3.3 in the main text). As shown in Figure 1, we classify textual annotations from existing datasets into global-based captions that describe the entire image, and region-based captions that describe a specific region within the image. In our experiments, we identify captions from Visual Genome (VG) [2] as region-based captions, while captions from MS-COCO [6] and Conceptual Captions 3M (CC3M) [8] are referred to as global-based captions.

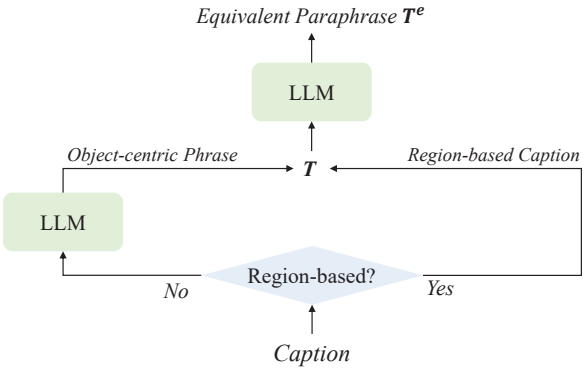


Figure 1. An overview of our two-level self-consistency data augmentation approach. For global-based captions, we use an LLM to chunk object-centric phrases T with our first-level prompting. The obtained phrases or region-based captions T are further input to an LLM in the second-level prompting to obtain equivalent paraphrases T^e.

A.1. Prompt Design

We adopt an in-context learning [1] strategy in our LLM-prompt design. Each pair of our prompts encompasses a query text **Q** and an expected answer **A**. The query texts **Q** were selected and modified based on generation quality and successful rate on a small validation subset.

```
Q: "a group of sheep with in grassy area next to trees"
A: ["a group of sheep", "grassy area", "trees"]
Q: "a train is on going down the track while people watch"
A: ["a train", "the track", "people"]
Q: "street sign in front of a clear glass building"
A: ["street sign", "a clear glass building"]
Q: "a woman sitting on a bus with a green purse looking at her cell phone"
A: ["a woman", "a bus", "a green purse", "her cell phone"]
Q: "a photo of a bus that is boarding passengers"
A: ["a bus", "boarding passengers"]
Q: "this is an image of a bathroom which is empty"
A: ["a bathroom"]
Q: "an individual covers himself under an umbrella on a rainy day"
A: ["an individual", "an umbrella"]
Q: "a city street scene with cars and a person crossing the street"
A: ["cars", "a person", "the street"]
Q: [T]
A:
```

Figure 2. In-context few-shot LLM-Prompt for our first-level self-consistency data augmentation. We leverage an LLM for phrase chunking to obtain object-centric captions from captions that describe images globally. $[T]$ is the query text. **A** in the last row is followed by the output.

Phrase Chunking. To obtain object-centric phrases, we prompt an LLM for global-based captions as shown in Figure 2. Compared to previous phrase chunking methods [9], this LLM-prompt-based approach aligns more closely with our objective. In conventional phrase chunking, the output chunks include nouns, verbs, and prepositional phrases. However, our SelfEQ method focuses on object-centric phrases rather than verbs or prepositions. Additionally, there are some abstract nouns such as "photo", "image" and "scene" that are commonly used in annotations but do not benefit visual grounding. We select several captions containing abstract nouns, verbs, and prepositions as our query texts in our prompts to guide the LLM to generate the chunks in a task-specific way.

Paraphrase Generation. Our second-level prompts aim to generate paraphrased captions through substitute words and for this purpose we use an in-context few-shot prompt with an LLM. Figure 3 shows the specific prompt we use for input region-based captions. As in-context examples, we select four captions with different sentence structures or where the primary object is playing different syntactic roles: A complex noun phrase (set 1), an existential structure (set 3), a subject followed by a passive verb phrase (set 4), and a noun phrase modified by a prepositional phrase (set 6). We also add two short noun phrases (sets 2, 5) to represent relevant captions. For the expected output **A**, we first detect the primary object (*i.e.*, the "group" field) in the query text, then use WordNet synsets [7] to generate relationships (*i.e.*, the "synonym", "antonym", "hypernym", and "meronym" fields) for the primary object. We further remove inaccurate

```
Q: "a smiling, holding something and standing man"
A: {"group": "man",
    "synonym": ["adult male","gentleman"],
    "antonym": ["woman"],
    "hypernym": ["adult","grownup","male","male person","human being","human"],
    "meronym": ["man body","man head","man arms","man legs"]}
Q: "a desk"
A: {"group": "desk",
    "synonym": ["table","writing desk"],
    "antonym": [],
    "hypernym": ["writer tool","furniture"],
    "meronym": ["desktop","desk leg"]}
Q: "there is dark grey carpet on the floor"
A: {"group": "carpet",
    "synonym": ["rug","carpeting"],
    "antonym": ["wall"],
    "hypernym": ["floor cover","floor covering","furnishing"],
    "meronym": ["carpet edging"]}
Q: "street is covered with snow"
A: {"group": "street",
    "synonym": [],
    "antonym": ["sky"],
    "hypernym": ["thoroughfare","road","path","way"],
    "meronym": []}
Q: "white books"
A: {"group": "books",
    "synonym": [],
    "antonym": ["computer", "pen"],
    "hypernym": ["publication","paper"],
    "meronym": ["book cover"]}
Q: "tree with sparse leaves"
A: {"group": "tree",
    "synonym": [],
    "antonym": [],
    "hypernym": ["ligneous plant","woody plant","plant","vegetation"],
    "meronym": ["tree bole","tree burl","tree crown","tree limb","tree stump",
                "tree branch","tree stump","tree trunk","treetop","tree trunk"]}
Q: [T]
A:
```

Figure 3. In-context few-shot LLM-Prompt for our second-level self-consistency data augmentation on Visual Genome (VG). We leverage an LLM to generate paraphrases for the given textual descriptions (region-based captions). $[T]$ is the regional-based caption, and **A** in the last row is followed by the expected output.

words as the final expected output **A** in our LLM-prompt.

Figure 4 shows our in-context few-shot LLM-prompt for object-centric phrases obtained from our first-level data augmentation. It is based on our LLM-prompt for region-based captions (Figure 3), but we further chunk the query texts into object-centric phrases for adaption and modify them (*e.g.*, articles, pronouns) empirically on a small validation subset.

A.2. Generated Data Examples

This section presents generated examples from our two-level self-consistency augmentation approach. We apply paraphrase generation for VG and both phrase chunking and paraphrase generation for MS-COCO and CC3M.

Visual Genome. Figure 5 shows generated example paraphrases for VG. Since our LLM-prompt contains various sentence structures in the provided in-context examples, the generated data showcases the successful detection of primary objects ("group") for a variety of input captions. To ensure the quality of equivalent paraphrases, we allow the LLM to leave blanks in relevant fields if no appropriate words are available. In total, the LLM generates equivalent paraphrases for 74.56% of captions for VG. The different types of equivalent paraphrases include general synonyms (*e.g.*, "bicycle", "sofa"), formal or technical terms

```
Q: "two men"
A: {"group": "men",
    "synonym": ["adult males","gentlemen"],
    "antonym": ["women"],
    "hypernym": ["adults","grownups","males","male people","human beings","humans"],
    "meronym": ["men bodies","men heads","men arms","men legs"]}
Q: "a desk"
A: {"group": "desk",
    "synonym": ["table","writing desk"],
    "antonym": [],
    "hypernym": ["writer tool","furniture"],
    "meronym": ["desktop","desk leg"]}
Q: "this dark grey carpet"
A: {"group": "carpet",
    "synonym": ["rug","carpeting"],
    "antonym": ["wall"],
    "hypernym": ["floor cover","floor covering","furnishing"],
    "meronym": ["carpet edging"]}
Q: "street"
A: {"group": "street",
    "synonym": [],
    "antonym": ["sky"],
    "hypernym": ["thoroughfare","road","path","way"],
    "meronym": []}
Q: "the white books"
A: {"group": "books",
    "synonym": [],
    "antonym": ["computers", "pens"],
    "hypernym": ["publications","paper"],
    "meronym": ["book cover"]}
Q: "tree with sparse leaves"
A: {"group": "tree",
    "synonym": [],
    "antonym": [],
    "hypernym": ["ligneous plant","woody plant","plant","vegetation"],
    "meronym": ["tree bole","tree burl","tree crown","tree limb","tree stump",
                "tree branch","tree stump","tree trunk","treetop","tree trunk"]}
Q: [T]
A:
```

Figure 4. In-context few-shot LLM-Prompt for our second-level self-consistency data augmentation on MS-COCO. We leverage an LLM to generate paraphrases for the given textual descriptions (object-centric phrases). $[T]$ is the object-centric phrase obtained from the first-level phrase chunking. **A** in the last row is followed by the expected output.

(*e.g.*, "pedal cycle"), colloquial or regional variants (*e.g.*, "pushbike", "settee") and descriptive synonyms (*e.g.*, "scattered trees").

MS-COCO. Generated examples for MS-COCO are shown in Figure 6. Captions in MS-COCO describe the entire images, so we augment them with our two-level phrase augmentation: Phrase chunking and equivalent paraphrase generation. In the first level, we prompt an LLM to generate object-centric phrases **A1**. In this way, we separate a broad descriptive region into several specific object-centric regions in a scenario, aligning more closely with the objective of visual grounding. Additionally, it may fix small issues with grammar and typos (*e.g.*, "siting at there desks"), potentially providing higher quality textual descriptions. Similar to VG, the second level prompting leads to equivalent paraphrases **A2**, including diverse types for 84.47% of object-centric phrases.

CC3M. Figure 7 showcases generated data from two-level self-consistency data augmentation for CC3M. Compared to manually annotated data from VG and MS-COCO (See Figures 5 and 6), web-crawled Alt-Text-image pairs in CC3M are noisier and potentially unaligned. We apply our first-level augmentation, phrase chunking **A1**, not only for object-centric phrases, but also as a filtering strategy to extract meaningful words from likely ambiguous sentences.

Q: "a stadium"
A: { "group": "stadium",
"synonym": ["arena", "sports stadium"],
"antonym": [],
"hypernym": ["sports facility", "outdoor sports venue", "place of sporting events"],
"meronym": ["stadium seating", "stadium lights", "stadium sound system"]}

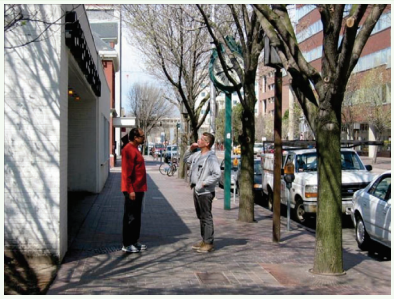
Q: "hills are behind the people"
A: { "group": "hill",
"synonym": ["mountain", "rising ground"],
"antonym": ["valley", "low land"],
"hypernym": ["landform", "terrain", "elevation"],
"meronym": ["hill crest", "hill top", "hill summit", "hillside", "hill foot"]}

Q: "a brown and wooden podium"
A: { "group": "podium",
"synonym": ["platform", "lectern", "speakers platform"],
"antonym": ["desk"],
"hypernym": ["stage", "platform", "speaking platform"],
"meronym": ["podium step", "podium base"]}

Q: "a grey couch"
A: { "group": "couch",
"synonym": ["sofa", "settee"],
"antonym": [],
"hypernym": ["furniture", "seating"],
"meronym": ["couch cushion", "couch cover", "couch pillow"]}

Q: "two bikes side by side"
A: { "group": "bike",
"synonym" : ["bicycle", "pushbike", "pedal cycle"],
"antonym" : ["car"],
"hypernym" : ["vehicle", "transport", "two wheeler"],
"meronym" : ["bike frame", "bike pedal", "bike tire", "bike handlebar"]}

Q: "sparse trees"
A: { "group": "trees",
"synonym": ["scattered trees"],
"antonym": ["dense forest"],
"hypernym": ["group of trees"],
"meronym": ["tree", "tree trunk"]}



Q: "palm trees near building"
A: { "group": "palm tree",
"synonym": [],
"antonym": [],
"hypernym": ["tree", "plant"],
"meronym": ["palm frond"]}

Q: "an orange fruit"
A: { "group": "orange fruit",
"synonym": ["citrus fruit", "orange"],
"antonym": [],
"hypernym": ["citrus fruit", "fruit"],
"meronym": ["orange peel", "orange juice", "orange segments"]}

Figure 5. LLM generated examples for VG. **Q** represents the query text associated with the image. **A** corresponds to the output of our second-level self-consistency data augmentation. "group" denotes the detected primary object, and it further generates relationships such as "synonym," "antonym," "meronym," and "hypernym" regarding with the identified "group."

Q: "People all siting at there desks with drinks ."
A1: "object-centric phrases": ["people", "desks", "drinks"]
A2: {"group": "people",
"synonym": ["human beings", "humans"],
"antonym": ["machine"],
"hypernym": ["living organism", "mankind", "humanity", "person", "creatures"],
"meronym": ["people body", "people head", "people arms", "people legs"]}
{"group": "desks",
"synonym": ["tables", "writing desks"],
"antonym": [],
"hypernym": ["writer tool", "furniture"],
"meronym": ["desk top", "desk leg"]}
{"group": "drinks",
"synonym": ["beverages", "refreshments"],
"antonym": ["food"],
"hypernym": ["nonalcoholic beverages", "alcoholic beverages", "beverage", "drink"],
"meronym": ["drink container", "drink cup"]}

Q: "Many colorful cupcakes are stacked below a similar cake."
A1: "object-centric phrases": ["many colorful cupcakes", "a similar cake"]
A2: {"group": "cupcakes",
"synonym": [],
"antonym": ["cake"],
"hypernym": ["sweet", "baked goods", "dessert", "pastry", "food"],
"meronym": ["cupcake wrapper", "cupcake liner", "cupcake filling", "cupcake frosting", "cupcake decoration"]}
{"group": "cake",
"synonym": ["dessert", "sweet"],
"antonym": ["salad"],
"hypernym": ["food", "edible", "baked goods"],
"meronym": ["cake ingredients", "cake decorations", "cake icing", "cake recipe", "cake mix", "cake baking", "cake layers", "cake pan", "cake recipes", "cake ingredients"]}

Q: "A white and gray cat laying underneath an umbrella."
A1: "object-centric phrases": ["a white and gray cat", "an umbrella"]
A2: {"group": "cat",
"synonym": ["feline", "cat animal", "cat creature"],
"antonym": [],
"hypernym": ["mammal", "animal", "pet"],
"meronym": ["cat fur", "cat eyes", "cat nose", "cat paws", "cat tail", "cat body", "cat head", "cat ears"]}
{"group": "umbrella",
"synonym": ["parasol", "brolly"],
"antonym": ["sun"],
"hypernym": ["covering", "shelter", "protection", "canopy"],
"meronym": ["umbrella handle", "umbrella frame", "umbrella fabric", "umbrella spike"]}

Q: "A group of people that are standing near a tennis net."
A1: "object-centric phrases": ["a group of people", "a tennis net"]
A2: {"group": "group of people",
"synonym": ["crowd", "assembly", "multitude"],
"antonym": ["solitude", "loneliness"],
"hypernym": ["collective noun", "noun", "group", "mob", "crowd", "assembly"],
"meronym": ["people in the group"]}
{"group": "tennis net",
"synonym": ["tennis court net", "nets for tennis"],
"antonym": [],
"hypernym": ["tennis court", "tennis netting"],
"meronym": ["tennis court net post", "tennis net mesh"]}

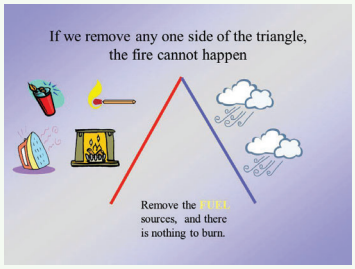
Q: "A white clock mounted to a white wall next to a curtain."
A1: "object-centric phrases": ["a white clock", "a white wall", "a curtain"]
A2: {"group": "clock",
"synonym": ["timepiece", "clockwork"],
"antonym": ["stopwatch"],
"hypernym": ["timekeeper", "time measuring instrument"],
"meronym": ["clock face", "clock hands", "clock numbers"]}
{"group": "wall",
"synonym": [],
"antonym": ["light"],
"hypernym": ["partition", "divider", "architecture"],
"meronym": ["wall hanging"]}
{"group": "curtain",
"synonym": ["drape", "curtains"],
"antonym": [],
"hypernym": ["window treatment", "window covering", "home furnishing", "textile"],
"meronym": ["curtain rod", "curtain tieback"]}

Q: "In this scene we see a person flying a kite with a flag attached."
A1: "object-centric phrases": ["a person", "flying a kite", "a flag"]
A2: {"group": "person",
"synonym": ["human being", "individual", "someone"],
"antonym": ["non person", "thing"],
"hypernym": ["human", "man", "woman", "personhood", "life form"],
"meronym": ["person body", "person head"]}
{"group": "kite",
"synonym": ["flying a kite", "kite flying"],
"antonym": ["not flying a kite"],
"hypernym": ["sport", "recreation", "hobby"],
"meronym": ["kite string", "kite tail", "kite stick"]}
{"group": "flag",
"synonym": ["banner", "standard"],
"antonym": ["country"],
"hypernym": ["national symbol", "emblem", "sign"],
"meronym": ["flag pole", "flag halyard"]}

Figure 6. LLM generated examples for MS-COCO. **Q** represents the query text associated with the image. **A1** is the object-centric phrase obtained from the first-level self-consistency data augmentation, while **A2** corresponds to the second level. For each object-centric phrase in **A1**, LLM detects primary objects "group" and generates relevant relationships in **A2**.

Q: "golden vintage greeting card on a black background"
A1: "object-centric phrases": ["golden vintage greeting card"]
A2: {"group": "greeting card",
"synonym": ["greeting", "postcard", "holiday card"],
"antonym": [],
"hypernym": ["communication", "message", "note"],
"meronym": ["card base", "card front", "card back"]}

Q: "if we remove any side of the triangle the fire can not happen"
A1: "object-centric phrases": ["triangle"]
A2: {"group": "triangle",
"synonym": ["3 sided shape", "triangle shape"],
"antonym": ["circle"],
"hypernym": ["polygon", "2 dimensional shape", "geometric shape"],
"meronym": ["triangle sides", "triangle angles"]}



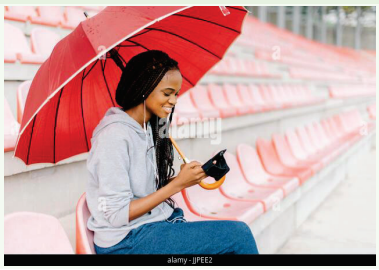
Q: "front view of passenger airplane flying in the sky vector art illustration"
A1: "object-centric phrases": ["passenger airplane"]
A2: {"group": "passenger airplane",
"synonym": ["airliner", "commercial airplane", "jet"],
"antonym": ["cargo airplane"],
"hypernym": ["aircraft", "aeroplane", "airship", "airboat", "airplane", "jetplane", "airliner", "jetliner"],
"meronym": ["passenger cabin", "cockpit", "wings", "tail", "undercarriage", "landing gear"]}

Q: "dry corn or maize plants in a farm field during harvesting with rows of uncut plants alongside stubble in a receding"
A1: "object-centric phrases": ["farm field"]
A2: {"group": "field",
"synonym": ["paddock", "meadow", "garden", "crop"],
"antonym": ["city"],
"hypernym": ["agricultural land", "growing area", "cultivated land", "arable land", "tillable land"],
"meronym": ["crop rows", "crop circles", "crop fields", "crop land"]}

Q: "an illustration of a gold genie lamp stock vector"
A1: "object-centric phrases": ["gold genie lamp"]
A2: {"group": "genie lamp",
"synonym": ["magic lamp", "wishing lamp"],
"antonym": [],
"hypernym": ["light source", "illuminant", "luminous object", "shining object"],
"meronym": ["genies lamp", "lamp base", "lamp shade"]}

Q: "person is pictured on his bicycle on a street that appears similar"
A1: "object-centric phrases": ["his bicycle"]
A2: {"group": "bicycle",
"synonym": ["bike", "pedal cycle"],
"antonym": ["car"],
"hypernym": ["means of transport", "vehicle", "two wheeled vehicle"],
"meronym": ["bicycle pedal", "bicycle tire", "bicycle chain", "bicycle saddle"]}

Q: "close up side portrait of the joyful afro american girl texting via the mobile phone while sitting under umbrella"
A1: "object-centric phrases": ["joyful afro american girl"]
A2: {"group": "girl",
"synonym": ["young lady", "miss"],
"antonym": ["boy"],
"hypernym": ["female", "woman", "young woman"],
"meronym": ["teen girl", "preteen girl"]}



Q: "sad homeless puppy falls asleep on a piece of burlap"
A1: "object-centric phrases": ["sad homeless puppy"]
A2: {"group": "puppy",
"synonym": ["pup", "puppy", "canine"],
"antonym": ["happy", "wealthy"],
"hypernym": ["animal", "mammal", "dog"],
"meronym": ["puppy paws", "puppy eyes", "puppy nose"]}

Figure 7. LLM generated examples for CC3M. **Q** represents the query text associated with the image. **A1** is the object-centric phrase obtained from the first-level self-consistency data augmentation, while **A2** corresponds to the second level. For each object-centric phrase in **A1**, LLM detects primary objects "group" and generates relevant relationships in **A2**.

Objective	Data Selection	RefCOCO+		Flickr30k	ReferIt
		Test A	Test B		
$\mathcal{L}_{\text{vl}}$	-	69.35	53.77	79.38	59.72
$\mathcal{L}_{\text{vl}}$	*ranked*	68.51	52.23	78.61	59.48
$\mathcal{L}_{\text{vl}}$	*random*	69.40	52.83	79.76	60.78
$\mathcal{L}_{\text{SelfEQ}}$	*ranked*	70.21	53.75	80.91	61.00
$\mathcal{L}_{\text{SelfEQ}}$	*random*	**71.59**	**54.19**	**81.53**	**63.04**

Table 1. Visual Grounding results when training with a subset of Conceptual Captions 3M (CC3M). The *ranked* subset corresponds to a set of image-text pairs filtered using the image-text matching score yielded by the base ALBEF. The *random* subset corresponds to a randomly selected subset. Applying SelfEQ on a *random* but relatively more noisy subset yields the best results.

The second-level augmentation further generates equivalent paraphrases **A2** for our SelfEQ tuning.

B. Effectiveness on Noisy Web-Crawled Data

We run additional experiments using two different subsets of data from the CC3M dataset, each containing $\sim 200k$ image-text pairs. The first subset, which we refer to as *ranked*, corresponds to a set of high-quality image-text pairs filtered using the image-text matching score yield by the base ALBEF model. The second subset, which we refer to as *random*, corresponds to a randomly selected set of arbitrary image-text pairs. We further generate paraphrases for each text caption using our two-level LLM-based augmentation and train our base model using the ALBEF baseline losses and our SelfEQ approach. Table 1 shows the effectiveness of our method with noisy web-crawled data.

C. Base Model Selection

We choose ALBEF [3] as our base model based on the off-the-shelf visual grounding ability through GradCAM under the pointing game setting as reported in the original work. We further compare it with the off-the-shelf performance of BLIP [4] and BLIP-2 [5] for reference. As shown in Table 2, ALBEF outperforms other methods by a large margin on visual grounding.

D. Object-Centric vs. Global-Based Captions

Table 3 shows the effect of different ways of chunking global-based captions. In our main paper, we demonstrate that shorter captions that are more object-specific lead to better results. Here we provide an additional chunking strategy that leads to captions that have a length between our short object-centric phrases P and the original long captions C, showing the benefits of gradually using shorter captions that are more likely to be object-centric. Specifically, we compare MS-COCO captions C and object-centric phrases

Method	RefCOCO+		Flickr30k	ReferIt
	Test A	Test B		
BLIP [4]	61.23	41.07	60.56	45.81
BLIP-2 [5]	50.09	42.26	64.86	45.34
ALBEF [3]	**69.35**	**53.77**	**79.38**	**59.72**

Table 2. Pointing game accuracy comparisons with other pre-trained vision-language models on off-the-shelf visual grounding via GradCAM.

```
Q: "a group of sheep with in grassy area next to trees"
A: ["a group of sheep with in grassy area"]
Q: "a train is on going down the track while people watch"
A: ["a train is on going down the track"]
Q: "a laptop sitting on a wooden table with a cord plugged in"
A: ["a laptop sitting on a wooden table"]
Q: "a woman sitting on a bus with a green purse looking at her cell phone"
A: ["a woman sitting on a bus"]
Q: "a photo of a bus that is boarding passengers"
A: ["a photo of a bus"]
Q: "this is an image of a bathroom which is empty"
A: ["this is an image of a bathroom"]
Q: "an individual covers himself under an umbrella on a rainy day"
A: ["an individual covers himself under an umbrella"]
Q: "a city street scene with cars and a person crossing the street"
A: ["a person crossing the street"]
Q: [T]
A:
```

Figure 8. LLM-Prompt for an alternative first-level self-consistency data augmentation (*i.e.*, phrase chunking) strategy. In contrast to object-centric phrases, the expected answer **A** further includes simple compositions.

P with long phrases P' whose length is between global captions and object-centric phrases. As shown in Figure 8, long phrases P' shorten the captions C by removing compound or descriptive sentences, while they still remain simple compositions compared to object-centric phrases P. Rows 4 and 5 in Table 3 supplement experiments in Table 4 in the main paper, demonstrating an increasing trend when more object-centric (*i.e.*, shorter). Notably, SelfEQ improves all formats of input text (C, P', P) compared to the base model and the vision-and-language objective ($\mathcal{L}_{\text{vl}}$).

D.1. Qualitative Results

Visual Grounding. Figure 9 presents more qualitative results for visual grounding than those shown in the main paper. SelfEQ excels in localizing input textual descriptions across a variety of challenging scenarios, including objects with prepositions (rows 1 to 4), intricate background context (row 5), numerical answers (row 6), distinguishing a single descriptive object from multiple similar ones (rows 7 and 8), dealing with occluded objects (row 9), and handling tiny objects (row 10). SelfEQ improves visual grounding through self-consistency tuning without any bounding boxes, while still achieving competitive performance com-

Figure 9. Qualitative comparisons on visual grounding. The reference text is on the top of each row. From left to right, it presents the image, our base model ALBEF, SotA box-supervised method AMC, and our method SelfEQ.

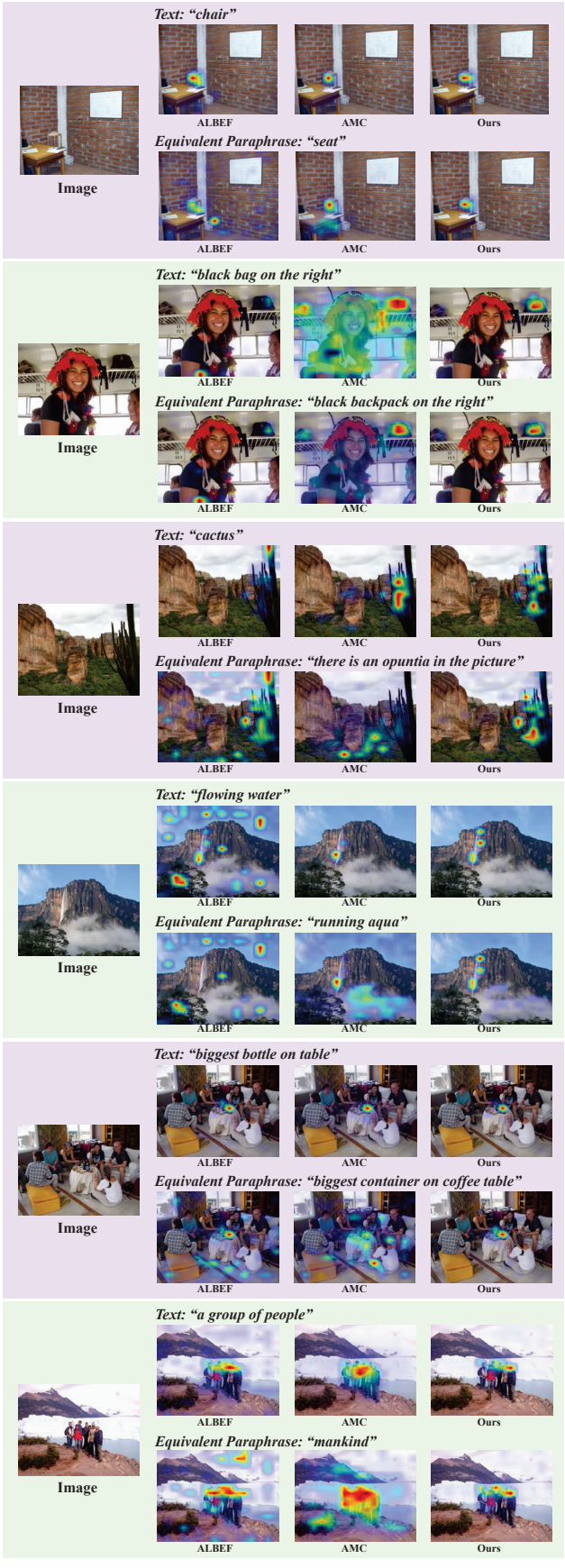


Figure 10. Qualitative comparisons on self-consistency. For each image, the first row is the reference text, and the second row is the equivalent paraphrase. Each column presents our base model ALBEF, SotA box-supervised method AMC, and our method SelfEQ.

Format	Objective	Flickr30k	ReferIt
-	$\mathcal{L}_{\text{vl}}$	79.38	59.72
C	$\mathcal{L}_{\text{vl}}$	79.90	60.64
C	$\mathcal{L}_{\text{SelfEQ}}$	81.28	62.04
P'	$\mathcal{L}_{\text{vl}}$	80.42	60.83
P'	$\mathcal{L}_{\text{SelfEQ}}$	82.09	62.12
P	$\mathcal{L}_{\text{vl}}$	81.18	61.18
P	$\mathcal{L}_{\text{SelfEQ}}$	**84.07**	**62.75**

Table 3. Trade-off between object-centric and rich context. The first row is the off-the-shelf base model performance. C is the caption, P is the object-centric phrase. P' is the long phrase, which can be defined as a shortened caption or an object-centric phrase with simple compositions.

pared to the state-of-the-art box-supervised method AMC.

Self-Consistency. Figure 10 shows more qualitative results, showcasing that our method generates more consistent results for paraphrases. SelfEQ leads to consistent results for various challenging scenarios such as handling general synonyms (row 1), synonym substitution (row 2), terminology and sentence extension (row 3), attributive and head noun substitution (row 4), multiple synonym substitution (row 5), and phrase-to-word transformation (row 6). Although our self-consistency data augmentation concentrates on synonym substitution, SelfEQ shows robust self-consistency in dealing with some non-trivial equivalent paraphrases.

E. Limitations and Future Work

We demonstrate that generating paraphrases based on noun substitutions leads to relatively reliable paraphrases. However, paraphrases generated in this way can be limited in terms of their diversity and complexity. Although our work shows encouraging results even for more complex forms of paraphrases at test time, investigating more reliable ways of generating visual paraphrases could lead to further gains. In addition, consistency can be imposed based on relations other than equivalence but also inclusion and exclusion relations. Generating automatic phrases that describe objects or regions with superordinate referring expressions or referring expressions that exclude content are possible paths for future work.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

[2] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123: 32–73, 2017. 1

[3] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34, 2021. 6

[4] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 6

[5] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 6

[6] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[7] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 1

[8] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 1

[9] Feifei Zhai, Saloni Potdar, Bing Xiang, and Bowen Zhou. Neural models for sequence chunking. In *Proceedings of the AAAI conference on artificial intelligence*, 2017. 1