

BACON: Supercharge Your VLM with Bag-of-Concept Graph to Mitigate Hallucinations

Zhantao Yang^{1,2*}, Ruili Feng^{2*}, Keyu Yan², Huangji Wang¹, Zhicai Wang^{2,3}
Shangwen Zhu¹, Han Zhang^{1,2}, Jie Xiao^{2,3}, Pingyu Wu^{2,3}, Kai Zhu², Jixuan Chen²
Chen-Wei Xie², Chaojie Mao², Yue Yang⁴, Hongyang Zhang⁵, Yu Liu², Fan Cheng^{1†}

¹Shanghai Jiao Tong University, ²Alibaba group

³University of Science and Technology of China

⁴University of Pennsylvania, ⁵University of Waterloo

Abstract

This paper presents **Bag-of-Concept Graph (BACON)** to gift models with limited linguistic abilities to taste the privilege of Vision Language Models (VLMs) and reduce hallucinations in the downstream tasks such as detection, visual question answering (VQA), and image generation. Since the visual scenes in physical worlds are structured with complex relations between objects, BACON breaks down annotations into basic minimum elements and presents them in a graph structure. Element-wise style enables easy understanding, and structural composition liberates difficult locating. Careful prompt design births the BACON captions with the help of public-available VLMs and segmentation methods. In this way, we gather a dataset with 100K annotated images, which endow VLMs with remarkable capabilities, such as accurately generating BACON, transforming prompts into BACON format, envisioning scenarios in the style of BACON, and dynamically modifying elements within BACON through interactive dialogue and more. Wide representative experiments, including detection, VQA, and image generation tasks, tell BACON as a lifeline to achieve previous out-of-reach tasks or excel in their current cutting-edge solutions.

1 Introduction

A single image can tell long stories, weaving tales as those told across myriad pages, akin to the masterful strokes seen in “The Last Supper” (Pitre, 2015) or the vibrant dynamism of “Liberty Leading the People” (Chaulagain, 2018). Yet emerging multi-modality tasks feed on describing images with language, which is bound to be lengthy and convoluted. Though Vision Language Models (VLMs) (OpenAI, 2023) can now easily produce those detailed annotations, downstream

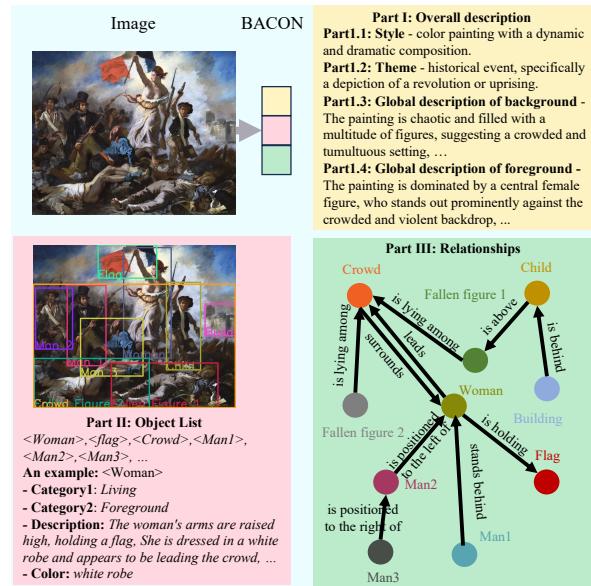


Figure 1: BACON representation of an image, including overall description, object list, and relationships.

multi-modality applications may not easily process them—understanding long and intricate context is still a privilege of VLMs and is far beyond the ability of most popular multi-modality methods, like those ResNet-based models (He et al., 2016a; Huang et al., 2017; He et al., 2016b), segmentation methods (Kirillov et al., 2023; Liu et al., 2023b; Ke et al., 2024), and Contrastive Language-Image Pre-training methods (CLIPs) (Radford et al., 2021; Lee et al., 2022; Li et al., 2021). Thus, popular multi-modality methods usually suffer from *model hallucinations*, where generative contents from these methods are inconsistent with input prompts. For example, off-the-shelf image generative models such as DALL-E 3 and SDXL cannot even count apples correctly and do not understand position relationships between objects, generating inconsistent images with human-provided captions (see Figure 5).

Challenges stem from two main issues: the complexity of long phrases and the difficulty in pinpointing specific information. Smaller mod-

* Corresponding author, † Equal contribution

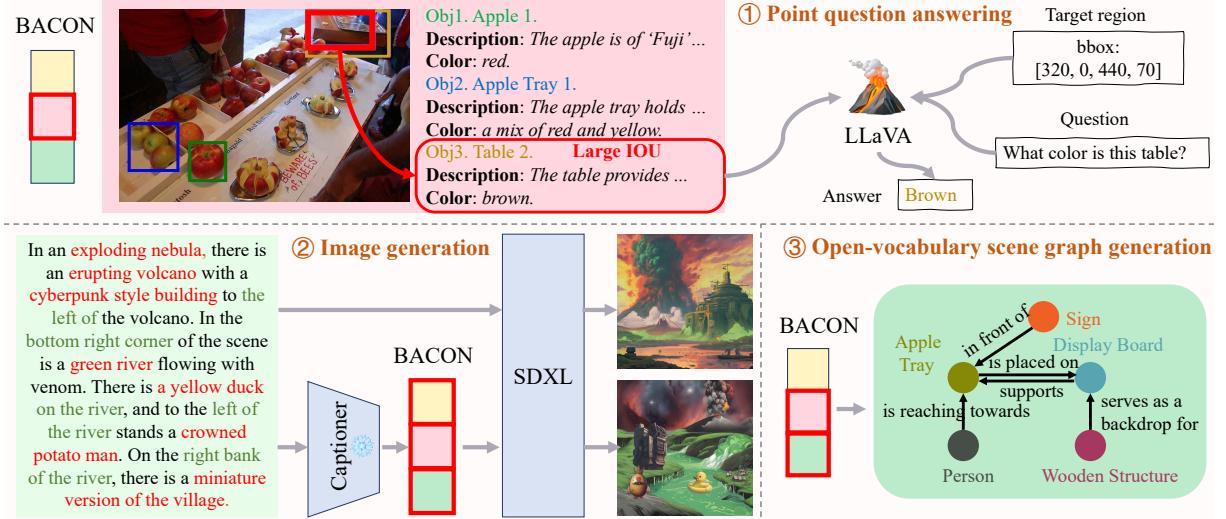


Figure 2: **Schematic diagram of multiple exemplary downstream tasks can benefit from BACON** Specifically, BACON can (1) enable VLMs to carry out the point question answering task previously beyond their scope; (2) assist text-to-image generative models such as SDXL in creating intricate images with higher precision as demanded by prompts; and (3) execute open-vocabulary scene graph generation tasks that were not feasible for other VLMs.

els like CLIPs and Segment Anything Models (SAMs) (Kirillov et al., 2023), constrained by their training data and resources, grasp only word-level or phrase-level language and struggle to train extensive attention mechanisms for detailed context analysis. Thus, a critical question emerges:

Shall the meticulous breakdown of VLM annotations into clearly distinguishable and comprehensible elements significantly mitigate model hallucinations and elevate the performance of downstream tasks?

Luckily, the answer is yes. This paper proposes BACON—a neat and efficient approach for VLMs to mitigate hallucinations and boost downstream tasks. BACON breaks down annotations into basic elements (like objects, style of picture, relationships, etc.) within an image, which can be easily understood by various methods. After that, In-Context Learning (ICL) (Brown et al., 2020) techniques are deployed to organize the annotation into a graph, with nodes representing fundamental visual elements and edges illustrating their connections. Illustrated in Figure 1, BACON composes of three key parts: 1) an **overall description** capturing the image’s overall style, theme, and key features; 2) a detailed **object list** with labels and descriptions for every item in the image; 3) the **relationships** between these objects. The structure of BACON allows models to flexibly extract and use the desired parts to comprehend complex information and accomplish tasks previously beyond their reach (see detailed examples as Figure 2).

In conclusion, the contribution of this work can be summarized as follows. 1) We introduce BACON, a rich and digestible caption method for boosting multi-modality tasks; 2) We collect a high-quality BACON-Dataset using human labor and advanced VLM models; the dataset and BACON-Captioner trained on it will be soon open-source; 3) The proposed method boosts wide and representative downstream tasks remarkably, demonstrating significant potential in promoting future research.

2 Bag-of-Concept Graph

In this section, we first explain the design of Bag-of-Concept Graph in Section 2.1, and then introduce how to get a BACON representation from an image in Section 2.2.

2.1 Design of BACON

BACON aims to tackle the issues faced by smaller downstream models in understanding long phrases and in locating specific pieces of information. For the former, BACON deconstructs the image annotations into basic elements, ensuring that even smaller downstream models can fully comprehend them. Subsequently, BACON employs a specific graph structure to amalgamate these elements, ensuring each element appears in a designated spot, allowing smaller downstream models to query and retrieve them easily. Specifically, the real world can often be explicitly represented using a scene graph composed of objects and their relationships (Miller, 1995; Doddington et al., 2004; Krishna et al., 2017;

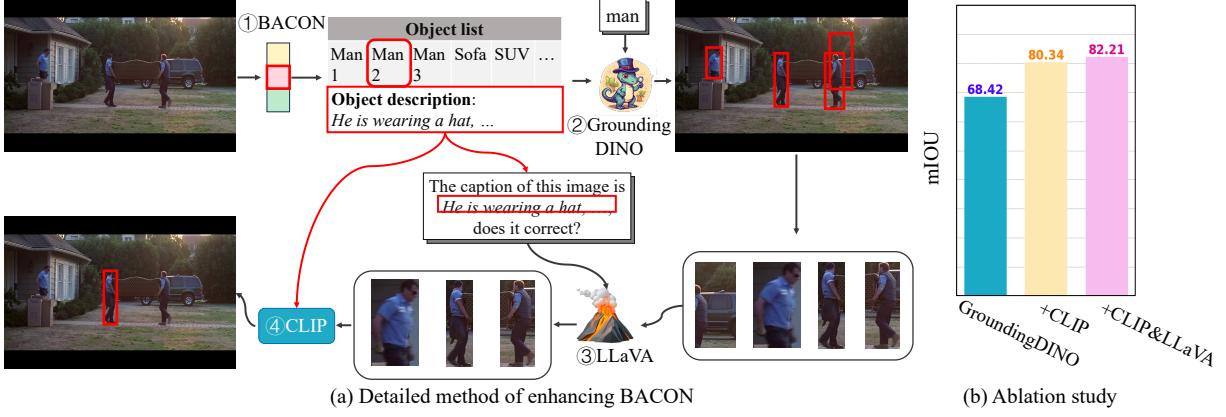


Figure 3: (a) **Detailed method** for graph grounding. The method contains four steps: 1) Extracting BACON from images using GPT-4V or BACON-Captioner; 2) Getting candidate regions using Grounding DINO given the name of the object; 3) Using LLaVA to discard blatant incorrect regions; 4) Select the region whose image feature matches the text feature of object description the most by CLIP. (b) **Ablation study** of method in (a), exploring the improvement of introducing CLIP and LLaVA, where the experiment is conducted on BACON benchmark.

Lu et al., 2016; Xu et al., 2017; Johnson et al., 2015, 2018). Inspired by this insight, we adopt the structure of a graph to deconstruct and reassemble basic elements. When selecting these basic elements, we have deliberately considered those that may prove beneficial for downstream tasks (listed in Appendix A.2.1). Finally, we have obtained BACON as shown in Figure 1.

Surprisingly, we discover that **such a design also significantly increases the number of correct objects included in captions** (see the higher recall score of BACON compared to GPT-4V in Section 4.2.2). We believe this is due to the structured object list emphasizing the distinction of separated entities, urging VLMs to separate complex objects into more detailed components.

2.2 Formulation of BACON

Given an image I , we aim to induce a structural representation $G = (D, O, R, B)$, where D is the textual description, O is the list of objects in the image, with R denotes their relationships and B as their bounding box positions. In practice, we optimize the construction of G in two stages: (1) **Graph Construction** (Section 2.3), which utilizes VLMs to generate the graph elements (D, O, R) from the image, and (2) **Graph Grounding** (Section 2.4) aligns the graph elements with the bounding boxes (B) extracted by an object detection model.

2.3 Graph construction

Deconstructing annotations. BACON assists downstream models in understanding complex texts by decomposing the annotations of VLMs into

basic elements and then combining them according to a specific structure. Our approach is divided into two key parts. Firstly, we design a reversible transformation method that converts BACON into a string format understandable by VLMs. Illustrative examples are provided in Appendix A.2.2. Then, we employ the ICL technique to teach VLMs to output following the string format we designed. We find a few simplified examples sufficient and execute the ICL learning process in one conversational exchange. In practice, we use GPT-4V as the VLM and provide the final instruction in Appendix A.2.2. Based on this approach, we develop the BACON dataset, detailed in Section 3.

BACON-Captioner. Obtaining BACON by GPT-4V is reliable but expensive. Therefore, we opt to fine-tune a 13B LLaVA model on the BACON dataset to serve as a specialized captioner, with training details in Section A.4.1. Our goal for this captioner is to match GPT-4V’s capability in generating BACON. We gather BACON outputs from BACON-Captioner and GPT-4V on a test set, analyzing the categories each detects and the root nouns and verbs of their outputs. The resulting distributions, displayed in Figure 9, show a high degree of similarity between them. Furthermore, the precision and recall score calculated by manual annotation (the metrics are detailed as Section 4.2.2) show BACON-Captioner achieve 91% of precision score and 90% of recall score of that of GPT-4V. Consequently, BACON-Captioner is a viable alternative to GPT-4V for producing BACON and helps us extend BACON dataset.

Beyond generating Bacon from images, we sur-

prisingly find that the trained captioner is also adept at performing additional useful tasks without fine-tuning, such as interactively modifying the items of BACON, transforming prompts into BACON format, envisioning scenarios in the style of BACON.

2.4 Graph grounding

Space information is important for representing an image. While VLMs have some level of grounding, specialized models like Grounding DINO (Liu et al., 2023b) significantly outperform them. Therefore, we did not pursue the option of VLMs in Section 2.3 delivering location details, but rather, we intend to leverage dedicated grounding models for this purpose. Fortunately, BACON’s structure provides a list of objects required by grounding models, enabling the combination of advanced VLMs for detailing and top-tier grounding models for precise localization within BACON.

Although Grounding DINO provides accurate object positions, names alone fall short of distinguishing objects within the same category. Here, BACON’s detailed node descriptions come into play, allowing for precise region identification when used in conjunction with CLIP. Moreover, we enhance grounding accuracy by first applying LLaVA to filter out incorrect bounding boxes before proceeding with the CLIP step. We conducted an ablation study on the BACON benchmark (with details in Section 3), and the findings, presented in Figure 3 (b), confirm the benefits of incorporating CLIP and LLaVA into our approach. See Figure 3 (a) for an illustration of this process.

3 BACON dataset

In this section, we introduce the BACON dataset proposed in this paper. BACON dataset is composed of two parts, the training set and the test benchmark, which share different collection methods.

Training set. For the training set, we employ the method as detailed in Section 2.3 to collect 110k BACON-image pairs. Then, we engage in a thorough manual re-annotation process to eliminate ambiguities and incorrectness (see details in Appendix A.3). The end product is a refined dataset of 100k high-quality image-BACON pairs. Notably, BACON’s structure greatly streamlines the annotation process. By breaking down GPT-4V’s output into manageable sections, we can assign specific segments to different annotators, which concentrates their efforts and enhances productivity. Of-

Table 1: **Comparison of open-vocabulary object detection** among BACON, Grounding DINO, open-vocabulary object detection models, and grounding caption models on BACON benchmark. GD represents Grounding DINO. We have calculated error bars for models that exhibit randomness.

Method	AP50(\uparrow)	Recall(\uparrow)	mIOU(\uparrow)
OV-DQUO	14.4%	17.1%	75.3%
DE-VIT	38.2%	32.5%	76.3%
GD	$35.0 \pm 0.5\%$	$20.5 \pm 0.2\%$	$75.9 \pm 0.1\%$
Next-Chat	$30.1 \pm 2.4\%$	$8.0 \pm 0.1\%$	$67.2 \pm 0.0\%$
Kosmos-2	$35.8 \pm 4.4\%$	$13.2 \pm 2.2\%$	$76.1 \pm 0.4\%$
GLaMM	36.49%	20.79%	80.06%
BACON	$38.9 \pm 0.9\%$	$36.2 \pm 0.7\%$	$80.6 \pm 0.4\%$

ten, their tasks involve simple verifications like checking colors or relationships. Furthermore, the comprehensive nature of BACON generally spares annotators from generating new information, allowing them to focus on editing or deleting existing content—tasks that are typically less complex.

Test benchmark. Despite the existence of graph datasets like Visual-Genome (VG) (Krishna et al., 2017) and PSG (Yang et al., 2022), there is still no dataset that concurrently offers open-vocabulary capabilities, detailed object attributes, and a comprehensive overall description. To address this deficiency, we develop the BACON benchmark.

Despite GPT-4V’s advanced capabilities, it may miss objects in images. For the test benchmark, aiming for the utmost accuracy, we utilize a rigorous method that relies heavily on human annotators in contrast to the training set. This method, as shown in Figure 11, starts by using SAM (Kirillov et al., 2023) to separate each object in an image. Next, it applies VLMs to describe all objects, list their attributes, and explain how these objects interact with each other. Finally, given the object list, VLM provides the overall description of the image. We leave the details in the appendix. Thanks to SAM’s excellent performance, it is rare to miss objects in the image. The challenge comes with VLM’s difficulty in recognizing objects given their masked images, which is still hard even for advanced VLMs like GPT-4V (OpenAI, 2023). This difficulty leads to a lot of manual annotation to correct mistakes. The high cost of manual corrections is the reason why this method is not applied to create our training set. Finally, we annotated a test benchmark containing around 4k images, 40k objects, and 200k relationships.

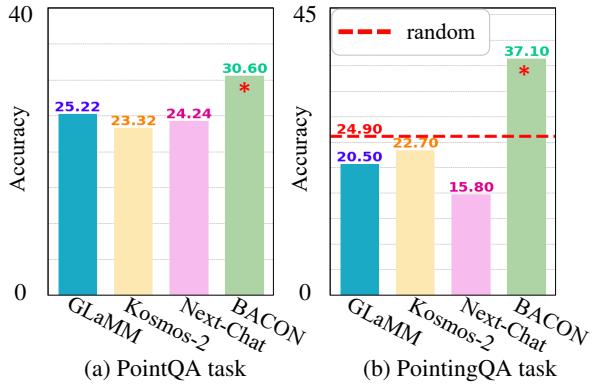


Figure 4: **Quantitative comparison on (a) PointQA and (b) PointingQA** between BACON and baselines.

Table 2: **Comparison on open-vocabulary scene graph generation task** between BACON and multiple baselines on VG dataset and BACON benchmark. The number of correct predictions is used as the metric (\uparrow).

Method	Visual-Genome	BACON dataset
PSGFormer	0.24K	8.57K
PSGTR	0.34K	10.13K
IMP	2.81K	17.45K
Gps-net	2.99K	17.65K
Motifs	2.97K	17.70K
VCTree	2.96K	17.80K
BACON	6.89K	64.30K

4 Experiments

In this section, we show that BACON can be applied to help multiple downstream tasks by flexibly utilizing desired parts of information and the remarkable capabilities of BACON-Captioner. First, in Section 4.1, we show BACON can help a wide range of downstream tasks by flexibly utilizing desired parts of BACON. Then, for completeness, we discuss some special cases in Section 4.2 where BACON can be directly utilized without any special operations. Next, in Section 4.3, we show some impressive capabilities of BACON-Captioner. Finally, in Section 4.4, we show BACON can be easily generalized to video dense-captioning task.

4.1 Downstream tasks benefiting from BACON

The adaptable nature of BACON’s structure enhances models’ comprehension of complex text and empowers them to undertake tasks previously beyond their reach. We conduct evaluations across five downstream tasks, including object detection (Section 4.1.1), point question answering (PointQA) (Section 4.1.2), Pointing question answering (PointingQA) (Section 4.1.3), scene graph generation (SGG) (Section 4.1.4), and image generation (Section 4.1.5).

Table 3: **Quantitative comparison of VQA task** between BACON and multiple VLM-based baselines, where the input image of the QA model is replaced by its caption to evaluate the performance of the captioner. The metric is the accuracy of answering questions (\uparrow).

Method	NLVR2	OK-VQA	VQAv1	VQAv2
LLaVA	56.3 %	30.9 %	50.0 %	64.1 %
ShareGPT-4V	57.5 %	31.4 %	50.7 %	65.4 %
Qwen-VL-max	56.8 %	29.0 %	46.0 %	61.1 %
BACON	59.1 %	32.1 %	52.6 %	66.4 %

4.1.1 Open-vocabulary object detection

Grounding DINO struggles to differentiate between multiple objects of the same category during open-vocabulary object detection (OVD) task (Examples are shown in Appendix A.4.2). However, thanks to the structural features of BACON, the approach described in Section 2.4 can be leveraged to address this issue, significantly enhancing Grounding DINO’s performance in OVD task. We compare the performance with grounding caption models, including GLaMM (Rasheed et al., 2024), Kosmos-2 (Peng et al., 2023), Next-Chat (Zhang et al., 2023), and two of the SOTA dedicated OVD models, OV-DQUO (Wang et al., 2024a) and DEVIT (Zhang et al., 2024) as baselines on BACON benchmark, with AP50, recall and mIOU used as the metrics. Results in Table 1 show that BACON enables grounding models to perform OVD task and outperform all evaluated baselines.

4.1.2 Point question answering

Point Question Answering (PointQA) (Mani et al., 2020) aims to answer region-related questions according to an input image. Inspired by the setting of zero-shot visual question answering (ZS-VQA) (Guo et al., 2023), **we replace the input of the QA model from images to their captions** to evaluate the performance of captioners. Intuitively, if a fixed QA model can correctly answer more questions using a given caption, it suggests that the caption contains more accurate information, indicating a superior captioning performance.

This task is extremely difficult for VLMs like LLaVA because they can’t match parts of the caption to image regions. But BACON aids LLaVA in finding relevant descriptions for a given area (we’ve detailed the method in Appendix A.4.3). We use grounding caption models, including GLaMM, Kosmos-2, and Next-Chat, as baselines. Besides, we apply LLaVA as the fixed QA model, and utilize the QA accuracy as the metric. Results on

Table 4: **Accuracy in depicting objects (A_o) and relationships (A_r) in images generated from text prompts**, as evaluated by human. We compare SDXL enhanced by BACON with SDXL and DALL-E 3.

Method	$A_o(\uparrow)$	$A_r(\uparrow)$
SDXL	$59.20 \pm 3.96\%$	$41.45 \pm 3.46\%$
DALL-E 3	$90.05 \pm 4.17\%$	$71.60 \pm 3.39\%$
BACON + SDXL	$95.20 \pm 1.13\%$	$76.65 \pm 0.92\%$

LookTwice-QA dataset (Mani et al., 2020) shown in Figure 4 (a) indicate BACON can help VLMs associate descriptions with regions to perform PointQA tasks and surpass all baselines.

4.1.3 Pointing question answering

Pointing Question Answering (PointingQA) (Zhu et al., 2016) asks models to choose the correct region from various options based on a given question, using only the input image. Following the settings in Section 4.1.2, we replace the input from images to corresponding captions to evaluate the performance of captioners. BACON can help LLaVA with the task, which is originally beyond its capabilities (details in Appendix A.4.4). Again, we chose the grounding caption models, including GLaMM, Kosmos-2, and Next-Chat, as baselines and used the accuracy of selecting regions as the metric. Results on Visual-7W dataset (Zhu et al., 2016) in Figure 4 (b) show that BACON can help LLaVA find the correct region according to the requirement and significantly outperform the baselines, whose performances are even worse than randomly selecting.

4.1.4 Scene graph generation

The scene graph generation (SGG) task focuses on identifying (subject-predicate-object) triplets in images. Traditional SGG models are trained to classify items in the triplet within a finite set of categories, which may not be suitable for real-world applications. However, the open-vocabulary (OV-SGG) setting, closer to the real situation, is extremely hard for those SGG models. Fortunately, BACON-Captioner can perform the OV-SGG task without any fine-tuning as BACON provides the relationships part. We compare the performance of OV-SGG between BACON-Captioner with multiple specialized SGG approaches including Motifs (Zellers et al., 2018), GPS-Net (Lin et al., 2020), VCTree (Tang et al., 2019), PSGTR, PS-GFormer (Yang et al., 2022), and IMP (Xu et al., 2017) on two datasets, VG (Krishna et al., 2017)

(see details in Appendix A.4.5) and BACON benchmark. To calculate the open-vocabulary metrics, we apply CLIP similarity to decide whether a prediction is correct. The number of correct predictions is used as the evaluation metric (as detailed in Appendix A.4.5). The results shown in Table 2 demonstrate that BACON can perform the OV-SGG task, which is out-of-reach for VLMs and substantially outperforms purpose-built SGG models.

4.1.5 Image generation

Advanced text-to-image models like SDXL (Podell et al., 2023) struggle to follow complex text prompts and accurately generate images. Fortunately, BACON allows generative models to separately generate the background and different objects and then merge them, thus splitting the challenge into several easy parts (Detailed method can be found in Appendix A.4.6). We show two examples in Figure 5 with more instances available in Appendix A.4.6. To assess the correlation between the text prompts and the generated images, we conduct a user study involving 10 human annotators and 100 samples. They are tasked with counting the significant objects and relationships mentioned in the text prompts and those accurately generated in the images. Utilizing the manual annotation, we compute the accuracy metrics for objects (A_o) and relationships (A_r), with the details in Table 4. The results demonstrate that BACON significantly enhances SDXL’s ability to understand and follow complex prompts. Remarkably, it enables SDXL to surpass DALL-E 3 in faithfully reproducing the details specified in the text descriptions.

4.2 Tasks directly using BACON

In this section, for completeness, we discuss downstream tasks where BACON can be used directly without special operations.

4.2.1 Visual question answering

Following the setting of ZS-VQA and Section 4.1.2, we replace the input of VQA from images to their captions to evaluate the performance of the captioner. We compare BACON-Captioner with multiple VLM-based captioners including LLaVA (Liu et al., 2023a), Qwen-VL-max (Bai et al., 2023), and the distillation version on caption task of GPT-4V, ShareGPT-4V (Chen et al., 2023a) on multiple commonly used VQA datasets including NLVR2 (Suhr et al., 2018), VQAv1 (Antol et al., 2015), VQAv2 (Goyal et al., 2017), and OK-

Caption

At the center of the screen is a man wearing a black suit and a red tie, riding on the back of a dinosaur. The dinosaur stands on the green grass, opening its mouth to reveal its sharp teeth. In the bottom right corner of the screen, a photographer wearing a black coat squats, focusing on capturing this unique scene with a camera. The mid shot features two trees with sparse leaves. In the background, there are two colorful hot air balloons flying in the sky. One is black, located in the upper right corner of the screen, and the other is colorful, located slightly higher in the middle of the screen

The full moon hangs high on the left side of the sky, with stars dotted around it. The background on the left side of the picture is a magnificent castle, standing on a distant hill. In the bottom left corner of the photo, two people are standing hand in hand on a rock by the river. Men are on the left, women are on the right, and the river winds from their feet towards the distance. On the right side of the picture is a male angel holding a trumpet and spreading his wings. Above the angel, there is a glowing flying saucer suspended in the air.



Figure 5: **Comparative examples of image generation** reveal that BACON enhances advanced generative models like SDXL, SDXL and DALL-E 3 struggle with complex text and fail to produce corresponding images. Remarkably, BACON not only elevates SDXL’s image quality but also markedly boosts its comprehension of intricate instructions, enabling it to surpass DALL-E 3 in terms of accurately generating images aligning with textual directives.

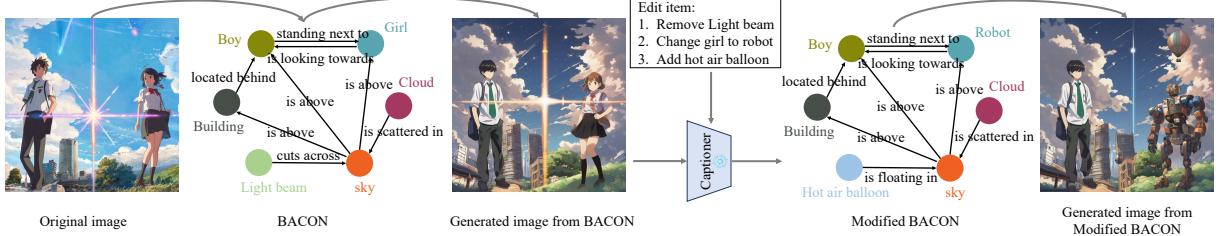


Figure 6: **An example of interactively modifying BACON using BACON-Captioner.**

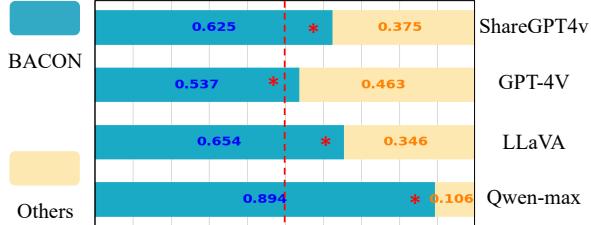


Figure 7: **Win rate of pairwise comparisons** between BACON-Captioner and other VLM-based captioners.

Table 5: **Precision & recall score calculated by manual annotation** between BACON-Captioner and other VLM-based captioners.

Method	Precision	Recall
LLaVA	$36.35 \pm 1.48\%$	$59.15 \pm 4.74\%$
ShareGPT-4V	$23.20 \pm 3.82\%$	$55.30 \pm 2.12\%$
Qwen-VL-max	$35.20 \pm 5.94\%$	$57.50 \pm 1.98\%$
GTP4v	$21.5 \pm 0.71\%$	$70.55 \pm 13.36\%$
BACON	$56.23 \pm 4.23\%$	$82.83 \pm 8.33\%$

VQA (Marino et al., 2019). For evaluation measurement, we follow the official evaluation metrics to measure the performance. As shown in Table 3, BACON outperforms all baselines.

4.2.2 Precision & recall and user study

To assess the performance of the BACON-Captioner, we conduct a user preference user study and examine the precision and recall score by manual annotation. We compare BACON-Captioner with various VLM-based captioners, including LLaVA, Qwen-VL-max, ShareGPT-4V, and GPT-4V, by analyzing captions produced for a randomly sampled set of 200 images from the MSCOCO dataset (Lin et al., 2014). We engage 10 human annotators for manual labeling. For the precision and recall scores, we first extract all important nouns existing in the captions (see details in Appendix A.4.7) and then ask

annotators to count the number of objects in the image and the number of correct predictions in the extracted nouns. Then, the precision and recall score can be calculated. In the user preference study, annotators select their preferred annotation in pairwise comparisons, ensuring structural aspects are neutralized to prevent any biases. The outcomes, as shown in Figure 7 and Table 5, indicate BACON outperforms all comparisons in general, **notably predicting more correct objects than multiple popular VLMs even containing GPT-4V**.

4.3 Additional capabilities of captioner

Beyond obtaining BACON from images, the trained captioner is also adept at performing additional useful tasks, including interactively editing BACON, transforming ordinary prompts into BACON format, and planning positions of objects in BACON. First,

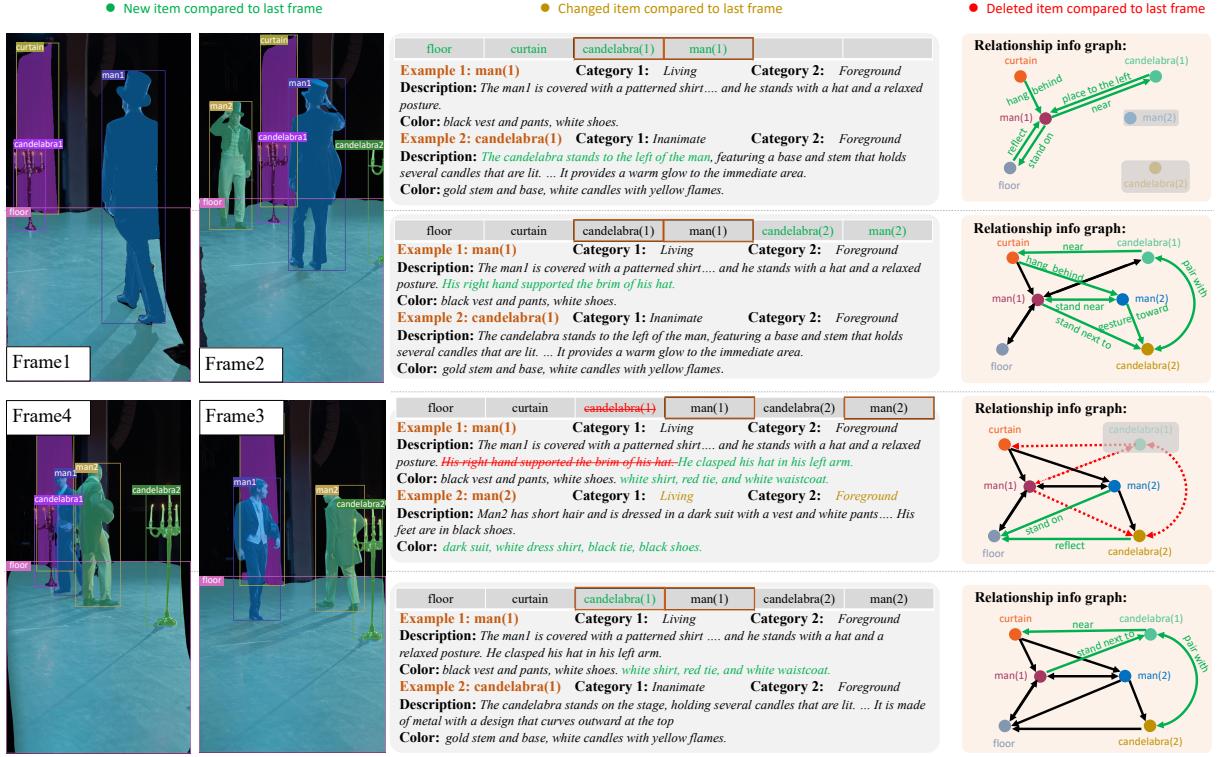


Figure 8: **An example of BACON on video captioning**, which includes three components: an overall description, an object list, and their relationships, each dynamically evolving over time. With respect to a prior frame, updates are color-coded: new elements in green, removed in red, altered in gold, and persistent ones in black. BACON thus adeptly captures the temporal changes and salient details of each video frame, while its structured nature potentially aids in downstream model comprehension.

as shown in Figure 6, BACON-Captioner allows us to interactively edit the BACON, and thus affect the generation of images.

Subsequently and remarkably, without any fine-tuning, BACON-Captioner can transform an ordinary prompt into a BACON. For short prompts, it can add details to create a BACON, and for longer, it can organize the given information into a BACON. Moreover, the BACON-Captioner can arrange the positions of objects of the BACON. We provide examples of both expanding and organizing prompts in Appendix A.4.8. We quantitatively compare BACON-Captioner’s capability of planning with LayoutGPT (Feng et al., 2024) on the MSCOCO dataset (Lin et al., 2014) and BACON datasets, using the mIOU, precision and recall metrics (Feng et al., 2024) as detailed in Appendix A.4.8. Results in Table 6 indicate that BACON-Captioner performs better than LayoutGPT on both evaluated datasets.

4.4 BACON on video captioning

While BACON is primarily developed for image data, it can be extended to create structured captions for videos with the help of additional techniques that address the temporal dimension of

Table 6: **Comparison of plan task** between BACON and LayoutGPT (Feng et al., 2024) on both MSCOCO (Lin et al., 2014) and BACON benchmark.

Dataset	Method	Precision	Recall	mIOU
MSCOCO	LayoutGPT	68.82%	38.30%	4.14%
	BACON	70.42%	39.90%	6.71%
Bacon Dataset	LayoutGPT	48.83%	27.57%	7.86%
	BACON	51.10%	39.80%	15.65%

video content. The principal challenge of adapting BACON for videos is distinguishing between elements that change over time and those that remain constant. To resolve this, we employ a tracking method (Cheng et al., 2023) to identify and match the same object across different frames. Furthermore, we use T5 (Raffel et al., 2020) as the text encoder to compare descriptions of the same object or scene segment across frames. Parts of the text with high similarity scores are considered stable, while those with low similarity are deemed to have changed. An example shown in Figure 8 demonstrates this approach efficiently captures the continuity and evolution of video content, providing a coherent and descriptive narration. Additional examples are available in Appendix A.4.9.

5 Conclusion

In this paper, we introduce BACON to address the difficulty of downstream models in understanding complex texts, by breaking down complex annotations into basic minimum elements and presenting them in a graph structure. We propose a novel method for obtaining the structural representation from images, and construct a dataset of 100k scale based on it. A captioner is trained on this dataset with multiple remarkably useful capabilities. Extensive experiments demonstrate that our method can effectively assist downstream models in accomplishing tasks they previously could not achieve or excel in their current cutting-edge solutions.

6 Limitations

This paper introduces a method designed to assist smaller models in comprehending complex texts and to facilitate their integration with VLMs, achieving remarkable performances across multiple benchmarks. However, despite these achievements, our approach still faces certain limitations. Firstly, given the absence of a fully automated method that guarantees reliable quality, our data collection process still necessitates human annotation involvement. Secondly, due to cost and resource constraints, the captioner’s localization capabilities remain insufficient, necessitating the combination of a grounding model to obtain high-quality positional information.

References

- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *IEEE Conf. Comput. Vis. Pattern Recog.*
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Int. Conf. Comput. Vis.*
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.
- James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn.openai.com/papers/dall-e-3.pdf*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Adv. Neural Inform. Process. Syst.*
- Yashoda Chaulagain. 2018. Visual position and juxtaposition: an analytical study of liberty leading the people and moon-woman cuts the circle. *Tribhuvan University Journal*.
- Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Duhua Lin. 2023a. ShareGPT4V: Improving large multimodal models with better captions. *arXiv preprint arXiv:2311.12793*.
- Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. 2024. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6593–6602.
- Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2023b. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.
- Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. 2023. Tracking anything with decoupled video segmentation. In *Int. Conf. Comput. Vis.*
- George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*.
- Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2024. Layoutgpt: Compositional visual planning and generation with large language models. *Adv. Neural Inform. Process. Syst.*
- Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc’Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. *Adv. Neural Inform. Process. Syst.*
- Hanan Gani, Shariq Farooq Bhat, Muzammal Naseer, Salman Khan, and Peter Wonka. 2024. LLM blueprint: Enabling text-to-image generation with complex and detailed prompts. *Int. Conf. Learn. Represent.*
- 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 *IEEE Conf. Comput. Vis. Pattern Recog.*

- Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. 2023. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *IEEE Conf. Comput. Vis. Pattern Recog.*
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016a. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016b. Identity mappings in deep residual networks. In *Eur. Conf. Comput. Vis.*
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*
- Justin Johnson, Agrim Gupta, and Li Fei-Fei. 2018. Image generation from scene graphs. In *IEEE Conf. Comput. Vis. Pattern Recog.*
- Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. 2015. Image retrieval using scene graphs. In *IEEE Conf. Comput. Vis. Pattern Recog.*
- Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798.
- Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. 2024. Segment anything in high quality. *Adv. Neural Inform. Process. Syst.*
- Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*
- Janghyeon Lee, Jongsuk Kim, Hyounguk Shon, Bum-soo Kim, Seung Hwan Kim, Honglak Lee, and Junmo Kim. 2022. Uniclip: Unified framework for contrastive language-image pre-training. *Adv. Neural Inform. Process. Syst.*
- Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. 2021. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Eur. Conf. Comput. Vis.*
- Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. 2020. GPS-Net: Graph property sensing network for scene graph generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *Adv. Neural Inform. Process. Syst.*
- Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023b. Grounding Dino: Marrying Dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.
- Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2016. Visual relationship detection with language priors. In *Eur. Conf. Comput. Vis.*
- Arjun Mani, Nobline Yoo, Will Hinthon, and Olga Russakovsky. 2020. Point and ask: Incorporating pointing into visual question answering. *arXiv preprint arXiv:2011.13681*.
- Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *IEEE Conf. Comput. Vis. Pattern Recog.*
- Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conf. Comput. Vis. Pattern Recog.*
- Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2021. SDEdit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.
- George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*.
- OpenAI. 2023. GPT-4V(ision) technical work and authors. <https://openai.com/contributions/gpt-4v/>.
- Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shao-han Huang, Shuming Ma, Qixiang Ye, and Furu Wei. 2023. Grounding multimodal large language models to the world. In *Int. Conf. Learn. Represent.*
- Brant Pitre. 2015. Jesus and the last supper.

- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn.*
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*
- Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdela Rahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S Khan. 2024. Glamm: Pixel grounding large multimodal model. *IEEE Conf. Comput. Vis. Pattern Recog.*
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*
- Vishnu Sarukkai, Linden Li, Arden Ma, Christopher Ré, and Kayvon Fatahalian. 2024. Collage diffusion. In *IEEE Winter Conf. Appl. Comput. Vis.*, pages 4208–4217.
- Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. 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*.
- Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2018. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*.
- Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. 2019. Learning to compose dynamic tree structures for visual contexts. In *IEEE Conf. Comput. Vis. Pattern Recog.*
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *IEEE Conf. Comput. Vis. Pattern Recog.*
- Junjie Wang, Bin Chen, Bin Kang, Yulin Li, YiChi Chen, Weizhi Xian, and Huifeng Chang. 2024a. Ov-dquo: Open-vocabulary detr with denoising text query training and open-world unknown objects supervision. *arXiv preprint arXiv:2405.17913*.
- Zhenyu Wang, Enze Xie, Aoxue Li, Zhongdao Wang, Xihui Liu, and Zhenguo Li. 2024b. Divide and conquer: Language models can plan and self-correct for compositional text-to-image generation. *arXiv preprint arXiv:2401.15688*.
- Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. 2017. Scene graph generation by iterative message passing. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5410–5419.
- Jingkang Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. 2022. Panoptic scene graph generation. In *Eur. Conf. Comput. Vis.*
- Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. 2024. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal LLM. *Int. Conf. Mach. Learn.*
- Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. Neural motifs: Scene graph parsing with global context. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5831–5840.
- Ao Zhang, Liming Zhao, Chen-Wei Xie, Yun Zheng, Wei Ji, and Tat-Seng Chua. 2023. Next-chat: An LLM for chat, detection and segmentation. *arXiv preprint arXiv:2311.04498*.
- Xinyu Zhang, Yuting Wang, and Abdeslam Boularias. 2024. Detect everything with few examples. *arXiv preprint arXiv:2309.12969*.
- Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *IEEE Conf. Comput. Vis. Pattern Recog.*

A Appendix

The appendix is organized into five distinct sections. It begins by presenting a comprehensive overview of related work in Appendix A.1. This is followed by an in-depth discussion of the methodology in Appendix A.2. The next section, Appendix A.3, focuses on the aspects of human annotation within the dataset collection process for BACON. Appendix A.4 then provides a thorough explanation of the experimental setup, detailing the metrics used and the methodology for implementing BACON in solving downstream tasks. Moreover, this section includes additional experimental findings.

A.1 Related works

Image description Transforming images into textual descriptions is a quintessential task, with many downstream applications relying on accurate text-based representations of visual data (Radford et al., 2021; Antol et al., 2015; Podell et al., 2023; Betker et al., 2023; Liu et al., 2023b; Frome et al., 2013). Traditional methods for image captioning usually suffer the problem of lacking details (Anderson et al., 2018; Mao et al., 2016; Kazemzadeh et al., 2014; Sharma et al., 2018; Vinyals et al., 2015). Recently, VLM-based models including GPT-4V (OpenAI, 2023), LLaVA (Liu et al., 2023a), ShareGPT-4V (Chen et al., 2023a), InternVL (Chen et al., 2023b), Qwen-VL (Bai et al., 2023) and so on, can describe images in very detail and output very long captions, which have been proved helpful in the downstream fields, such as image generation (Betker et al., 2023).

Decomposing description However, overly complex texts pose significant challenges to the comprehension abilities of downstream models. Take the domain of text-to-image generation as an example, where representative models like SDXL (Rombach et al., 2022) struggle to accurately generate images based on detailed textual guidance. This has spurred a plethora of efforts aimed at decomposing and reassembling texts to enhance models' fidelity to intricate instructions (Yang et al., 2024; Wang et al., 2024b; Gani et al., 2024). There are also works trying to describe the image in a graph structure, including grounding caption models (Peng et al., 2023; Zhang et al., 2023; Rasheed et al., 2024) and scene graph generation (SGG) models (Xu et al., 2017; Zellers et al., 2018; Lin et al., 2020; Tang et al., 2019; Yang et al., 2022).

A.2 Supplementary of BACON

In this section, we will delve into the details of the BACON method as a supplement to the main text. We will organize this section following the structure of the main text, including additional insights into the design philosophy behind BACON in Appendix A.2.1, as well as showcasing complete examples. In Appendix A.2.2, we will present the details and examples of the graph construction.

A.2.1 Design of BACON

As mentioned in the main text, when breaking complex annotations down into basic elements, we specifically considered information that could be beneficial for downstream tasks as these basic elements. We then listed these basic elements alongside the potential domains that might benefit from them, as illustrated in Section 6. Besides, we provide several complete examples of BACON in Figures 13 and 14, which are omitted in the main paper to save space.

A.2.2 Graph construction

In the main paper, we introduced the method of constructing graphs using VLMs but omitted some details due to space constraints. The method encompasses two key steps. The first is to design a string format along with a reversible conversion method that can transform BACON into the string format understandable by VLMs, while also ensuring that this string format can be converted back to BACON without loss. The second step involves utilizing the ICL technique to teach the VLMs to describe images using our designed string format. We will now introduce these two steps in detail.

BACON in string format We translate the graph structure of BACON into a string format interpretable by VLMs (we use GPT-4V in practice), demarcating various sections with unique symbols, where an example is shown in Figure 15. Specifically, we label main titles with %% and subtitles with &&. When listing objects, we enclose extra details like category, description, and color in brackets (). Each detail is separated by a semi-colon ;. We mark the name of an object with <>. During the description of relationships, we use <> for showing objects and [] for the predicate. Additionally, we use <> to highlight important objects within the object, serving multiple purposes. One such function is to post-process the GPT-4V output results. This involves removing foreground information from the background description by

Part of information	Potential areas of benefit
Style	Art classification, Style transfer, Aesthetic evaluations
Theme	Recommendation systems, Thematic analysis
Background	Scene reconstruction, Virtual reality
Foreground	Focus analysis, Subject tracking, Interactive applications
Object name	Object detection, Labeling, Database organization, Language-based image retrieval
Object category	Semantic segmentation, Hierarchical classification
Life status	Behavior Prediction, human-computer interaction, Living identification
Depth cues	Depth estimation, Layered generation, Compositing
Object description	Fine-grained recognition, region captioning, region multi-modal understanding, Image generation
Object color	Color-based retrieval, Fashion industry applications, Image coloring
Relationships	Scene understanding, Action recognition

Table 7: **List** of different important parts of information and the potential fields that may benefit from them. Depth cues represent the foreground-background information of objects.

deleting sentences where the foreground objects appear, or similarly, eliminating background information from the foreground description. By using these special symbols to separate different sections, we can effortlessly organize the string format of BACON into a dictionary using regular expressions. This makes it easy for downstream tasks to extract various pieces of information without any hassle.

Instruction for GPT-4V to obtain BACON Then, we discuss the method of applying ICL technique to obtain BACON via VLMs (we use GPT-4V in practice) in detail. As described in Section 2.3, we discover that GPT-4V does not require exhaustive examples to master the desired format. We simply need to insert a few important examples in the right spots within the instruction, which then play a key role. You can see the final instruction in Figure 16, where we’ve highlighted the critical examples in orange. Among the examples used, some are specific and others are more general. We’ve observed that for straightforward structural elements, general examples are quite effective. For instance, just a few lines, like ‘lines 3-4’ or ‘lines 8-9’, can adequately indicate the use of special symbols in a section, eliminating the need for a full-fledged example. In lines 21-22, we present a general example that clearly delineates the structure of each object, which significantly minimizes GPT-4V’s errors. To keep object details easy to grasp, we use a general example lines 23-24, which are sufficient for producing simple sentences. Regarding lines 27-28, a general example is enough to instruct GPT-4V on the basic pattern for depicting relationships. Lastly, a general example set out in line 29 aids in preventing GPT-4V from repeatedly generating two-way relationship pairs.

However, our high demands on the content and structure are extremely hard even for GPT-4V.

Therefore, GPT-4V sometimes gets details wrong, like missing special symbols, even when we use general examples. That’s why we need to use specific examples to make sure GPT-4V really gets the structure. Take numbering items in the same category, for instance, we introduce a specific example in lines 14-15. Without this example, GPT-4V tends to forget to number the items correctly, even though we’ve already required it in lines 13-14. Also, we noticed GPT-4V does well with the format of the first section but often slips up with the second and third parts, which complicates turning the data into a dictionary. By providing only one clear example for these sections, GPT-4V is much more likely to produce the right structure. The ICL technique has helped ensure that nearly all of the 110k data entries we’ve gathered are formatted correctly and can be translated into a dictionary format.

A.3 Supplementary of BACON dataset

In this section, we begin by examining the output distribution of BACON-Captioner, as detailed in Appendix A.3.1. Subsequently, we introduce the details of human annotation, covering both the collection of training data and the establishment of test benchmarks.

A.3.1 Output distribution of BACON-Captioner

We show the analysis of the root words and categories detected in the outputs of BACON-Captioner, which can be seen in Figure 9. The result clearly shows that the output pattern of BACON-Captioner is very close to that of GPT-4V. Notably, there’s a 100% overlap in the top 100 frequent nouns, 99% for verbs, and 97% for categories detected by GPT-4V and BACON-captioner. This similarity confirms

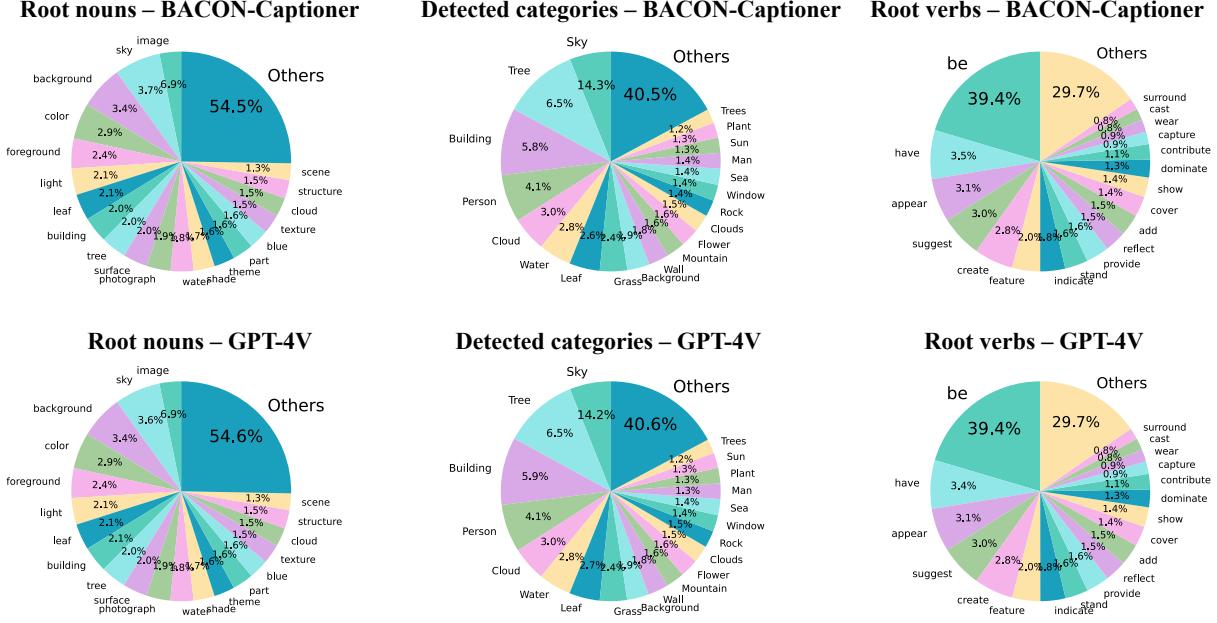


Figure 9: Analyzing the root words and detected categories in BACON’s output on testset: We compare the root words and detected categories generated by BACON-Captioner and GPT-4V, with certain sections magnified for clearer visualization. The results reveal that the output distribution of BACON closely resembles that of GPT-4V.

that BACON-Captioner can effectively take over from GPT-4V in generating BACON from images and extend our BACON dataset.

A.3.2 Human annotation

As we’ve mentioned in Sections 2.3 and 3, creating the BACON dataset’s training and test sets involves human annotations.

Collecting training data. In the process of collecting training data, BACON significantly reduces the workload of annotation. It breaks down the complex descriptions into basic elements, for many of which annotators simply need to make a straightforward judgment of right or wrong, a task that is remarkably simple. For large pieces of information such as background or foreground descriptions, annotators are asked to separately determine if each sentence is correct according to the image. Besides, the annotators are asked to add objects missed by GPT-4V. In this process, the structure we designed for objects can help annotators simplify the description process. They only need to fill in the corresponding information according to the structure.

Collecting test benchmark. In the method of collecting the test set of BACON, annotators are involved in four parts. For the first part, they are expected to correct the result returned by VLMs to recognize the object name given the masked image. In the second and third parts, annotators are asked

Table 8: Complete hyper-parameters of training BACON-Captioner.

Hyper-parameter	Value	Hyper-parameter	Value
Lora rank	128	Learning rate	2×10^{-4}
Epochs	3	Warmup ratio	0.03
Batch size	16	Max length	2048

to separately determine if each sentence is correct. They don’t have to add objects as Segment anything (SAM) (Kirillov et al., 2023) in this method has ensured that there will be no omissions. At the last stage, they have to determine if a relationship is correct and add an important relationship omitted by VLMs.

A.4 Supplementary of experiments

In this section, we provide supplementary explanations for the experimental details omitted in the main text (Section 4), including the training details of BACON-Captioner, the specific manner in which BACON aids downstream tasks, the exact calculation methods for metrics, and any special processing applied to the datasets. We will organize this section following the structure of the main text (Section 4) to facilitate readers in quickly locating the corresponding section for each experiment.

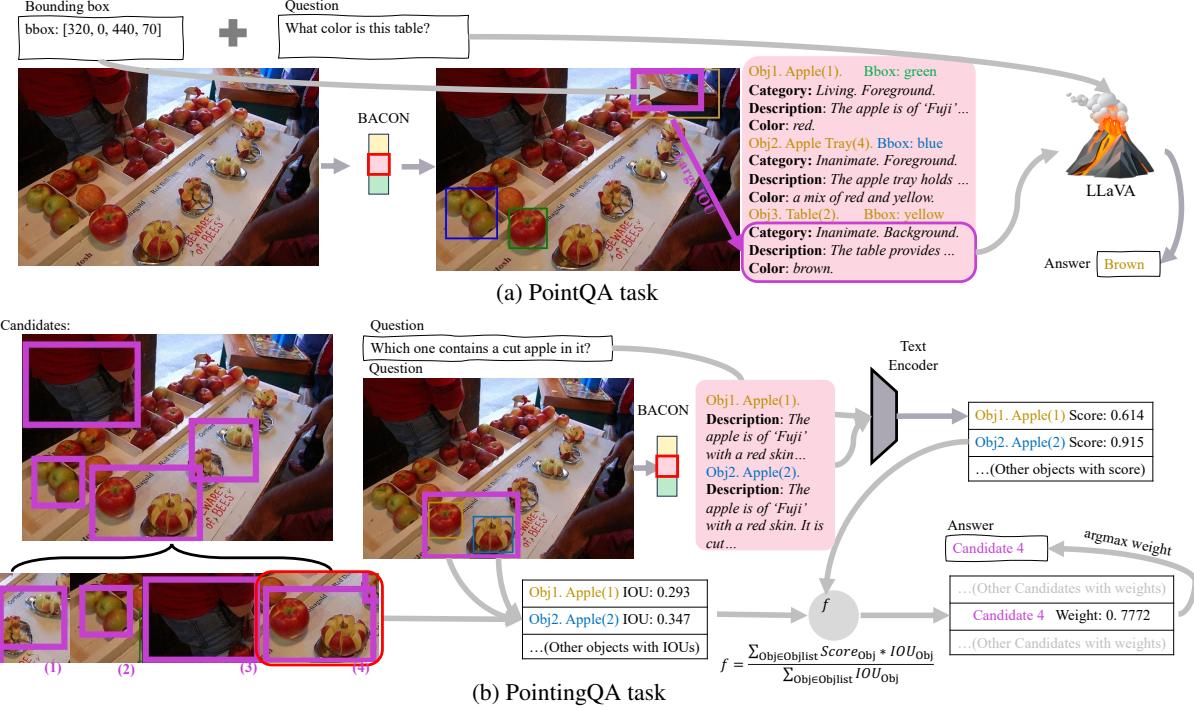


Figure 10: **An illustrative diagram depicting how BACON aids downstream models in executing PointQA and PointingQA tasks.** In (a) the PointQA task, a list of objects and their corresponding descriptions provided by BACON are utilized. The description of the object with the large overlap with the target region is used to represent the description of that region; this regional description is then fed into a QA model to answer questions related to the region. In (b) the PointingQA task, object descriptions provided by BACON are used to calculate similarity scores with the input question, generating scores for each object. Based on the overlap between object positions and candidate regions, a weighted sum of all object scores is computed to assign scores to candidate regions; the region with the highest score is then selected as the prediction.

A.4.1 Details of training BACON-Captioner

BACON-Captioner is fine-tuned on a pre-trained 13B LLaVA model using Low-Rank Adaptation (LoRA) (Hu et al., 2021) technique, where the number of parameters of LoRA adapter is around 0.5B. We follow the default settings of LLaVA-lora fine-tuning, where the detailed hyper-parameters are in Table 8. The captioner is trained on NVIDIA A100 GPUs, taking around 100 GPU hours. We’ve done this in hopes of making it easier for others to reproduce BACON-Captioner.

A.4.2 Open-vocabulary object detection

Although Grounding DINO can carry out open-vocabulary object detection task, it still faces some issues. There are primarily two problems. First, the core step of Grounding DINO requires a noun as input to locate the position of that noun in the image. Moreover, it introduces methods to extract a series of nouns from a sentence description, enabling it to perform object detection tasks. However, the method of extracting nouns can sometimes err, leading to Grounding DINO producing some bizarre

labels. For example, as illustrated in Figure 12, Grounding DINO outputs ambiguous labels such as “one”, “four men one one gray”, “another”.

The second issue, which is more severe, is Grounding DINO’s difficulty in distinguishing between different individuals of the same category. As shown in Figure 12, although Grounding DINO identifies four people, it is challenging to determine which individual is represented by which bounding box with vague labels like “four men one”. Note that the BACON benchmark serves as such a complex benchmark, incorporating numerous scenarios that more closely mirror real-life situations where it is necessary to distinguish different objects within the same or similar categories.

Benefiting from BACON’s powerful capabilities, Grounding DINO can overcome these two issues with the aid of BACON. For the first problem, BACON inherently possesses the ability to identify important objects in an image, allowing Grounding DINO to receive a list of objects from BACON, resulting in a more accurate and comprehensive list of nouns. Regarding the second issue, as intro-

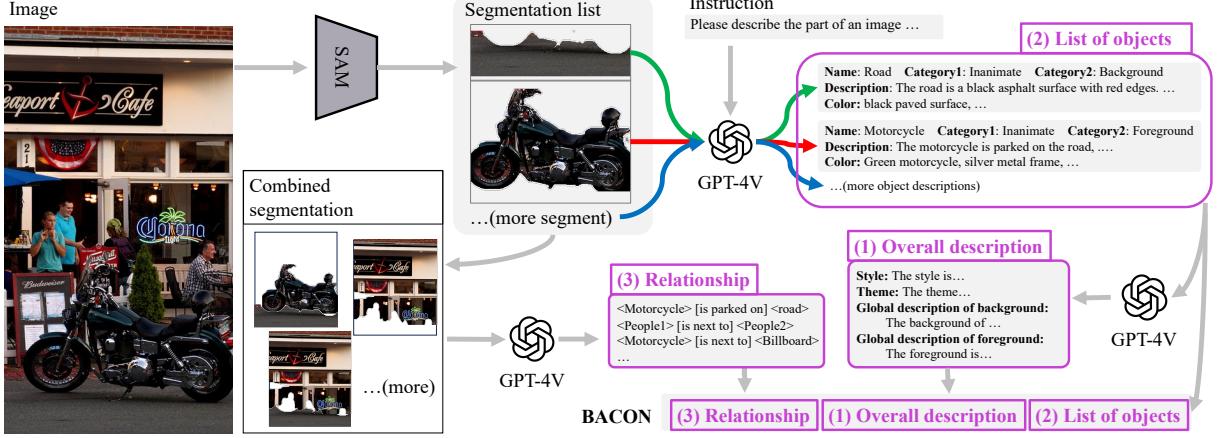


Figure 11: A detailed overview of the method used to collect the BACON benchmark, segmented into five distinct steps. 1) The SAM model segments all components within the image. 2) VLMs identify the names of objects in the masked image obtained from the first step. 3) Using the names identified in the second step, VLMs annotate each object in detail. 4) VLMs generate an overall description of the image based on the list of objects derived from the above steps. 5) images created by randomly pairing two masked images from the first step are fed to VLMs to identify the relationship between the combined segments. It is important to note that human annotation is required to correct and verify the outputs from steps two through five.

duced in Section 2.4, by utilizing the list of objects provided by BACON, along with detailed descriptions of each object, it is possible to post-process Grounding DINO’s predictions. This enables the precise distinction of different individuals within the same category label.

A.4.3 Point question answering

Method of applying BACON. In our experiment, PointQA is designed to answer questions related to image regions based on the description of the image. Most descriptions provided by Visual Language Models (VLMs) cannot accomplish this task as their descriptions lack positional information. However, BACON provides both the positional information of objects within the image and their corresponding descriptions. Given a target area, by combining descriptions of different objects based on their positional relationships, one can create a description relevant to the location. Specifically, as illustrated in Figure 10, we compute the Intersection Over Union (IOU) between the target area and the positions of all objects. By combining the descriptions of objects with high overlap, we obtain a description that is closely related to the target area. Then, we feed this description to the question-answering model to answer the question.

A.4.4 Pointing question answering

Method of applying BACON. The PointingQA task requires selecting the most appropriate region from a set of candidate areas based on a textual

prompt. VLMs struggle to complete this task because they often lack the ability to perceive input location information. However, since BACON decomposes image descriptions into a series of basic elements, each with its corresponding location, we can leverage this feature to accomplish the task. As shown in Figure 10, the method is divided into three steps. First, we calculate the CLIP similarity between each object’s description and the input textual prompt, obtaining scores for each object. The more relevant an object is to the text description, the higher its score. Secondly, we calculate scores for each candidate region by weighting the sum of object scores based on the overlap between the candidate region and the object’s location. The greater the overlap with the candidate area, the larger the proportion of that object’s score. In the third step, the region with the highest score is selected as the answer.

A.4.5 Open-vocabulary scene graph generation

Visual Genome dataset. Visual Genome is an open-vocabulary dataset. However, since most current scene graph generation (SGG) models only consider a certain number of categories, researchers often treat it as a dataset with a limited set of categories. Specifically, they usually identify the most frequent 70 or 150 noun classes, along with the 50 most common predicates, to create a filtered dataset. In our case, as we are working on an

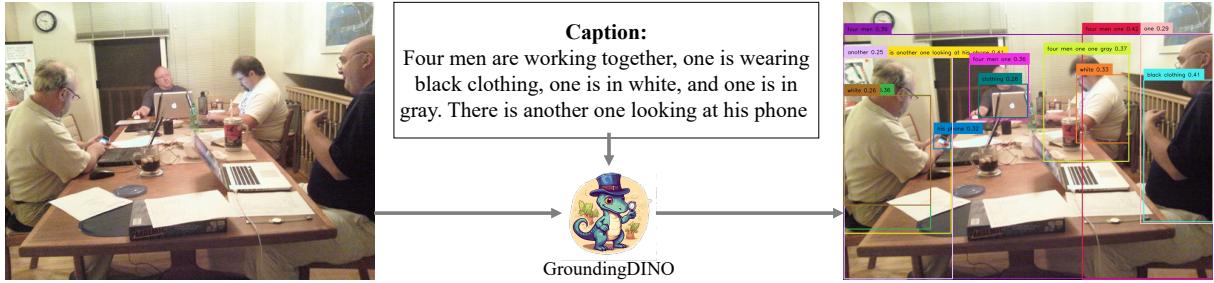


Figure 12: **An example of Grounding DINO undertaking an open-vocabulary task**, where it encounters issues with ambiguous labels and faces challenges in distinguishing between different individuals within the same category.

open-vocabulary scene graph generation (OV-SGG) task, we treat the VG dataset as an open-vocabulary dataset, retaining all of its annotations.

Evaluation metrics. Traditional SGG tasks often use recall-related metrics to evaluate performance, that is, how many (subject-predicate-object) triplets are predicted in an image. However, these metrics cannot be directly applied to assess the performance of open-vocabulary tasks. This is because previous metrics involve performing classification tasks within a set of fixed categories and using the confidence of those classifications to obtain the top K predictions with the highest likelihood. However, in an open-vocabulary scenario, there are theoretically an infinite number of possibilities, making it impractical to calculate and sort all scores. Therefore, we use the number of correctly predicted triplets in the dataset as the evaluation metric. The more triplets are accurately predicted, the better the performance.

Additionally, we use CLIP to determine the correctness of predictions on an open-vocabulary dataset. Given a prediction, it is considered correct as long as the CLIP similarity of its subject, predicate, and object to the corresponding ground truth exceeds a certain threshold (0.85 here), and the IOU between the positions of the subject and object with the ground truth also surpasses a threshold (0.5 here).

A.4.6 Image generation

Method of enhancing SDXL by BACON. Even as one of the most renowned models for text-to-image generation, SDXL often struggles to understand complex prompts and generate precise images accurately. This is primarily because SDXL employs CLIP for text understanding, which limits its ability to comprehend the text. However, each basic element within a complex prompt is not complicated for SDXL to understand and gener-

ate. Therefore, by breaking down complex texts into basic elements, BACON can significantly assist SDXL in simplifying complex tasks. Specifically, SDXL can first create the background, then sequentially generate each object, and finally assemble the different parts. Currently, there are many methods that can be utilized for image stitching, such as Anydoor (Chen et al., 2024), Collage Diffusion (Sarukkai et al., 2024), etc. Sometimes, images can also be directly stitched together and then refined using SDXL as the base model, with SDEdit (Meng et al., 2021) for refining the images, but this typically requires the images to be relatively simple. Aside from generating individual parts of the image and then stitching them together, another approach is to sequentially inpaint (Rombach et al., 2022) objects onto the image using inpainting methods.

More results. We provide more examples in Figure 19

A.4.7 Precision & recall and user study

When calculating precision and recall, it involves identifying which objects have been predicted by different captioners. For other captioners, this can be challenging because directly extracting nouns would include many nouns that cannot be considered objects. Therefore, we utilize VLMs to accomplish this task. Specifically, we input the model’s captions into the VLMs, requesting them to extract the important objects contained within. For BACON-Captioner, this process is straightforward because BACON explicitly provides a list of objects. This also highlights the advantages of BACON.

A.4.8 Additional capabilities of captioner

Evaluation metrics. Evaluating the performance of the planning task is a subject that hasn’t been widely discussed. As one of the pioneers, LayoutGPT (Feng et al., 2024) collected some images from the COCO dataset (Lin et al., 2014), which

have varying numbers of objects of the same category and used precision and recall as evaluation metrics to assess whether the quantity of objects planned is accurate. Inspired by their approach, we have slightly expanded the concepts of precision and recall. We randomly sample 1000 images from COCO and use their official captions as input for either LayoutGPT or BACON-Captioner. Then, we apply precision and recall metrics to assess how many of the objects predicted by different planning methods actually exist in the images, and how many objects present in the images are predicted.

It’s important to note that both the captioner and LayoutGPT operate in an open-vocabulary manner. Hence, we used CLIP to map the open-vocabulary predictions to COCO’s fixed set of categories. Specifically, for an open-vocabulary prediction, we compute its similarity to all categories in COCO, treating the similarity as logits, and then use a softmax function to map it to a category in COCO. If the softmax score for the most likely category exceeds a threshold (0.9 here), we consider the prediction to be correct; otherwise, it is deemed incorrect. In BACON dataset, the situation is quite similar. A slight difference is that the model’s predictions are mapped onto the list of ground truth objects for the current image, rather than a fixed set of categories. Similarly, when the softmax score exceeds a certain threshold, it is considered a correct prediction. Given that BACON benchmark is significantly more challenging than COCO, if the threshold is set too high, almost all predictions would be incorrect; hence, we lowered the threshold to 0.5.

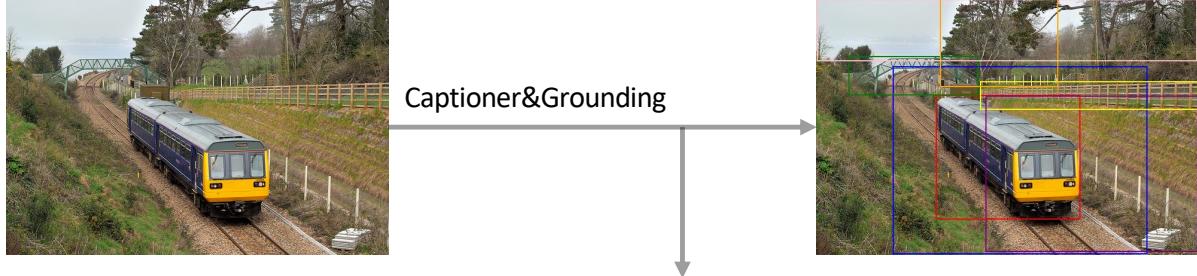
Precision and recall do not take into account the positioning of the planning. This is because evaluating whether a position is appropriate is a subjective task, and so long as it is reasonable, it should suffice. Nonetheless, since the positional distribution in the original images is assuredly reasonable, we can also use the positions in the original images as a certain reference. Therefore, we calculated the mean Intersection Over Union (mIOU) of the positions of the objects in the planning compared to those in the original images, and used this as an evaluation metric.

Examples. We provide two examples (as shown in Figures 17 and 18) to demonstrate how BACON transforms a standard prompt into the format we need. Example in Figure 17 features a simple prompt; in this scenario, the BACON-Captioner often imagines additional objects not present in the

prompt to enrich the overall scene and converts it into the required format. Conversely, Example in Figure 18 involves a more complex input; in such cases, the likelihood of the BACON-Captioner adding new objects is reduced.

A.4.9 BACON on video captioning

We provide more examples (as Figures 20 and 21) as a supplementary of the main paper.



Overall description:

Style: 'The image is a photograph with a realistic style.'

Theme: 'The theme of the image is transportation, specifically a train traveling through a rural landscape.'

Background description: 'The background of the image features a rural landscape with elements of nature and infrastructure. There is a bridge with green metal railings crossing over the train tracks. Beyond the bridge, a fence made of wooden posts and rails encloses a field. The field appears to be grassy with some patches of bare earth. The sky is overcast, with a pale, diffused light suggesting an overcast or cloudy day.'

Foreground description: 'In the foreground, a train is captured in motion on the tracks. The train is painted in a blue and yellow color scheme. The train has multiple carriages, and the windows reflect the surrounding environment. The tracks are made of steel rails with wooden sleepers, and they run parallel to a grassy embankment on the left side of the image.'

Object list:

Train:

Category: inanimate, foreground
 Description: 'The <train> 's body is long and sleek, with <windows> lined along its side. The front <car> has a curved nose with a destination <sign> and <headlights>. The <train> is composed of several <carriages> connected together.'
 Color: blue and yellow
 Position: [200, 160, 441, 367]

Track:

Category: inanimate, foreground
 Description: 'The <track> consists of parallel <steel rails> supported by wooden <sleepers>. It stretches into the distance, guiding the <train>.'
 Color: rusty brown rails, brown sleepers
 Position: [128, 112, 553, 425]

Bridge:

Category: inanimate, background
 Description: 'The <bridge> spans over the <tracks> with a structure made of metal <beams> and <railings>. It appears functional and unadorned.'
 Color: green railings
 Position: [54, 95, 271, 160]

Fence:

Category: inanimate, background
 Description: 'The <fence> is constructed of wooden <posts> and <rails>, enclosing the <field> and providing a boundary.'
 Color: natural wood tone
 Position: [274, 137, 638, 184]

Field:

Category: inanimate, background
 Description: 'The <field> is predominantly grass-covered, with some areas of bare <soil>. It is bordered by the <fence> and <trees>.'
 Color: green grass, brown soil
 Position: [283, 161, 638, 421]

Tree:

Category: inanimate, background
 Description: 'The <trees> have bare branches, indicating a lack of <leaves> which could suggest a seasonal change.'
 Color: dark brown branches
 Position: [207, 0, 404, 146]

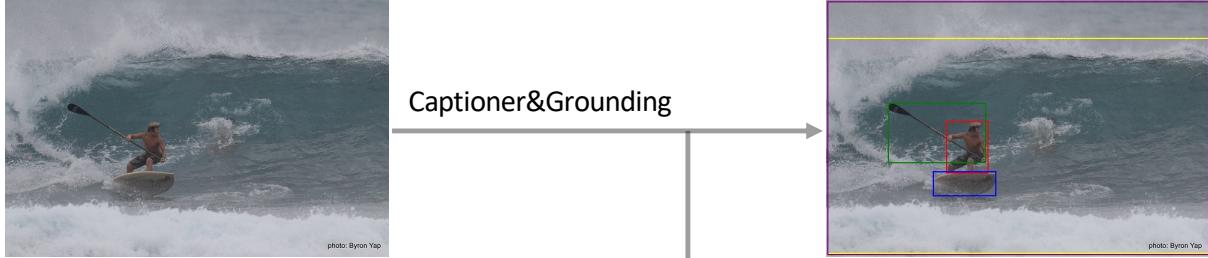
Sky:

Category: inanimate, background
 Description: 'The <sky> is overcast, with a uniform light grey color, suggesting cloudy weather.'
 Color: light grey
 Position: [1, 0, 636, 103]

Relationship:

- <Train> [is traveling on] <track>
- <Train> [is passing under] <bridge>
- <Bridge> [spans over] <track>
- <Fence> [encloses] <field>
- <Field> [is bordered by] <tree>
- <Field> [is bordered by] <Fence>
- <Tree> [is standing in] <Field>

Figure 13: A complete example of BACON.



Overall description:

Style: 'The image is a candid, action-oriented photograph.'

Theme: 'The theme of the image is water sports, specifically surfing.'

Background description: The background of the image is dominated by the ocean with its varying shades of blue and white due to the churning waves and foam. The water appears dynamic and powerful, suggesting a strong surf. There is no clear horizon line, but the waves occupy the majority of the background space, creating a sense of immersion in the ocean environment. The light seems natural, likely from the sun, and it enhances the texture of the water and waves. The ocean is the only element in the background, emphasizing the vastness and the solitary nature of the surfing activity.'

Foreground description: In the foreground, a man is captured in the midst of surfing a wave. He is wearing a cap, shorts, and is shirtless, which indicates a warm climate or season. The man is in a semi-crouched position on a surfboard, navigating the wave with a paddle in his left hand. His facial expression shows concentration and determination. The surfboard is mostly white with a design on its surface, and it cuts through the water, leaving a trail of spray behind. The man's musculature and stance suggest that he is experienced and in control of the surfboard. The water around the surfboard is turbulent, with foam and spray being generated by the movement of the surfboard and the wave.'

Object list:

Man:

Category: living, foreground
 Description: 'The <man>'s <torso> is bare, and he is wearing a <cap> on his <head>. His lower body is covered by <shorts>. He is in a semi-crouched position on the <surfboard>, holding a <paddle> in his left hand. His facial expression shows focus.'
 Color: skin tone, green shorts, white cap
 Position: [200, 201, 270, 288]

Surfboard:

Category: inanimate, foreground
 Description: 'The <surfboard> is under the <man>, supporting him as he rides the <wave>. It has a design on its surface and is cutting through the <water>.'
 Color: predominantly white with a design
 Position: [178, 285, 283, 326]

Paddle:

Category: inanimate, foreground
 Description: 'The <paddle> is held by the <man> in his left hand, assisting him in navigating the <wave>.'
 Color: black shaft, white blade
 Position: [103, 171, 266, 271]

Wave:

Category: inanimate, background
 Description: 'The <wave> is large and powerful, with <water> churning and creating <foam> and <spray> as it breaks.'
 Color: shades of blue and white
 Position: [2, 63, 638, 422]

Ocean:

Category: inanimate, background
 Description: 'The <ocean> fills the background, characterized by its dynamic <waves> and <foam>.'
 Color: various shades of blue
 Position: [2, 2, 638, 424]

Relationship:

- <Man> [is riding] <Surfboard>
- <Man> [is holding] <Paddle>
- <Surfboard> [is cutting through] <Wave>
- <Wave> [is breaking around] <Man>
- <Man> [is surfing on] <Ocean>
- <Ocean> [is supporting] <Surfboard>

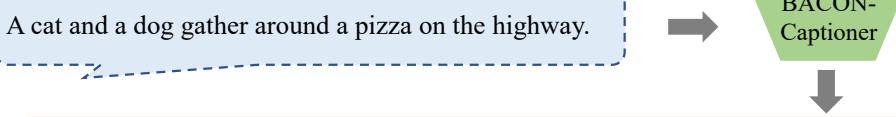
Figure 14: A complete example of BACON.



Figure 15: An example of BACON in string format obtained by GPT-4V.

1 Hello, I would like to ask for your help in describing an image. Please note that I would like the description to be as
 2 detailed as possible. Please strictly respond following my instructions and do not print any redundant words.
 3 This description needs to include three parts. The title of each part should be '%Part1: Overall description%', '%Part2:
 4 List of objects%', and '%Part3: Relationships%'. All important nouns in your response have to be bounded by '<' and
 5 '>'.
 6 The first part is an overall description of the image. Your answer to this part should consist of three parts, one sentence to
 7 describe the style of the image, one sentence to describe the theme of the image, and several sentences to describe the
 8 image. The titles of these parts are '&&Part1.1: Style&&', '&&Part1.2: Theme&&', '&&Part1.3: Global description of
 9 background&&', 'Part1.4: Global description of foreground&&'. The global description should be as detailed as possible
 10 and at least 150 words in total. If there is text content in the image, you can also describe the text, which should be bound
 11 by quotation marks. All important nouns in your response have to be bounded by '<' and '>'.
 12 The second part is to list all the objects in the image, as many as possible, in order of importance. Note that any object
 13 should not be a part of other objects. Note that the listed object should not be the plural. If there are multiple individuals
 14 of the same category of objects, please list them separately. For example, if there are three apples in the picture, they
 15 should be listed as 'Apple 1,' 'Apple 2,' and 'Apple 3.', respectively. Additionally, the objects should be classified into two
 16 categories: living and inanimate objects. Living refers to creatures such as humans, cats, dogs, and plants, while other
 17 lifeless objects belong to the category of inanimate objects. Finally, each object should have a very detailed description,
 18 with more important objects receiving more detailed descriptions. Each description should be at least 30 words and the
 19 important nouns in it have to be bounded by '<' and '>'. You should also identify whether this object belongs to the
 20 foreground or background. You should additionally provide a sentence to describe the color information of the object.
 21 Therefore, the format for listing each object should be 'Object Name (Category (Living/Inanimate);
 22 foreground/background; Description; Color information)'. Specifically, the detailed description of an object should focus
 23 on its part and its action. All descriptions should be in the forms of, object's + part + verb + object/adjective or object + is +
 24 present participle. The description should be detailed as well as possible, and try to describe all parts of this object. You
 25 should specifically notice if there is a sky, tree, sun, or other object in the background of the environment. All important
 26 nouns in your response have to be bounded by '<' and '>'.
 27 The third part is to describe the relationships between all the objects in pairs. Please list them one by one. Additionally,
 28 please describe the relationship between object A and object B in the format of 'Object A' + 'Action' + 'Object B.' Please
 29 don't print the same relation twice. For example, if there is "A relation B", you shouldn't print 'B relation A' again. All
 30 important nouns in your response have to be bounded by '<' and '>'.
 31 I will provide you with an example of the last two parts of a description to show you the desired format. You should only
 32 focus on the format of this example instead of the content of it. You should use the same format to respond.
 33 "%Part2: List of objects%"
 34 <Woman> (Living; foreground; The <woman>'s <hair> is bundled in a <scarf>. Her <torso> is covered with a <black shirt>.
 35 Her <lower body> is clad in <blue jeans>. Her <legs> move through the <water>. Her <right hand> holds a pair of <shoes>;
 36 Color information: <black> shirt, <blue> jeans, <orange> scarf.)
 37 <Water> (Inanimate; foreground/background; The <water> floods the <street>, reflecting the <sky> and <surrounding
 38 objects>; Color information: <murky blue-grey>.)
 39 <Building 1> (Inanimate; background; The <building> has a <façade> with <doors> and <windows>, showing signs of
 40 <water damage>; Color information: <pale yellow>.)
 41 <Building 2> (Inanimate; background; This <building> is similar to <Building 1> but with a <red> roof visible above the
 42 <flood>; Color information: <light orange> walls, <red> roof.)
 43 <Vehicle 1> (Inanimate; background; A <vehicle> is partially submerged, showing only the <roof> and <upper parts>; Color
 44 information: <white>.)
 45 <Vehicle 2> (Inanimate; background; Another <vehicle>, also partially submerged, with a <visible logo>; Color information:
 46 <silver>.)
 47 <Sky> (Inanimate; background; The <sky> is filled with <clouds>, implying recent or ongoing <precipitation>; Color
 48 information: <gray>.)
 49 "%Part3: Relationships%"
 50 <Woman> [is walking through] <Water>.
 51 <Woman> [is moving away from] <Camera>.
 52 <Water> [reflects] <Sky>.
 53 <Water> [surrounds] <Vehicles>.
 54 <Buildings> [line] <Street>.
 55 <Vehicle 1> [is submerged by] <Water>.
 56 <Vehicle 2> [is submerged by] <Water>.

Figure 16: **The instruction** for GPT-4V to obtain BACON from an image. We highlight the parts involving specific examples in orange.



Overall description

Style: The image is a candid, real-life photograph with a humorous setup.
Theme: The theme of the image is an unexpected interaction between animals and human food.
Background description: The background of the image features a highway setting with a large, dark-colored **rock** on the left side and a **scrapped car** on the right. The sky is overcast, with a dull, uniform gray tone suggesting an overcast or cloudy day. The **highway** itself is made of asphalt and has visible lane markings. The **rock** and **car** appear to be out of place on the **road**, indicating a possible scene of an accident or roadside debris. There are no visible people or moving vehicles, which adds to the stillness of the scene.
Foreground: In the foreground, a **cat** and a **dog** are positioned on the road, closely inspecting a **pizza** that lies between them. The **cat**, on the right, is leaning forward with its head down, sniffing or showing interest in the **pizza**. The **dog**, on the left, mirrors the **cat's** posture, with its nose also directed towards the **pizza**. The **pizza** itself is large, with visible toppings that include cheese and pepperoni. The animals' proximity to the **pizza** and their focused attention on it create a humorous and unexpected scene, as it is unusual to see wild animals interacting in such a human-like manner. The **road** surface is a faded gray, and the animals' fur appears well-defined against this backdrop.

Object list

- ① **Name:** Cat; **Category:** Animal, foreground; **Description:** The cat's body is arched slightly as it leans over the pizza. Its head is lowered towards the pizza, its eyes seem to be focused on the food, and the cat's fur is dark blue; **Color:** dark blue; **Position:** [677, 178, 981, 361]
- ② **Name:** Dog; **Category:** Animal, foreground; **Description:** The dog is standing with its body angled towards the pizza. Its head is also lowered to get a closer look at the pizza. The dog has fur that appears to be short and smooth; **Color:** dark gray; **Position:** [45, 167, 375, 423]
- ④ **Name:** Pizza; **Category:** Inanimate, foreground; **Description:** The pizza is round with a crust and several toppings that include cheese and pepperoni. It is placed directly on the road surface; **Color:** golden crust, red pepperoni, yellow cheese; **Position:** [428, 183, 632, 324]
- ⑤ **Name:** Rock; **Category:** Inanimate, background; **Description:** The rock is large and irregularly shaped, with a rough texture and a dark color that contrasts with the asphalt of the highway; **Color:** dark gray; **Position:** [281, 537, 488, 649]
- ⑥ **Name:** Scrapped Car; **Category:** Inanimate, background; **Description:** The scrapped car is heavily damaged, with twisted metal. It is positioned on its side, suggesting it was involved in a severe accident; **Color:** rusty and metallic tones; **Position:** [584, 478, 852, 655]

Relationship

- | | |
|--|--|
| ① < Dog > on the left side of < pizza > | ② < Cat > on the right side of < pizza > |
| ③ < Pizza > is lying on < highway > | ④ < Rock > is blocking < highway > |
| ⑤ < Rock > is adjacent to < Scrapped Car > | ⑥ < Scrapped Car > is blocking < highway > |

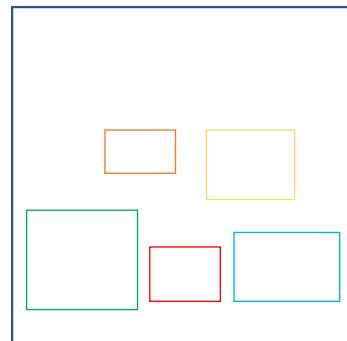


Figure 17: An example of how BACON-Captioner transforms a brief prompt into the distinctive style of BACON.

Under the bright moon, there is an icy sky and snow. In the center of the scene, a bonfire blazes. To the left of the bonfire are an arctic fox and a polar bear, to the right is a penguin, and above the bonfire flutters a small ice butterfly.

BACON-Captioner

Overall description

Style: The image is a color photograph with a clear focus and a balanced composition.

Theme: The theme of the image is a serene, wintery scene featuring various arctic animals gathered around a warm bonfire.

Background description: The background of the image is mainly the bright moon in the upper part of the screen. Below the sky, the horizon is lined with a layer of snow that blankets the ground, reflecting the light from the bonfire and creating a contrast with the dark sky. The snow appears to be undisturbed except for the area immediately around the fire, where it is melted.

Foreground: In the foreground, the central focus is the **bonfire**, with vibrant flames leaping upwards, casting a warm glow on the surrounding snow. The fire's light illuminates the figures of the animals and the snow around them. To the left of the fire, an **arctic fox** sits attentively, its fur appearing thick and white. To the right, a **polar bear** sits with a relaxed posture, its white fur contrasting with the darker elements of the scene. In front of the fire, a **penguin** stands out with its black and white coloring, with its back facing the camera. Above the fire, an **ice butterfly** with outstretched wings adds a whimsical touch to the scene. The animals are positioned on a flat surface, likely a frozen body of water, as indicated by the presence of the ice butterfly and the reflection of the fire on the snow.

Object list

- ① **Name:** Arctic Fox; **Category:** Animal, foreground; **Description:** The arctic fox's body is facing the camera, with its head turned slightly to the side. Its fur is thick and white; **Color:** White; **Position:** [153, 235, 337, 419]
- ② **Name:** Polar Bear; **Category:** Animal, foreground; **Description:** The polar bear is sitting with its body oriented away from the camera. Its fur is thick and white, and its head is turned to the side, giving a profile view; **Color:** White; **Position:** [266, 264, 481, 627]
- ③ **Name:** Penguin; **Category:** Animal, foreground; **Description:** The penguin stands with its body turned away from the camera, its head slightly tilted. It has black back and white belly; **Color:** Black back, white belly; **Position:** [471, 154, 614, 416]
- ④ **Name:** Bonfire; **Category:** Inanimate, foreground; **Description:** The bonfire consists of flames leaping upwards from a pile of wood. The fire is bright and orange, casting light on the surrounding snow; **Color:** Orange flames, brown and black wood; **Position:** [307, 204, 614, 409]
- ⑤ **Name:** Ice Butterfly; **Category:** Animal, foreground; **Description:** The ice butterfly is positioned above the bonfire, with its wings spread wide. It appears to be made of ice and is translucent with a white hue; **Color:** Translucent with a white tint; **Position:** [378, 768, 450, 896]

Relationship

- | | |
|--|---|
| ① < Arctic Fox > on the left side of < Bonfire > | ② < Arctic Fox > sits near < Polar Bear > |
| ③ < Polar Bear > on the right side of < Arctic Fox > | ④ < Penguin > in front of < Bonfire > |
| ⑤ < Ice Butterfly > floats above < Bonfire > | ⑥ < Polar Bear > sits beside < Bonfire > |

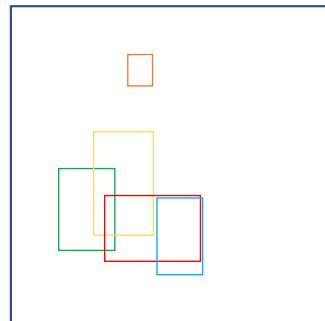
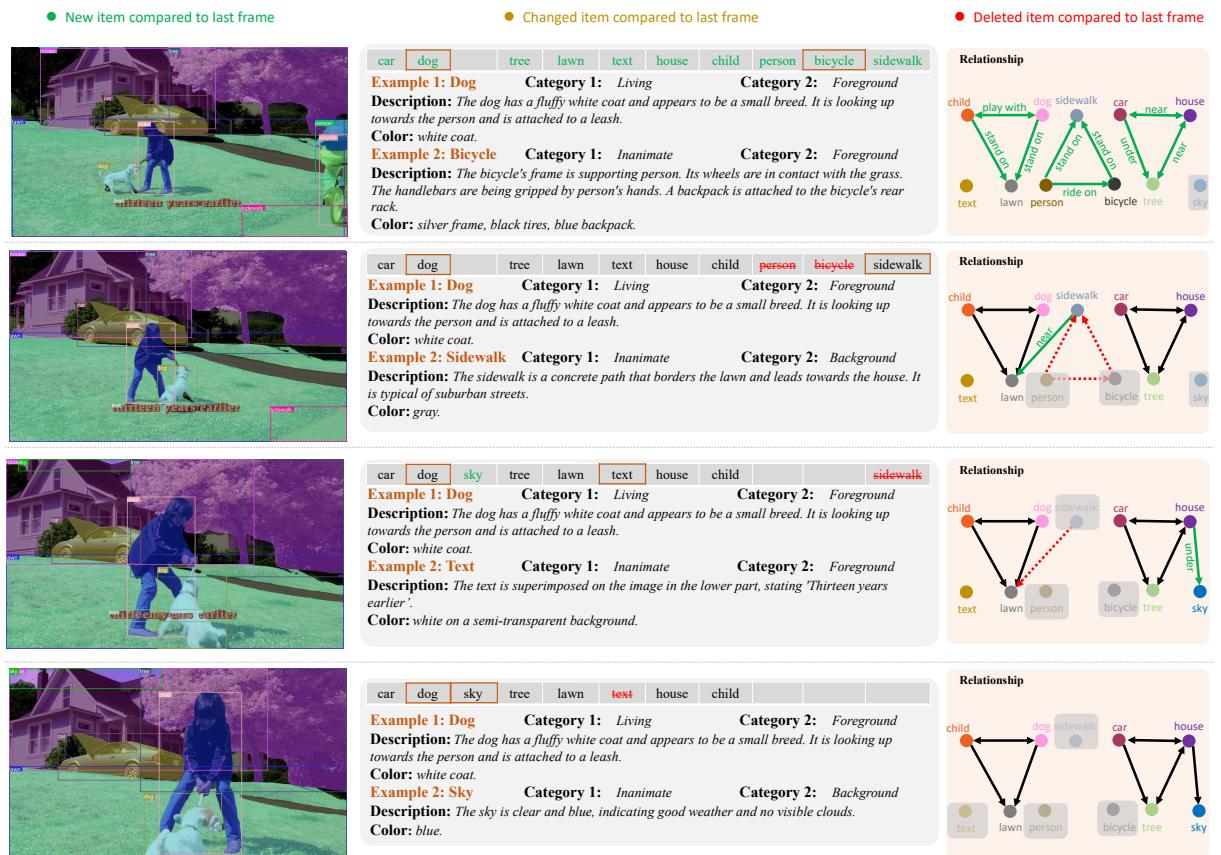


Figure 18: An example of how BACON-Captioner transforms a relatively complex prompt into the distinctive style of BACON.

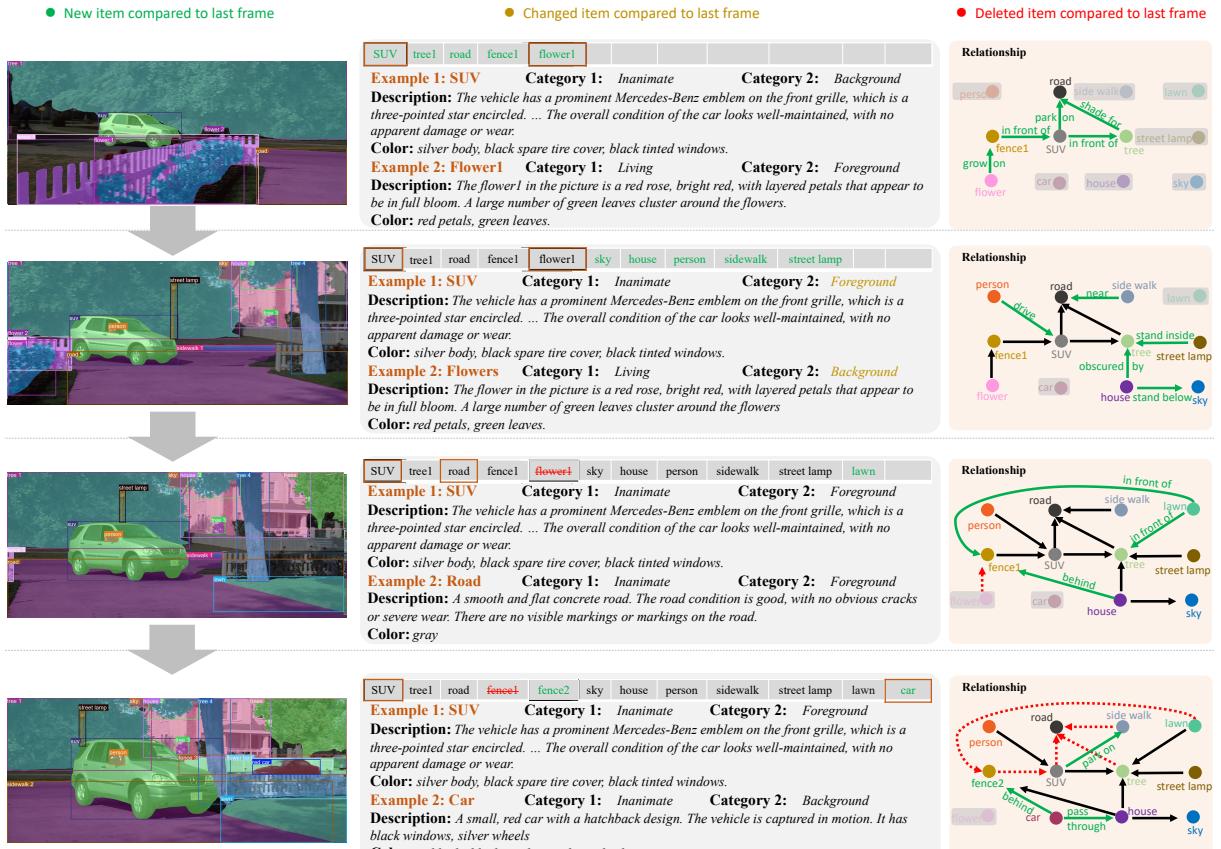
Caption	BACON + SDXL	DALL-E 3	SDXL
In the deep sea, an abandoned large ship full of marine life sank to the bottom of the sea. There are two blue balloons floating in front of the ship. There is a dolphin swimming below the balloon. There is a drifting bottle floating in the deep sea, inside which is a sailboat			
In a yoga studio, there is an artwork of a green jade dragon, with a white cat lying on the right side of the artwork. On the distant ground, against the wall, there is a painting depicting war			
In an abandoned factory building, sunlight filtered in. A technologically advanced spaceship flies over the factory building. Listening to a motorcycle below the spaceship, there is a pink guitar on the ground to the right of the motorcycle.			
On a pink night, there was a pool in the center of the lawn, and a purple sports car was floating on the pool. There was a light bulb on the hood of the sports car, and there was an orange goldfish in the bulb. On the left side of the car is a small, colorful robot			
There is a small river in the forest, and there is a stone bridge on the river. There is a golden praying mantis on the bridge. There is a mongoose standing by the riverbank, and to its right lies a turtle			
In an old-fashioned subway station, there is a emerald green lion, a gray white wolf, and a colorful paper crane standing together waiting for the subway			

Figure 19: Additional examples of BACON on image generation.



Note: Considering the limited drawing space and beauty of graph structure, we have performed some simplification for relationships that appear more than once.

Figure 20: An additional example of BACON on video captioning.



Note: Considering the limited drawing space, some similar objects with different names are omitted here. For relationships that appear more than once, we have performed some simplification.

Figure 21: An additional example of BACON on video captioning.