
Graph-Based Captioning: Enhancing Visual Descriptions by Interconnecting Region Captions

Yu-Guan Hsieh^{1*} Cheng-Yu Hsieh^{2*†} Shih-Ying Yeh³
 yghsieh@apple.com cydhsieh@cs.washington.edu kblueleaf@gapp.nthu.edu.tw

Louis Béthune¹ Hadi Pour Ansari¹ Pavan Kumar Anasosalu Vasu¹
 {l_bethune, mpouransari, panasosaluvasu}@apple.com

Chun-Liang Li¹ Ranjay Krishna^{2†} Oncel Tuzel¹ Marco Cuturi¹
 ranjay@cs.washington.edu, {chunliang_li, otuzel, cuturi}@apple.com

¹ Apple ² University of Washington ³ National Tsing Hua University

Abstract

Humans describe complex scenes with compositionality, using simple text descriptions enriched with links and relationships. While vision-language research has aimed to develop models with compositional understanding capabilities, this is not reflected yet in existing datasets which, for the most part, still use plain text to describe images. In this work, we propose a new annotation strategy, graph-based captioning (GBC) that describes an image using a labelled graph structure, with nodes of various types. The nodes in GBC are created using, in a first stage, object detection and dense captioning tools nested recursively to uncover and describe entity nodes, further linked together in a second stage by highlighting, using new types of nodes, *compositions* and *relations* among entities. Since *all* GBC nodes hold plain text descriptions, GBC retains the flexibility found in natural language, but can also encode hierarchical information in its edges. We demonstrate that GBC can be produced automatically, using off-the-shelf multimodal LLMs and open-vocabulary detection models, by building a new dataset, GBC10M, gathering GBC annotations for about 10M images of the CC12M dataset. We use GBC10M to showcase the wealth of node captions uncovered by GBC, as measured with CLIP training. We show that using GBC nodes’ annotations—notably those stored in composition and relation nodes—results in significant performance boost on downstream models when compared to other dataset formats. To further explore the opportunities provided by GBC, we also propose a new attention mechanism that can leverage the entire GBC graph, with encouraging experimental results that show the extra benefits of incorporating the graph structure. Our datasets are released at <https://huggingface.co/graph-based-captions>.

1 Introduction

The availability of huge paired image/caption datasets has revolutionized our ability to produce joint vision-language embedding, paving the way for tasks like efficient caption-guided image generation within powerful multimodal foundation models [2, 36, 38, 50]. The quality and granularity of these datasets plays, therefore, a crucial role. While quality can be addressed by filtering out data [17, 22, 53] or, inversely, by improving caption quality through *recaptioning* [14, 16, 35, 45],

*Equal contribution

†Work done at Apple

there is ample interest in the community to provide more detailed, fine-grained information for each image [3, 7, 15, 42]. To obtain better annotations, we draw inspiration from compositionality, a fundamental characteristic of human perception that is reflected in the natural language used to describe our surroundings [4, 10, 12, 20, 29, 30]. Compositionality plays an especially important role when examining larger images found in the wild, which have a rich coarse-to-fine, hierarchical structure, commonly represented as a scene graph [31]. While scene graphs have been successfully applied to image retrieval [31, 54], generation [18, 44], and pre-training [26, 28], the scale of scene graph dataset is typically small. For instance, Visual Genome [33] only contains around 100k images.

Contributions. To overcome the limitations of existing datasets and annotation formats that either struggle to represent the hierarchical nature of scenes or are of small size and lack flexibility in their description, this paper makes a series of contributions as summarized below.

1. **We propose graph-based captioning (GBC)**, a new vision-language data format that captions images with a graph-based structure akin to scene graphs while retaining the flexibility and intuitiveness of plain text description. GBC contains four types of nodes: (1) an image node with captions of the entire image, (2) entity nodes that contain descriptions of individual objects, (3) composition nodes that link objects in the images of the same type, and (4) relation nodes that describe the spatial (“the tree is to the left of the tower”) or semantic (“The branch is covered in snow”) relationships between objects of different types (§ 3.1).
2. **We design a workflow to produce GBC annotations at scale.** Inspired by recent recaptioning approaches [14, 16, 35, 45], we generate GBC annotations using the OSS LLaVA-1.6 [42, 43]. First, LLaVA generates short and long captions for the entire image, used to extract entities. Then, an object detection model (YOLO-World [9]) is employed to find bounding boxes for each entity. Subsequently, the same procedure is recursively invoked to produce a GBC for each proposal. Finally, LLaVA-1.6 is prompted to produce composition and relation captions that connect multiple entity nodes (§ 3.2).
3. **We create large-scale GBC dataset containing 10 million images with ≈ 534 words per image using that workflow.** While ours is the first vision-language dataset that contains structured captions, a few recent datasets contain dense annotations, and only [65] has a scale that is similar to ours. Drawing on graph-encoders, we also design a baseline architecture to utilize this structure, using a new attention mechanism, structure-aware hierarchical attention (SAHA) (§ 4).
4. **We demonstrate experimentally in § 5 that the diversity of captions found in GBC nodes improves CLIP model performance** across image-to-text retrieval, text-to-image retrieval, compositionality, and semantic segmentation tasks, while retaining comparative performance on zero-shot ImageNet classification. Our ablations highlight that it is not the density or scale of our dataset, but the structured nature itself, that leads to better performance. Remarkably, we observe that composition and relation nodes, which can only be obtained through the GBC workflow, boost performance. Finally, we perform ablation on the influence of annotation format on retrieval performance using a set-aside test set from GBC. In this case, we see that SAHA, when fed GBC annotations, provides comparable or even better performance than that obtained when describing an image with detailed captions, suggesting that GBC can be a promising alternative to traditional image captioning formats.

2 Related works

In this section, we discuss related works on vision-language datasets. We refer the readers to Appendix A for works that are specific to CLIP [50] training.

Vision-language datasets. First vision-language datasets were manually built using human annotations, such as Flickr30k [69], COCO [41] and Visual Genome [33]. This yielded annotations of high quality, but unfortunately of short length, and in limited amounts (with no dataset containing more than 130k images). Several studies have then demonstrated the benefits of using larger scale datasets obtained by crawling the web, such as YFCC100M [59], RedCaps [13], or Wikipedia-based image-text dataset (WIT) [57]. The quality of these data became a concern when it was noticed that in some situations the caption was only loosely related (or not related at all) with the image, which can be detrimental to the overall performance [52]. This motivated researchers to use automatic filtering procedures to select higher-quality data samples, like in Localized Narratives [48] or Conceptual Captions (CC3M) [56], and its successor CC12M [6]. These efforts have reached billion scale with LAION-5B [53], and LAION-CAT [49]. In a similar vein, Meta-CLIP [68] reproduces the processing of the seminal CLIP paper [50] on a subset of the Common Crawl dataset, SemDeDup [1] relies on

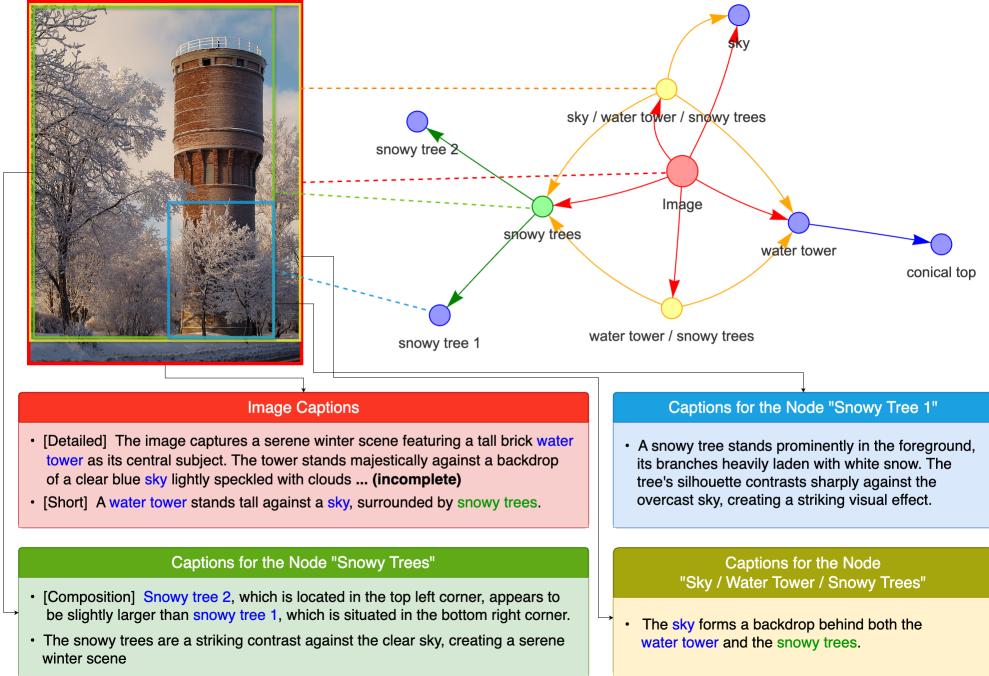


Figure 1: An illustration of our proposed graph-based captions. The image node, entity nodes, composition nodes, and relation nodes are respectively colored in red, blue, green, and yellow. The color texts in the captions correspond to the labels of the outgoing edges. More examples are provided in Appendix C.3.

the embeddings provided by foundation models to filter data and remove duplicates, while DFN [17] uses filtering networks trained on high quality data to extract subsets of Common Crawl.

VL datasets with dense captioning. It was noticed recently that using entirely generated captions from raw images, as in DAC [14] and AS-1B [65], could improve results over filtering approaches. These datasets are characterized by their long and detailed captions that describe every element within a scene. Complementing these efforts, Urbanek et al. [61] introduced DCI, a dataset featuring similarly dense annotations but curated by humans and on a smaller scale. Alternatively, DOCCI [46] focuses on a set of only 15k high quality, high resolution, paired image-captions, manually selected and annotated by one of the authors, with typical caption length of more than 135 words. In the ImageInWords [21] dataset, captions are iteratively improved by humans, on top of previously human or machine annotated captions, yielding 9K densely captioned images.

3 Improving image annotations with graph-based captioning

We introduce in this section our new captioning format to represent an image, explain how we can use any off-the-shelf multimodal large language model (MLLM) and open-vocabulary detection model to obtain such captions, and briefly describe the two datasets GBC1M, and GBC10M that we construct following the proposed workflow. Additional details about the data preparation process and the datasets can be found in the Appendices B and C.

3.1 Representing an image with graph-based captions

To encode the structured information contained in an image, we propose to represent each image as a directed acyclic graph (DAG), denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Each node of the graph $v \in \mathcal{V}$ is associated with a bounding box. Starting with the root node, which corresponds to the entire image (image node), other nodes can either hold a set of objects (composition node and relation node), or a single object in the image (entity node). Moreover, to benefit from the expressive power of natural language descriptions and to ensure smooth integration of our annotations into the existing ecosystems of methods that rely primarily on image-text pairs, we label each node v with a set of captions $\mathcal{C}^v = \{C_1, \dots, C_{n^v}\}$.

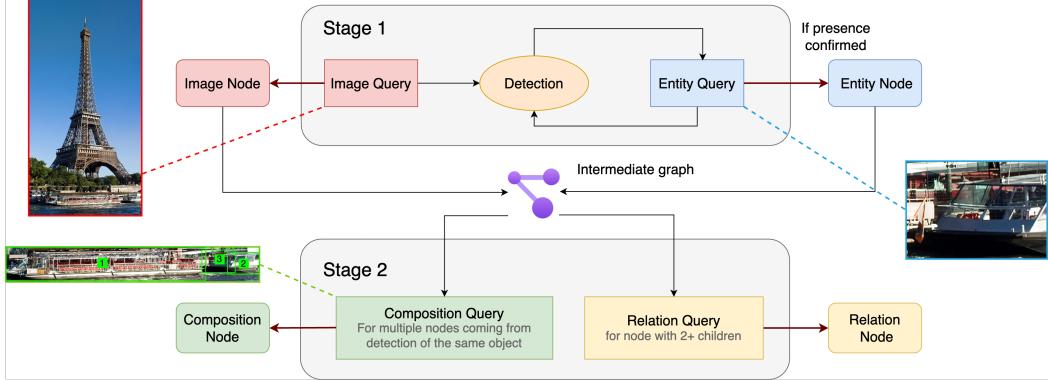


Figure 2: Our image annotation process involves four types of queries that are performed in two separate stages, with the detection model serves to single out the regions that are used for different queries.

The edges, on the other hand, are used to encode the *hierarchy* between the nodes. More specifically, there is an edge $e \in \mathcal{E}$ from u to v only if the content associated to v is part of the content associated to u . This relation is also reflected by the edge label L^e which should appear in the captions of the source node u and be able to represent the object(s) associated to the target node.³

An exemplar GBC, generated automatically through our workflow is provided in Figure 1. It should be noted that the only manual addition in that graph comes from the "title label" of each node, which is obtained by taking the union of labels found in its incoming edges. Compared to the standard scene graph annotation, the use of node captions provides flexibility to describe complex concepts, while the underlying graph still captures the inherent structure of the image. Our dataset, whose construction is detailed in Section 3.2 next, includes several different types of captions tailored to the structure of the DAG. At the root image node, we provide both detailed and short captions to cater to varying levels of granularity. Captions at composition nodes and relation nodes explicitly describe the arrangement and interaction of multiple objects, while the captions at the entity nodes provide detailed description of a single object.

3.2 GBC dataset construction workflow

We show how to produce GBC annotations automatically, using any pre-trained MLLM and open-vocabulary detection model. This results in a workflow that is comparable, in compute time and complexity, to that of other widespread recaptioning approaches. At a high level, we use a MLLM model to provide captions and identify potential entity nodes, followed by a detection model to provide bounding box coordinates for these entities.

Data annotation. Our overall process to annotate a single image is shown in Figure 2. To account for the different types of nodes, we design four query templates as listed below:

- **Image query:** We ask the model to provide detailed caption for the image, identify prominent elements, and summarize the long caption with a concise one that contains all these elements. The identified elements are then passed to the detection model to obtain the bounding boxes.
- **Entity query:** For each bounding box, we crop out the region and ask the model whether a specific object appears in the cropped image. Moreover, we also ask the model to describe the object and identify prominent elements of the object when it is present. The identified elements are again passed to detection models for detection.
- **Composition query:** In the case where multiple bounding boxes are returned for a single type of object, we ask the model to describe the composition of these objects with an annotated image.
- **Relation query:** For image or entity nodes with more than two children, we ask the model to describe the relations between its children.

Provided that there is no guarantee that all the detected objects would end up as a node in the graph—consider the case where the MLLM says that the object is not present or just fails to reply in

³Ideally, we would also like to distinguish between multiple appearances of the same text in a caption. However, this is not explicitly handled by our current dataset construction workflow so we omit it here.

	GBC1M	GBC10M
# Images	1,013,592	10,138,757
# Vertices / Image	12.12	12.24
# Edges / Image	22.28	21.81
# Captions / Image	17.40	17.67
# Words / Image	593.14	533.98
Average Graph Diameter	4.55	4.41

Table 1: Key statistics of the GBC1M and GBC10M datasets. We report number of images, average number of vertices, edges, captions, and words per image, and average graph diameter.

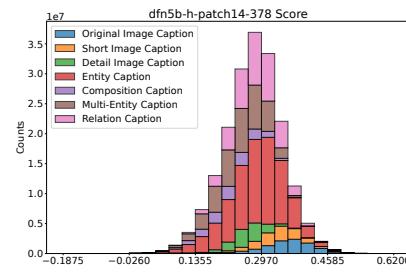


Figure 3: CLIP score distribution of different types of captions in the GBC10M dataset. Multi-entity caption correspond to the non-composition captions at composition nodes.

the correct format—we split the entire process into two stages, and we only perform composition queries and relation queries *after* discovering all the entity nodes. Finally, to improve efficiency and to reduce redundant information, we train two dedicated classifier on top of Jina Embeddings [23] to decide whether a piece of text is suitable for object detection and whether two texts can represent the same object in an image. The former is applied to every identified element while the later results in merging of nodes when a new query targets a region that has already been queried with similar texts.

3.3 GBC1M and GBC10M

Following the process outlined in [Section 3.2](#), we annotate the CC12M dataset [6] with graph-based captions using LLaVA-1.6 [42, 43] as the MLLM and Yolo-World [9] as the open-vocabulary detection model. Specifically, we construct two sets of annotations: GBC1M for a subset of around 1M of images, with all the queries performed with the Yi-34B version of LLaVA-1.6, and GBC10M for a subset of around 10M of images, with LLaVA-1.6 Yi-34B for image and composition queries, and LLaVA-1.6 Mistral-7B for entity and relation queries.⁴

We provide statistics of the above two datasets in [Table 1](#). We note that these two datasets have very similar per-image statistics, with the number of words being the only exception, as LLaVA-1.6 Yi-34B tends to provide longer descriptions than LLaVA-1.6 Mistral-7B. Moreover, our datasets use an average number of around 500 words to describe each image. This is comparable to other dataset with rich annotations such as DCI (1111 words/img) [61] and DOCCI (136 words/img) [46]. We also compute the CLIP scores between the captions and their corresponding regions using the DFN-5B CLIP model [17], and we report their distribution for the GBC10M dataset in [Figure 3](#). We note that the original CC12M caption achieves the highest CLIP scores, followed by the short synthetic caption for the entire image. This can be explained by the fact that in these two cases, the involved image-caption pairs more closely align with the training data of standard CLIP models.

4 Encoding GBC via structure-aware hierarchical attention

Alongside many ways to leverage GBC annotations, as we shall present in [Section 5](#), we propose a simple text encoder architecture to incorporate structural information encoded in GBC graph along with node captions. Specifically, we present structure-aware hierarchical attention (SAHA) block which treats each caption as an individual sample, and introduces an additional cross-attention layer that enforces the captions to attend to their children.

Formally, we consider a caption graph $\mathcal{G}^C = (\mathcal{C}, \mathcal{E}^C)$ with vertices $\mathcal{C} = \bigcup_{v \in \mathcal{V}} \mathcal{C}^v$ and edges $\mathcal{E}^C \subseteq \mathcal{C} \times \mathcal{C}$ such that $(C, C') \in \mathcal{E}^C$ if and only if $C \in \mathcal{C}^u$, $C' \in \mathcal{C}^v$, $e = (u, v) \in \mathcal{E}$, and the label L^e is included within the caption C . In words, each vertex in the graph represents a caption from a node of the original graph and there is an edge from one caption to another only if the second caption describes part of the first caption. After tokenization of the captions, we can map the edge labels to a

⁴Our larger dataset does not cover the entire CC12M both because some images were no longer accessible at the time we accessed the images, and because we discard images for which the MLLM model’s reply to the image query does not comply with the prescribed format.

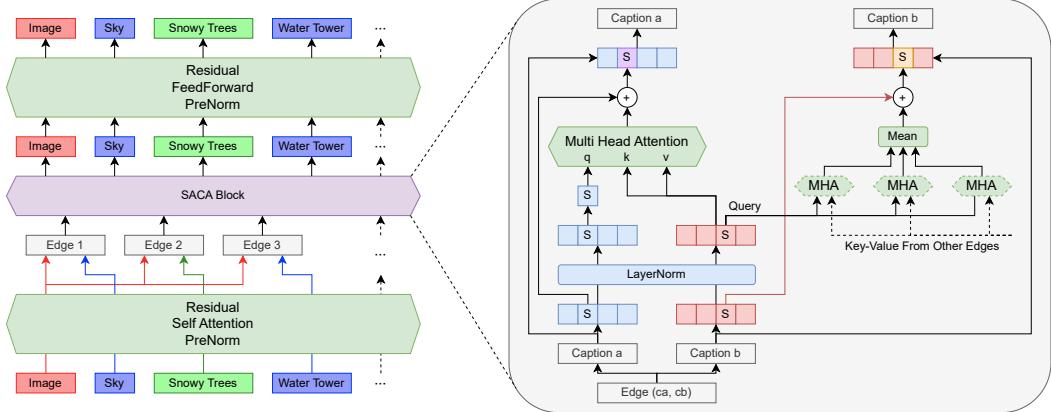


Figure 4: Illustration of the proposed SAHA block when applied to the graph shown in Figure 1. For the sake of simplicity, we assume here there is only one caption per node.

set of token positions of the source caption, which we write as \mathcal{P}^e . Then, the target caption *annotates* the source caption via the tokens at positions \mathcal{P}^e . Therefore, we can simply consider cross-attention with queries from these tokens and keys and values from the target caption. We illustrate this idea in Figure 4, where we zoom in on the additional cross-attention layer (SACA) on the right side of the figure. For completeness, we also provide the corresponding mathematical formula in Appendix D.

Our text encoder then stacks a number of SAHA blocks, effectively interleaving the vanilla self-attention layers that process, *local, intra-caption* information, with the structure-aware cross-attention (SACA) layers that process *global, inter-caption* information in a structure-aware manner. Furthermore, the model acts as a classic text encoder in the absence of edges in the graph, i.e., when $\mathcal{E}^C = \emptyset$, dropping all SACA blocks.

Complexity analysis. To estimate the computational complexity of our approach, we assume that the captions have a fixed sequence length n . Then, implementing SACA using masked cross-attentions between captions leads to a total complexity of $\mathcal{O}(|\mathcal{E}^C|n^2)$. Additionally, we must account for the self-attention operations, resulting in a combined complexity of $\mathcal{O}(|\mathcal{C}|n^2 + |\mathcal{E}^C|n^2)$. Provided that many of the graphs in our dataset are trees, we have $|\mathcal{E}^C| = |\mathcal{C}| - 1$ and the complexity simplifies to $\mathcal{O}(|\mathcal{C}|n^2)$. In contrast, a naive approach that performs self-attention on the concatenated set of all captions would result in a significantly higher complexity of $\mathcal{O}(|\mathcal{C}|^2n^2)$.

5 Experiments

We present in this section a comprehensive set of experiments to benchmark different image annotation schemes. Specifically, using CLIP model training as the main task, we show that GBC annotations can bring improvements on a range of benchmarks across classification, retrieval, and dense prediction tasks, compared to existing annotation schemes (Section 5.3). On retrieval tasks, we demonstrate how GBC allows one to encode denser, more descriptive textual information to better represent images as shown by the performance gain compared to existing annotation formats (Section 5.4). Missing experimental details are provided in Appendix E.

5.1 Annotation formats

We outline below the different types of image annotations that are considered in our experiments, each providing different opportunities to leverage information from the image.

Short caption. Each image is paired with a short caption, as in common image-text datasets.

Long caption. One can improve image description using a longer caption. We use long captions in our dataset. They are of 110 words on average, as compared to short captions, of only 28 words on average. We extend the context length of text encoders in CLIP models from 77 to 512 for this setup.

Annotation	Flickr-1k		MSCOCO-5k		ImageNet	SugarCrepe	ADE20K
	T2I	I2T	T2I	I2T			
CC12M	46.4	64.6	25.0	39.4	39.2	72.9	41.7
Short	56.3	73.2	30.7	46.7	38.8	76.0	42.0
Long	56.4	75.2	31.8	<u>50.1</u>	<u>39.6</u>	77.0	42.8
Region	<u>58.3</u>	<u>76.6</u>	31.5	49.1	38.5	75.6	43.5
GBC-captions	60.6	79.3	34.1	51.9	40.8	<u>76.7</u>	45.0
GBC-concat	56.1	76.0	31.4	48.5	39.0	75.7	42.1
GBC-graph	58.1	76.2	<u>32.2</u>	49.5	38.2	75.2	<u>43.8</u>

Table 2: Comparative performance on various existing benchmarks when trained using different annotation schemes. For retrieval tasks we report Recall@1, and for ADE20K we report the mIOU. As a baseline, we also report performance of a model trained on the same set of images using original CC12M captions. The highest scores for each task are highlighted in bold, and the second-highest scores are underlined.

Region captions. Alternatively, more captions can be provided for an image, especially those that describe a specific region of the image. While this format includes all region captions, it does not include the relational information between region descriptions found in GBC.

Graph-based captions. Finally, we consider the GBC format as proposed in Section 3.1. The GBC format includes region captions, but also provides additional information, stored in relation and composition nodes. With this in mind, we explore three different ways to leverage GBC annotations:

- A direct way to leverage GBC is to treat captions for all nodes in the graph as positive texts for the image, i.e. as what we do for region captions. We refer to such method as **GBC-captions**.
- Another strategy is to traverse from the root image node through the graph and concatenate the captions at the visited vertices into a single long caption. We then train a CLIP model with 512 context length in the standard fashion. We refer to this method as **GBC-concat**.
- To fully benefit from the graph information, when available, we leverage the SAHA block proposed in Section 4. This allows us to encode the entire graph into text embeddings that also contain information from their respective subgraph. We refer to this method as **GBC-graph**.

As GBC annotation encapsulates all existing *short*, *long*, and *region* caption formats, we are able to instantiate all the setups by using only a subset of annotation available in our curated GMC10M dataset. Specifically, taking only the short or detailed caption at the root image node creates the *short* and *long* caption setup, respectively. To mimic the *region* caption setup, we drop the relation and composition captions from GBC annotations. By turning GBC into these settings, we ensure the same quality of text annotations across different methods.

5.2 Experimental setup

We perform CLIP training on our GBC10M dataset, while leaving out 10,151 samples as the test set. Following common practice, we use the CLIP score computed by a pre-trained CLIP model [17] to filter our training set, discarding the 5% of captions with the lowest scores for each type. In addition, we retain the original CC12M captions associated with each image. Specifically, in all setups, both the original caption and the short synthetic caption are consistently used as positive texts for the image during training. This prevents the severe distribution shifts that could occur from using only long or region captions when evaluating on standard benchmarks.

Objective. To pair an image with multiple captions in training CLIP models, we adopt a multiple-positive contrastive loss in the spirit of LaCLIP [16] and DAC [14]. Briefly speaking, compared to standard CLIP objective, the multiple-positive loss sums over the loss on each positive captions of an image while all the captions from the images in the same batch are used in the normalization term.

Model and hyperparameters. We use the standard CLIP ViT-B/16 model, with the only difference of longer context length of text encoder for long caption and GBC-concat, and a replacement of the vanilla transformer block by the SAHA block in text encoder for GBC-graph. We fix the global batch size (i.e., number of images in each batch) to 4,096 for all the methods. The models are trained for 45,000 steps with AdamW and cosine scheduler at a learning rate of 10^{-3} . This roughly correspond to 20 epochs of training. We evaluate at the EMA checkpoint at epoch 10, as we observe that further training provides little to no improvement in performance across the benchmarks.

Annotation	Short		Long		GBC-captions		GBC-concat		GBC-graph	
	T2I	I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I	I2T
Short	85.8	86.2	85.0	87.2	57.3	37.0	87.4	88.2	-	-
Long	86.4	87.5	95.4	95.7	44.3	33.1	90.5	91.4	-	-
Region	85.3	86.1	85.5	88.2	91.5	79.3	89.5	90.0	-	-
GBC-captions	86.8	87.6	87.2	89.6	91.3	80.9	90.1	91.0	-	-
GBC-concat	86.1	86.5	92.7	93.5	57.5	37.9	94.6	94.9	-	-
GBC-graph	84.3	85.2	84.9	87.6	90.8	79.3	89.1	89.7	95.7	96.1

Table 3: Image and text retrieval performance on GBC test set when trained and evaluated using different types of annotations (Rows: models trained from different annotations; Columns: Evaluations on different annotations). For the GBC-captions column, we include all the captions from our graph by default except in cases where training is done solely on the region annotations. In this case, excluding relation and composition captions, as done during training, results in better performance.

5.3 Evaluations on existing benchmarks

We compare the CLIP models derived from different annotation schemes on an array of evaluation benchmarks, including: **Flickr-1k** [47] and **MSCOCO-5k** [41] for zero-shot retrieval, **ImageNet** [51] for zero-shot classification, **SugarCrepe** [27] for compositional understanding evaluation, and **ADE20k** [76] for semantic segmentation that measures models’ dense prediction performances. Table 2 illustrates our results, from which we draw the following two key insights.

GBC annotation leads to clear performance gains by encoding relational information. Table 2 demonstrates that training models with more detailed textual information, such as long captions or region captions, consistently enhances downstream performance, particularly in retrieval tasks and dense prediction. However, the most significant improvements are seen with GBC-captions, which augment traditional region captions with relational and compositional descriptions. Given that the GBC workflow is uniquely positioned to provide these, this result demonstrates the soundness of GBC, capturing valuable insights not present in conventional captions.

How the captions are used matters. Compared to GBC-captions, the improvements achieved by GBC-concat and GBC-graph on these benchmarks are of a smaller margin. This indicates that the way GBC annotations is used significantly impacts performance. Specifically, this worse performance is likely due to a mismatch between training and evaluation. For instance, the graph information that would benefit GBC-graph most is not provided in any of these benchmarks. We address this discrepancy below.

5.4 Evaluation on GBC test set

To assess the effectiveness of different annotation formats, we provide the model with these annotations at *test time*. Annotations that better describe the images should, ideally, result in better retrieval performance when they are used. Specifically, we use our own test set and consider performing retrieval with the various types of annotations presented in Section 5.1. Note, however, that when using region captions or GBC-captions, no single text embedding can naturally encompass all the relevant information. To address this limitation, we perform retrieval based on the average CLIP score between the image embedding and the text embeddings of the provided captions in this setup. We report our results in Table 3, where the rows correspond to the annotations at training time and the columns correspond to the annotation at test time. Unsurprisingly, we see a strong tendency that when a model is trained with a certain annotation format, it performs the best when we use the same format for retrieval. Among the few exceptions, we note that models trained to pair with shorter captions may have better performance when concatenation of short captions is provided at test time. This leads us to the following two observations.

Denser textual information improves retrieval performances. The table clearly shows that training with richer annotations—such as long captions, GBC-concat, or GBC-graph—enhances retrieval performance. This improvement suggests that these methods provide a more effective representation of the images. Specifically, GBC-graph yields the best performance, indicating that the proposed GBC format consists in a viable alternative to the commonly used detailed captions.

Simple augmentation during training does not allow to exploit additional information when available. Our observations from Section 5.3 show that treating all captions as independent positives

Annotation	T2I	I2T
Short	71.5	78.1
Long	72.9	80.1
GBC-captions	86.4	87.3
GBC-concat	73.3	79.7
GBC-graph	85.1	85.7

Table 4: Image and text retrieval performance on GBC test when using max CLIP score over all the captions.

Annotation	DCI-Long		DCI-concat		ShareGPT4V-15k	
	T2I	I2T	T2I	I2T	T2I	I2T
Long	53.6	53.3	63.3	65.8	93.4	93.9
GBC-captions	42.5	43.3	64.3	63.8	78.7	82.1
GBC-concat	51.4	52.3	69.0	70.8	89.5	91.4

Table 5: Image and text retrieval performance on DCI and a 15k subset of ShareGPT4V-cap100k of our models trained on longer captions. We also include GBC-captions that by design can only handle short captions as a baseline for comparison.

yields the best performance on existing benchmarks. However, there is no evidence that this method could harness the richer information from multiple captions when they are provided together in test time. Indeed, whether we use average CLIP score or concatenation, the retrieval performance of these methods significantly lags behind those methods that are trained directly with captions that individually encompass rich information.

5.5 Ablation studies

In addition to our main experiments, we have conducted extensive ablation studies on both the training and evaluation of our models. We present two of them here and defer the remaining to [Appendix F](#).

Retrieval with multiple captions using maximum CLIP score. An alternative to the mean CLIP score we considered in [Table 3](#) is to take the maximum, for which we report the results in [Table 4](#). Compared to taking the average, using the maximum is more robust to low CLIP scores, and thus gives better results when the model is not trained to match the image with all local captions, as with Short, Long caption, and GBC-concat. Nonetheless, despite these differences, the overall retrieval performance still significantly lags behind that achieved using a single caption.

Retrieval with long captions. To complement the results presented in [Table 2](#), we evaluate the retrieval performance of our extended context models on datasets with dense annotations. We focus specifically on ShareGPT4V [8], which offers GPT-style detailed captions closely resembling those obtained from LLaVA, and DCI [61], containing human-annotated detailed and region captions. The latter allows us to perform retrieval using either detailed captions or concatenated short captions, as we did in Section 5.4. Our results shown in [Table 5](#) demonstrates that the close caption distribution with ShareGPT4V effectively enables strong retrieval results for models trained on our long captions. However, potentially due to the distribution shift, all the models perform badly on DCI retrieval with long captions. In this setup, using concatenated captions for training and retrieval significantly outperformed other baselines, indicating the broader benefit of the concatenation approach.

6 Conclusion

We propose graph-based captioning (GBC) as a new image-text annotation format, and curated GBC1M and GBC10M datasets. Grounding on CLIP model training, we propose various baseline methods to utilize the GBC datasets. Via our experiments, we demonstrate that training with GBC leads to improvements in various benchmarks compared to models derived from traditional annotation formats. This suggests that GBC provides richer textual information than existing annotation schemes and could hence be a valuable foundation for developing more advanced vision-language models across various applications.

Acknowledgements

The authors are thankful to Alaa El-Nouby, Enrico Fini, Rick Chang, Shuangfei Zhai, Jiatao Gu, Yizhe Zhang, Hanlin Goh, Josh Susskind, and Vaishaal Shankar for the valuable feedback that have been provided throughout the course of the project.

References

- [1] Amro Kamal Mohamed Abbas, Kushal Tirumala, Daniel Simig, Surya Ganguli, and Ari S Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023.
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.
- [3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions, 2023.
- [4] Léon Bottou. From machine learning to machine reasoning. *Machine learning*, 94(2):133–149, 2014.
- [5] Likun Cai, Zhi Zhang, Yi Zhu, Li Zhang, Mu Li, and Xiangyang Xue. Bigdetection: A large-scale benchmark for improved object detector pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4777–4787, 2022.
- [6] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021.
- [7] Junsong Chen, Jincheng YU, Chongjian GE, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-\$\alpha\$: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024.
- [8] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.
- [9] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2024.
- [10] Noam Chomsky and Morris Halle. Some controversial questions in phonological theory. *Journal of linguistics*, 1(2):97–138, 1965.
- [11] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. <https://github.com/open-mmlab/mmsegmentation>, 2020.
- [12] MJ Cresswell. *Logics and Languages*. Methuen, 1973.
- [13] Karan Desai, Gaurav Kaul, Zubin Trivadi Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.
- [14] Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-Bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogerio Feris, Shimon Ullman, and Leonid Karlinsky. Dense and aligned captions (DAC) promote compositional reasoning in VL models. In *Neural Information Processing Systems*, 2023.
- [15] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024.

- [16] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving CLIP training with language rewrites. In *Neural Information Processing Systems*, 2023.
- [17] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander T Toshev, and Vaishaal Shankar. Data filtering networks. In *International Conference on Learning Representations*, 2024.
- [18] Azade Farshad, Yousef Yeganeh, Yu Chi, Chengzhi Shen, Böjrn Ommer, and Nassir Navab. Scenegenie: Scene graph guided diffusion models for image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 88–98, 2023.
- [19] Enrico Fini, Pietro Astolfi, Adriana Romero-Soriano, Jakob Verbeek, and Michal Drozdal. Improved baselines for vision-language pre-training. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. Featured Certification.
- [20] Jerry A Fodor and Zenon W Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988.
- [21] Roopal Garg, Andrea Burns, Burcu Karagol Ayan, Yonatan Bitton, Ceslee Montgomery, Yasumasa Onoe, Andrew Bunner, Ranjay Krishna, Jason Baldridge, and Radu Soricut. Imageinwords: Unlocking hyper-detailed image descriptions. *arXiv preprint arXiv:2405.02793*, 2024.
- [22] Sachin Goyal, Pratyush Maini, Zachary C Lipton, Aditi Raghunathan, and J Zico Kolter. Scaling laws for data filtering–data curation cannot be compute agnostic. *arXiv preprint arXiv:2404.07177*, 2024.
- [23] Michael Günther, Louis Milliken, Jonathan Geuter, Georgios Mastrapas, Bo Wang, and Han Xiao. Jina embeddings: A novel set of high-performance sentence embedding models. *arXiv preprint arXiv:2307.11224*, 2023.
- [24] Hasan Abed Al Kader Hammoud, Hani Itani, Fabio Pizzati, Philip Torr, Adel Bibi, and Bernard Ghanem. Synthclip: Are we ready for a fully synthetic clip training? *arXiv preprint arXiv:2402.01832*, 2024.
- [25] Laura Hanu and Unitary team. Detoxify. Github. <https://github.com/unitaryai/detoxify>, 2020.
- [26] Roei Herzig, Alon Mendelson, Leonid Karlinsky, Assaf Arbelle, Rogerio Feris, Trevor Darrell, and Amir Globerson. Incorporating structured representations into pretrained vision \& language models using scene graphs. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.
- [27] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in Neural Information Processing Systems*, 36, 2024.
- [28] Yufeng Huang, Jiji Tang, Zhuo Chen, Rongsheng Zhang, Xinfeng Zhang, Weijie Chen, Zeng Zhao, Zhou Zhao, Tangjie Lv, Zhipeng Hu, et al. Structure-clip: Towards scene graph knowledge to enhance multi-modal structured representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2417–2425, 2024.
- [29] Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795, 2020.
- [30] Theo MV Janssen and Barbara H Partee. Compositionality. In *Handbook of logic and language*, pages 417–473. Elsevier, 1997.
- [31] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015.

- [32] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.
- [33] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.
- [34] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.
- [35] Zhengfeng Lai, Haotian Zhang, Bowen Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jiulong Shan, Chen-Nee Chuah, Yinfei Yang, and Meng Cao. Veclip: Improving clip training via visual-enriched captions, 2024.
- [36] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.
- [37] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.
- [38] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR, 23–29 Jul 2023.
- [39] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.
- [40] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23390–23400, 2023.
- [41] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.
- [42] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2023.
- [43] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL <https://llava-vl.github.io/blog/2024-01-30-llava-next/>.
- [44] Rameshwar Mishra and AV Subramanyam. Scene graph to image synthesis: Integrating clip guidance with graph conditioning in diffusion models. *arXiv preprint arXiv:2401.14111*, 2024.
- [45] Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. Improving multimodal datasets with image captioning. In *Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.
- [46] Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, et al. Docci: Descriptions of connected and contrasting images. *arXiv preprint arXiv:2404.19753*, 2024.

- [47] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.
- [48] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 647–664. Springer, 2020.
- [49] Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenbende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. Filtering, distillation, and hard negatives for vision-language pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6967–6977, 2023.
- [50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [51] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.
- [52] Shibani Santurkar, Yann Dubois, Rohan Taori, Percy Liang, and Tatsunori Hashimoto. Is a caption worth a thousand images? A study on representation learning. In *The Eleventh International Conference on Learning Representations*, 2022.
- [53] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.
- [54] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pages 70–80, 2015.
- [55] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019.
- [56] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.
- [57] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449, 2021.
- [58] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.
- [59] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.
- [60] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

- [61] Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. *arXiv preprint arXiv:2312.08578*, 2023.
- [62] Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. Fastvit: A fast hybrid vision transformer using structural reparameterization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.
- [63] Pavan Kumar Anasosalu Vasu, Hadi Pouransari, Fartash Faghri, and Oncel Tuzel. Clip with quality captions: A strong pretraining for vision tasks. *arXiv preprint arXiv:2405.08911*, 2024.
- [64] Pavan Kumar Anasosalu Vasu, Hadi Pouransari, Fartash Faghri, Raviteja Vemulapalli, and Oncel Tuzel. Mobileclip: Fast image-text models through multi-modal reinforced training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [65] Weiyun Wang, Min Shi, Qingyun Li, Wenhui Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. In *The Twelfth International Conference on Learning Representations*, 2023.
- [66] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- [67] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *European conference on computer vision*, 2018.
- [68] Hu Xu, Saining Xie, Xiaoqing Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. In *The Twelfth International Conference on Learning Representations*, 2023.
- [69] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.
- [70] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.
- [71] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *International Conference on Machine Learning*, pages 25994–26009. PMLR, 2022.
- [72] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.
- [73] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. *arXiv preprint arXiv:2403.15378*, 2024.
- [74] Kecheng Zheng, Yifei Zhang, Wei Wu, Fan Lu, Shuailei Ma, Xin Jin, Wei Chen, and Yujun Shen. Dreamlip: Language-image pre-training with long captions. *arXiv preprint arXiv:2403.17007*, 2024.
- [75] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022.
- [76] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019.

Appendix

Table of Contents

A Related works, limitations, and societal impact	16
A.1 Additional related works	16
A.2 Limitations and perspectives	16
A.3 Societal impact	17
B Dataset construction	17
B.1 Query templates	17
B.2 Text classifiers	17
B.3 Other details for data annotation	18
B.4 Computation cost	18
C Dataset information	26
C.1 Data release and licensing	26
C.2 Dataset statistics	26
C.3 Examples from GBC10M	33
D Algorithm details	41
D.1 Structure-aware cross attention	41
D.2 Multi-positive contrastive loss	41
E Experimental details	41
E.1 Data filtering	41
E.2 Dynamic batch size	42
E.3 Hyperparameters for CLIP training	42
E.4 Computation cost	42
E.5 Evaluation details	43
F Additional results and experiments	43
F.1 Matching compute resource for training with short captions	43
F.2 The importance of multi-positive contrastive loss	44
F.3 Impact of caption type on CLIP training	45
F.4 Impact of the underlying graph on retrieval	45
F.5 Evaluating at non-EMA checkpoints	46
G Image attributions	46

A Related works, limitations, and societal impact

This appendix delves deeper into the broader context of our study, examines additional related works, discusses the limitations of our methodologies, and explores its potential societal impacts.

A.1 Additional related works

In this section, we include works related to CLIP training.

CLIP with recaptioning. CLIP [50] is a seminal vision-language model that utilizes text and image encoders to generate joint latent representations. While there is an extensive body of literature on CLIP training—ranging from modifications in the objective [37, 70], data augmentation techniques [19, 40], to training procedures [58, 72]—it is impossible to cover all developments comprehensively here. Among these, particularly relevant to our work is the recent trend that highlights the benefits of enhancing caption quality through dedicated models. For instance, VeCLIP [35] enriches image alt-text with outputs from LLaVA, while similar recaptioning strategies have also been explored by Doveh et al. [14], Nguyen et al. [45], and Vasu et al. [64]. On the other hand, LaCLIP [16] employs LLaMA [60] to rewrite captions. Going further, SynthCLIP [24] leverages a dataset with entirely generated captions and images for CLIP training.

CLIP with additional annotations. There has been a plethora of research on training CLIP models with diverse annotations such as long captions, region captions, and scene graphs. As for long captions, DreamLIP [74] proposes to sample sub-captions from the long description to construct multiple positive pairs, while Long-CLIP [73] addresses CLIP’s 77-token limitation by modifying the positional encoding to accommodate longer text sequences during fine-tuning. Meanwhile, region annotations with varying granularity have been considered by works including GLIP [39], X-VLM [71], and RegionCLIP [75]. Their objectives match features of image crops to their specific descriptions. Efforts that aim to improve CLIP training with the help of scene graphs include CLIP-SGVL [26] and Structure-CLIP [28]. The former integrates scene graphs to define additional objective for image encoder, while the later uses scene graphs to guide the generation of negative captions, and to enrich the text encoder with additional contextual information.

A.2 Limitations and perspectives

We discuss below the limitations of our works from three different perspectives, the procedure and format, the datasets, and the experiments. These limitations also naturally point to several future directions that are to be explored.

A.2.1 Limitation concerning the GBC procedure and format

While GBC remains a versatile high-level annotation format that in principle applies to any image, its design is inherently tied to the coarse-to-fine and compositional nature of natural images. This design orientation means that GBC is not necessarily the most suitable for certain types of images such as scientific imagery, homogeneous patterns, or abstract art. Specifically, scientific imagery often requires annotations that convey precise, quantifiable data rather than relational or descriptive text. This limitation highlights the need for tailored approaches to different visual content categories to address their unique characteristics.

A.2.2 Limitation concerning the GBC datasets

Our datasets are curated with the help of LLaVA and Yolo-World, and hence inherit their limitations. This includes but is not limited to, the bias and hallucination from LLaVA captioning, incorrectly identified objects from Yolo-World, and the inability of Yolo-World to recognize certain object category (see Appendix C.3 for concrete examples). Moreover, our approach mainly distinguishes between objects of the same type via composition nodes. Yet, we believe that there is a more effective strategy than merely assigning numbers to these objects.

A.2.3 Limitation concerning our experiments

Our experiments, which focus on CLIP training and retrieval tasks, demonstrate the benefits of our method and dataset from several perspectives. Nonetheless, we believe this only represents a small part of what this new dataset and annotation method can offer. Moving forward, we commit to

exploring broader applications, particularly in text-to-image and image-to-text problems, to fully leverage our dataset’s potential.

A.3 Societal impact

Our paper introduces the GBC datasets and procedure, both aimed at advancing the development of multimodal models. Specifically, the structured approach of GBC, designed to provide detailed descriptions, may help overcome representational biases inherent in existing captioning pipelines, offering more accurate descriptions of images. The potential benefits of these advancements extend across a range of applications, such as assistive technologies and scientific research. However, alongside these benefits, there are challenges including the potential spread of misinformation and concerns about privacy. A comprehensive discussion of these broader societal impacts, both positive and negative, extends beyond the immediate focus of our methodological study.

B Dataset construction

In this appendix, we provide all the missing details about our dataset construction process that are not mentioned in [Sections 3.2](#) and [3.3](#).

B.1 Query templates

To make the MLLM models fulfill the tasks described in [Section 3.2](#), we perform COT prompting [66] with few-shot examples. The four templates for our queries are shown in [Figure 5](#) to [11](#). We make the following remarks concerning the design of our prompts.

Prompt structure. We craft these prompts with the help of ChatGPT, which results in prompts that might be more complicated than necessary. Meanwhile, we did notice that the inclusion of few-shot examples is crucial for the model to adhere to the required output formats. Given that using always the same few-shot examples might significantly bias the model’s output, it could be beneficial to randomly retrieve examples from a diverse pool for each query, but we did not pursue this exploration.

[Single] and [Multiple] annotations. Since a detection model could output multiple candidate bounding boxes for an input text, we ask the MLLM to annotate each identified element with either [single] or [multiple]. We then proceed with slightly different algorithms in the two cases, to encourage the selection of either only one, or multiple bounding boxes. In particular, we use respectively an NMS threshold of 0.05 and 0.2 for objects labeled with [single] and [multiple]. However, these labels do not necessarily dictate the final count of bounding boxes; multiple boxes may still be selected for items labeled [single], and vice versa.

Dynamically filled-in elements. To ensure that the response of the MLLM is relevant, the prompts are dynamic and reflect the content of the current image (the image query being the only exception). Such information comes from previous queries and can be naturally retrieved for different queries. The only nonobvious part is the *hard coded hints* for composition queries, which we explain below.

Hard coded hints for composition queries. After numerous attempts, we observe that LLaVA-1.6 struggles with accurately describing the composition of multiple objects in a scene, even when these objects are annotated with bounding boxes. To overcome this limitation, we guide the models with hints generated programmatically using a set of predefined rules. Specifically, we begin by constructing a Euclidean minimum spanning tree based on the centers of the bounding boxes. We then select a random node as the root and perform a Depth-First Search (DFS) on the tree. During this search, we interleave descriptions of the edges, which detail the geometric relations between two objects based on the positions of their bounding boxes, with node descriptions. These node descriptions are added when an object is located at a particular extremity of the composition, such as the rightmost or top-left position.

B.2 Text classifiers

Both of our text classifiers are trained for binary classification using logistic loss. To determine whether a piece of text is suitable for object detection, we utilize a single linear layer added on top

of the Jina Embedding.⁵ For the task of assessing whether two texts can represent the same object, we concatenate their Jina embeddings and process them through an MLP. This MLP includes layer normalization, a hidden layer that expands the input dimensionality by a factor of four, followed by SiLU activation and the final linear layer. The dataset for the training of our text classifiers are prepared with the help of ChatGPT.

B.3 Other details for data annotation

We incorporate LLaVA-1.6 into our pipeline using `llama.cpp`.⁶ Moreover, to speed up the annotation process, we utilize models quantized at different precision levels: the vision encoders at 6-bit precision, the LLM component of LLaVA-1.6 Mistral-7B at 5-bit precision, and the LLM component of LLaVA-1.6 Yi-34B at 3-bit precision.⁷ We use the default hyperparameters for inference except for a temperature of 0.1 and context window of size of 5952 (note that LLaVA-1.6 can use up to 2880 image tokens). We discard any responses that do not comply with our required format.

As for the object detection model, we use YOLO-Worldv2-X trained with input resolution of 640×640 .⁸ We set the confidence threshold to 0.05 and retain a maximum of six bounding boxes for each input text, selecting those with the highest confidence scores. We exclude any region whose size is smaller than 5,000. To prevent repetitive descriptions of the same element, we keep only those bounding boxes that occupy less than 80% of the current image region for detections arising from entity queries. Regarding node merging, we consider two bounding boxes to be overlapping if their intersection occupies more than 85% of the area of each bounding box involved.

B.4 Computation cost

We list below the major computation cost of our data preparation process.

- GBC1M: With our processing pipeline, it takes an average of around 3 minutes to annotate each image on an A100 80G when all the queries are performed with LLaVA-1.6 Yi-34B. As a result, annotating 1 million images took us around 6 days with 300 A100 80Gs.
- GBC12M: The average annotation time per image on an A100 80G is improved to 1 minute when relation and entity queries are performed with LLaVA-1.6 Mistral-7B. This process is about twice slower on a V100 32G. In this regard, our GBC12M dataset was compiled in roughly 6 days using 500 A100 80Gs and 1,000 V100 32Gs.

We also compute CLIP score for each caption using the DFN-5B model.⁹ This computation takes around 3 hours for every 10,000 images on a V100 32G.

⁵<https://huggingface.co/jinaai/jina-embeddings-v2-small-en> Accessed: 2024-05-01

⁶<https://github.com/ggerganov/llama.cpp> Accessed: 2024-05-01

⁷<https://huggingface.co/cmp-nct/llava-1.6-gguf> Accessed: 2024-05-01

⁸https://huggingface.co/wondervictor/YOLO-World/blob/main/yolo_world_v2_x_obj365v1_goldg_cc3mlite_pretrain-8698fbfa.pth Accessed: 2024-05-01

⁹<https://huggingface.co/apple/DFN5B-CLIP-ViT-H-14-378> Accessed: 2024-05-01

Query Template for Image Query

System message

As an AI visual assistant, your role is to conduct an in-depth analysis of images and articulate a detailed account of the visible elements. You must then distill this information into a precise and concise caption that accurately reflects the content of the image.

Step-by-Step Process:

Detailed Caption:

- Conduct a thorough examination of the image to note all elements present, including main subjects, minor objects, background details, and any text.
- Prepare a detailed caption that accounts for all these elements, emphasizing the whole objects within the scene.

Top-Level Element Identification:

- Identify and format concrete objects: Begin by identifying concrete objects within the image that are detectable by object recognition models. Each identified object should be formatted as [object_name][node_type] where [node_type] is either [single] or [multiple]:
 - [single]: Applied to items that appear only once in the image, represented as a unique entity within its context, such as a [cat][single] or a [chair][single]. This category is used regardless of the object's size or location in the frame and is intended for items that are not repeated elsewhere in the image. For example, a [stop sign][single] on a street corner or a [tree][single] in a field.
 - [multiple]: Applied to items that are present more than once within the image, emphasizing their plurality. Examples include [dogs][multiple] playing in a park, [chairs][multiple] in a café, [park benches][multiple] along a pathway, [girls][multiple] on a street, [pillows][multiple] on a couch, [paintings][multiple] on a wall, and [lights][multiple] across a ceiling.
- Entire objects only: When identifying elements within an image, only include objects that stand alone as the main subjects. Avoid breaking down the top-level objects into smaller components.
- Grouping similar items: When general items, such as houses, trees, players, or people, appear multiple times in the image, they should be grouped together under a single [multiple] label rather than described separately. This approach applies even if these items might have been described individually in the detailed caption.
- No abstraction: Do not include abstract qualities like colors (blue, red, white), patterns, or expressions.
- No numbering: Do not use any number to label objects. Just use [houses][multiple].
- No directional description: Do not use positional terms for individual elements. Instead, group similar items under a single [multiple] label, like [cowboys][multiple].

Concise Formatted Caption:

- Use the identified elements to construct a concise formatted caption. Use brackets to denote each identified object, following the [object_name][node_type] format. The object name should only appear in the bracket.
- Restrict the number of elements mentioned in the concise caption to avoid overcrowding and ensure clarity. Prioritize the inclusion of key elements that define the scene or the subject's essence.
- The concise caption should contain at most two sentences.

Example Adjustments:

- Character attributes: When analyzing an image featuring a person with distinctive attributes such as armor or tattoos, focus on the person as a whole rather than the individual attributes. The correct annotation would be [person][single], encompassing all aspects of the person appearance without breaking them down into separate elements.
- Architectural features: In the case of architectural elements, avoid itemizing components like the roof, windows, or door if they contribute to the overall structure of a building. For a singular building in the image, use [house][single]. If the image depicts a series of buildings, such as a row of houses with varying designs, annotate them collectively as [houses][multiple], regardless of their individual features.
- Groups of similar objects: For scenes containing groups of similar objects or individuals, such as girls playing in a park, group them under a single [multiple] label. Even if the individuals are engaged in different activities or have distinct appearances, they should be annotated as [girls][multiple] to emphasize their collective presence within the scene. Similarly, even if multiple dogs or chairs have different colors, they should be labeled as [dogs][multiple] and [chairs][multiples].

Figure 5: The system prompt used for image query (first half).

Query Template for Image Query

System message (continued)

Example Captions:

For an image featuring multiple elements like a logo:

Detailed Caption: A design showcasing a prominent grey 'N' at the top, with three smaller NEO Business Bank logos directly below it, two colored squares positioned to the bottom left, and a line of text to the bottom right detailing the availability of various file formats for the design. Top-Level Element Identification:

- ['N'][single]
- [Logos][multiple]
- [Squares][multiple]
- [Text][single]

Concise Formatted Caption: A design showcasing grey ['N'][single] positioned over NEO Business Bank [logos][multiple], accompanied by colored [squares][multiple] and [text][single] at the bottom.

For an illustration of a zebra:

Detailed Caption: An animated zebra stands upright on two legs, waving in a welcoming manner, next to a wooden signpost at the beginning of a dirt path. This path leads to a quaint wooden cabin with a thatched straw roof, surrounded by a simple wooden fence. In the background, there's another similar cabin. The scene is completed by a clear sky overhead and multiple trees dotting the landscape, contributing to the lush greenery. Contextual Considerations: The zebra's legs are part of its overall form and should not be listed separately.

Top-Level Element Identification:

- [Zebra][single]
- [Signpost][single]
- [Dirt path][single]
- [Cabins][multiple]
- [Trees][multiple]
- [Sky][single]

Concise Formatted Caption: An animated [zebra][single] waves next to a wooden [signpost][single] on a [dirt path][single] that leads towards wooden [cabins][multiple], with [trees][multiple] enhancing the lush greenery under a clear [sky][single].

For a photo of two men on street:

Detailed Caption: A photo of two men standing side by side on a city street. The man on the left has long hair and is wearing a beige blazer over a white shirt with black trousers. He is smiling and looking directly at the camera. The man on the right has short hair and is dressed in a gray blazer over a black shirt with gray trousers. He also smiles at the camera. They are standing on a sidewalk lined with shops and buildings, suggesting they are in a commercial or urban area. The lighting suggests it might be late afternoon or early evening. Contextual Considerations: The two men, despite their distinct appearances and attire, should be grouped together under a single label since they both fall under the category of "men".

Top-Level Element Identification:

- [Two men][multiple]
- [Sidewalk][single]
- [Shops][multiple]
- [Buildings][multiple]
- [City street][single]

Concise Formatted Caption: [Two men][multiple] stand side by side on a [sidewalk][single] along a [city street][single], lined with [shops][multiple] and [buildings][multiple], each dressed in coordinated blazers and trousers.

User message:

Following the instruction, please provide a detailed caption and a concise formatted caption for the given image. Note that it is crucial for you to put elements that can be detected and should be further described in picture in brackets as [object_name][node_type] in the concise formatted caption.

Figure 6: The system and user prompts used for image query (second half).

Query Template for Entity Query

System message

Your task is to perform an in-depth analysis of a cropped image focusing on a requested object, like a "house". The process involves a step-by-step evaluation to identify the object's presence, describe its features, craft concise captions, and assess any prominent objects.

Process Overview:

Verify Object Presence:

- Examine the image to determine if the specified object, or any instance of it, is present.
- State the presence with "Object Present: Yes" or "Object Present: No".

Provide Appropriate Caption (If Object Is Present):

- Provide a detailed description of the object, focusing solely on its features without reference to other elements in the image.
- The description should contain at most 50 words.

Assessment of Prominent Objects:

- Evaluate the described features to determine if any stand out for further description and are detectable by an object detection model. This is crucial for complex objects such as 'man', 'woman', 'family', 'couple', 'cat', or 'house', where components or distinctive attributes are significant. For example, when analyzing 'woman', consider elements like dress [single], skirt [single], or hair [single] as prominent features. For simpler objects like 'cup' or 'chair', additional descriptions may not be needed.

Identification of Prominent Features (If Applicable):

- If there are prominent features identified, list and format these features for potential detection by an object detection model.
- Ensure these features are parts or components of the main object and not the entire object itself.
- Use [single] for unique, standalone items, and [multiple] for features present more than once, such as roof [single] or windows [multiple].
- Group similar items under a single [multiple] label rather than describing them separately, even if individual descriptions were provided in the detailed caption. For example, multiple distinct windows in a house should be labeled as windows [multiple] rather than individually enumerated.
- For groups like families or couples, identify members separately (e.g., man [single], woman [single]) rather than as a collective unit. This contrasts with grouping similar inanimate objects (e.g., windows [multiple]), where individual distinction isn't necessary.
- Consistency with the caption: Ensure that the features identified as [single] or [multiple] are also mentioned in the caption.

Example Responses:

Example 1: Object Not Present

Object Presence: No

Detailed Caption: N/A

Prominent Features: N/A

Identification of Prominent Features: N/A

Example 2: Object Present Without Prominent Features (requested object: "cup")

Object Presence: Yes

Detailed Caption: A simple ceramic cup on a wooden table. The cup has a smooth, unadorned surface and a standard curved handle on one side.

Prominent Features: No

Identification of Prominent Features: N/A

Figure 7: The system prompt used for entity query (first half).

Query Template for Entity Query

System message (continued)

Example 3: Object Present With Prominent Features (requested object: "family")

Object Presence: Yes

Detailed Caption: A family of four is captured enjoying a sunny day in the park. The father, in casual attire, is engrossed in setting up a picnic spot, while the mother, donned in a summer dress, is laying out a feast on a blanket. Nearby, two children, a boy and a girl, engage in playful antics; the boy is kicking a football with fervor, and the girl, adorned in a light frock, is gleefully chasing bubbles.

Prominent Features: Yes

Identification of Prominent Features:

- Father: [single]
- Mother: [single]
- Boy: [single]
- Girl: [single]

Example 4: Object Present With Prominent Features (requested object: "car")

Object Presence: Yes

Detailed Caption: A vintage car in pristine condition, with shiny chrome bumpers and classic spoke wheels. The car's body is painted in a vibrant red, and the leather interior is visible through the clear windows. A unique hood ornament adorns the front, adding to the car's elegance.

Prominent Features: Yes

Identification of Prominent Features:

- Chrome bumpers: [single]
- Wheels: [multiple]
- Hood ornament: [single]

User message:

Please assess the image focusing on '{ }'. Start by confirming its presence with 'Object Present: Yes' or 'Object Present: No'. If present, describe its key features in a detailed caption with at most 50 words. Then, evaluate if any aspects stand out for further emphasis, stating 'Prominent Features: Yes' or 'No' while preferring "Yes". If yes, list a few notable features in brackets, applying [single] or [multiple] as appropriate. Importantly, do not include '{ }' in features. Instead, you should break it down.

Example filled-in elements

- boat
- boat



Figure 8: The system and user prompts used for entity query (second half). The placeholders '{ }' are dynamically filled with the name of the object, i.e., the label of the associated incoming edge.

Query Template for Composition Query

System message

Your role is to analyze images containing objects within pre-labeled bounding boxes and describe the compositional arrangement of these objects based on provided hints. You will then provide general descriptions that apply to all the objects collectively.

Input Image Format Explanation:

- The image will feature objects of interest, each enclosed within a bounding box.
- Each bounding box will be numbered centrally to uniquely identify it.
- The objects will be similar in nature (e.g., all dogs) and positioned within a scene.

Utilizing Hints for Analyzing Composition:

- Begin by reviewing the hints provided regarding the spatial arrangement of the objects.
- These hints may specify the relative positions of objects (e.g., "Object 3 is in the top right corner").
- Use the hints to guide your description of how the objects relate to each other within their bounding boxes.

Output Format:

- Composition Description: Start with "Composition:" followed by a description informed by the hints and using the bounding box numbers. This description should elucidate the spatial arrangement of the objects as per the hints.
- General Descriptions: Provide observations that apply to all objects within the specified group, excluding unrelated elements or background details. Preface this section with "General descriptions:".

Additional Guidelines:

- Describe the spatial arrangement of objects without inferring spatial relations from the sequence of numbers.
- Utilize clear spatial language to articulate the composition.
- The description should reflect the actual visual composition, not the order of numbers in the bounding boxes.

Examples:

Example for 3 Dogs in Bounding Boxes:

Query Prompt: "Please describe the composition of the 3 dogs in the bounding boxes, followed by some general descriptions that apply to all dogs."

System Response:

Composition: Dog 3 is in front, with dog 2 to the left and dog 1 to the right.

General descriptions:

- The three dogs are aligned in a row on the grass.
- They share similar sizes and features, suggesting they may be from the same breed.

Additional Examples:

For 5 Flowers in a Garden Bed in Bounding Boxes:

Composition: Flower 4 takes a central position, flanked by flower 2 and flower 3 on either side, while flower 1 and flower 5 bookend the arrangement at the outer edges.

General descriptions:

- Each flower is in full bloom, indicating a peak growing season.

For 2 Cats in a Window in Bounding Boxes:

Composition: Cat 1 is positioned on the left side of the window sill, while cat 2 is curled up on the right.

General descriptions:

- Both cats are basking in the sunlight coming through the window.
- Their relaxed postures suggest a shared sense of comfort and tranquility.

Figure 9: The system prompt used for composition query.

Query Template for Composition Query

User message:

Please describe the composition of the {} in the bounding boxes, followed by some general descriptions that apply to all {}. The composition should include {} and be based on the following hints (do not mention hints or bounding boxes in the response).

{}

Example filled-in elements

- boats
- boats
- boat 2, boat 3, boat 1
- - boat 2 is on the right side of the composition
 - boat 2 is to the right of boat 3
 - boat 3 is to the right of boat 1
 - boat 1 is on the left side of the composition



Figure 10: The user prompt used for composition query. The placeholders ‘{}’ are dynamically filled with the name of the object, the labels of the out edges (in the form of the name of the object plus number), and hard coded hints that are obtained using the positions of the bounding boxes.

Query Template for Relation Query

System message

Your role involves analyzing the spatial and direct interactions between pre-identified elements within an image, described through annotations like [beach], [turquoise waters], [people], [shoreline], [lush vegetation]. Your task is to objectively describe how these elements are related or positioned relative to each other within the scene.

Steps for Identifying Objective Relations:

1. Review Annotated Elements: Start by examining the list of annotated elements. Understand the nature of each element as it is presented in the image.
2. Identify Spatial Positions: Determine the spatial positioning of these elements in relation to each other. Focus on direct relationships such as touching, overlapping, or proximity without implying any further interpretation.
3. Describe Direct Interactions: Look for and describe any direct interactions between elements, such as one element supporting another, blocking, or leading into another.
4. Format Relations List: Provide your findings as a list of bullet points. Each bullet should detail a direct and observable relationship between two or more elements, using their annotated identifiers for reference.

Example Relations Based on Annotated Elements:

For elements: [beach], [turquoise waters], [people], [shoreline], [lush vegetation], you might reply:

- The [people] are standing on the [beach], with the [lush vegetation] to their left.
- [Turquoise waters] lap against the [beach] at the [shoreline], with [people] scattered along its edge.
- [Lush vegetation] flanks the left side of the [beach], providing a natural border.
- The [shoreline] separates the [beach] from the [turquoise waters].
- To the right of the [lush vegetation], the [beach] stretches towards the [turquoise waters].

For another set of elements: [eagle], [snake], [wings], you might reply:

- The [eagle] has its [wings] spread above the [snake].
- The [snake] is positioned below the [eagle].
- The [eagle]'s claws are near or in contact with the [snake].

Guidelines for Reporting Relations:

1. Ensure descriptions are based solely on visible or directly inferable relationships without adding interpretations or assumptions.
 2. Maintain clarity and precision in articulating the spatial and interactional dynamics between elements.
 3. Stick to objective descriptions that reflect the physical and observable aspects of the elements' relationships.
 4. Only answer the list of bullet points without anything before or after.
 5. Do not include any bullet point with 1 or even 0 elements.
- Visible Relationships Only: Report relationships that are clearly depicted in the image. If no clear relationships are visible, state "No visible relationships."
 - Objective Descriptions: Keep descriptions factual and based solely on what can be seen in the image.
 - Avoid Assumptions: Do not infer or assume any relationships that aren't clearly shown in the image.
 - Bullet Point Format: Present each observable relationship as a separate bullet point, avoiding any descriptive text not related to the direct relationships.
 - No Relation Inference: Refrain from implying relationships or positions that are not explicitly shown. If elements are simply present without any discernible interaction, it is acceptable to say "Elements are present without visible interaction."
 - Avoid Single Element Points: Do not include bullet points that mention only one element or have no elements at all. Each bullet point must reference the relationship between two or more elements.

User message:

Please infer and list of at most {} relations between {} in this images.

Example filled-in elements:

- 4
- eiffel tower, sky, trees, boat, water



Figure 11: The system and user prompts used for relation query. The placeholders ‘{}’ are dynamically filled with a random number between 2 and the number of involved objects, and the names of these objects.

C Dataset information

In this appendix, we provide information about dataset release, dataset statistics, and visualizations of a few examples from our GBC10M dataset.

C.1 Data release and licensing

Our datasets are available at <https://huggingface.co/graph-based-captions>, released under the Apple Sample Code License.¹⁰ Following CC12M, we include URLs to images along with captions generated through our GBC procedure, all stored in JSON lines format. Comprehensive documentation including a dataset card and croissant metadata is provided in the data repository. We are committed to maintaining the dataset to ensure its long-term public accessibility.

Personal identifiable information and offensive content. Our dataset comprises only captions generated by MLLM models (LLaVA 1.6 Yi-34B and LLaVA 1.6 Mistral-7B), which were trained on carefully curated data. The images, sourced from CC12M, are generally free from offensive content. In particular, CC12M is the result of a filtering operation involving *adult content detection* on images and their captions. While CC12M images may include human faces, we do not host the images directly; only the URLs are provided. Additionally, we conduct toxicity check with Detoxify [25] on a subset of examples in GBC dataset and find no harmful contents. While it was not possible to manually examine all the samples produced by GBC pipeline, we believe that the protective measures of the source dataset and model are sufficient to avoid both harmful content, and privacy leakages.

Author statement. We, the authors, accept full responsibility for any violation of rights.

C.2 Dataset statistics

In this section, we provide statistical insights into the GBC1M and GBC10M datasets. In particular, we zoom in on the statistics at image, vertex, edge, and caption levels, and present distributions of several key metrics including for example caption length, region size, and CLIP score. Since most of these metrics exhibit long-tailed distributions, we often group excessively large values into a single histogram bin for better visualization.

C.2.1 Image and graph statistics

We first look at the sizes of the images and of the annotation graphs, i.e., the numbers of vertices and edges in these graphs and their diameters (which is measured as the length of the longest path in a directed graph). The distributions of these metrics are shown in Figures 12 and 13. We see that the image size has a very long-tailed distribution, with the majority of images having around 786×786 pixels. Conversely, the distributions of graph diameters are more similar to that of a Poisson or a binomial distribution, with most of the graphs having a diameter between 3 and 6. Finally, as one could expect, the numbers of vertices and edges share quite similar distributions.

While we expect the size of a graph to reflect the inherent complexity of an image, we acknowledge that our annotations are influenced by the biases of the used models. In particular, we observe that our annotation process tends to yield larger graph for natural images compared to other types of images such as artworks or graphic designs.

C.2.2 Vertex statistics

We have shown previously that our datasets contain an average of 12 vertices per graph. This translates to 11 regions per image after excluding the root node that represents the entire image. We compare this number with several other vision-language datasets with region-based annotations in Table 6. As one can see, this number aligns well with many of these datasets, particularly those used for detection, such as COCO and Object365. However, it lags behind compared to Visual Genome and more recent datasets with dense annotations, such as AS-1B and DCI. We believe this discrepancy can be attributed to both the top-down design of our annotation process, which tends to overlook less significant components of the images, and the limitations of the detection model used. Notably, both AS-1B and DCI utilize Segment Anything [32] to identify regions of interest. Segment Anything is

¹⁰Should we receive the necessary approvals, we may transition to a less restrictive license in the future.

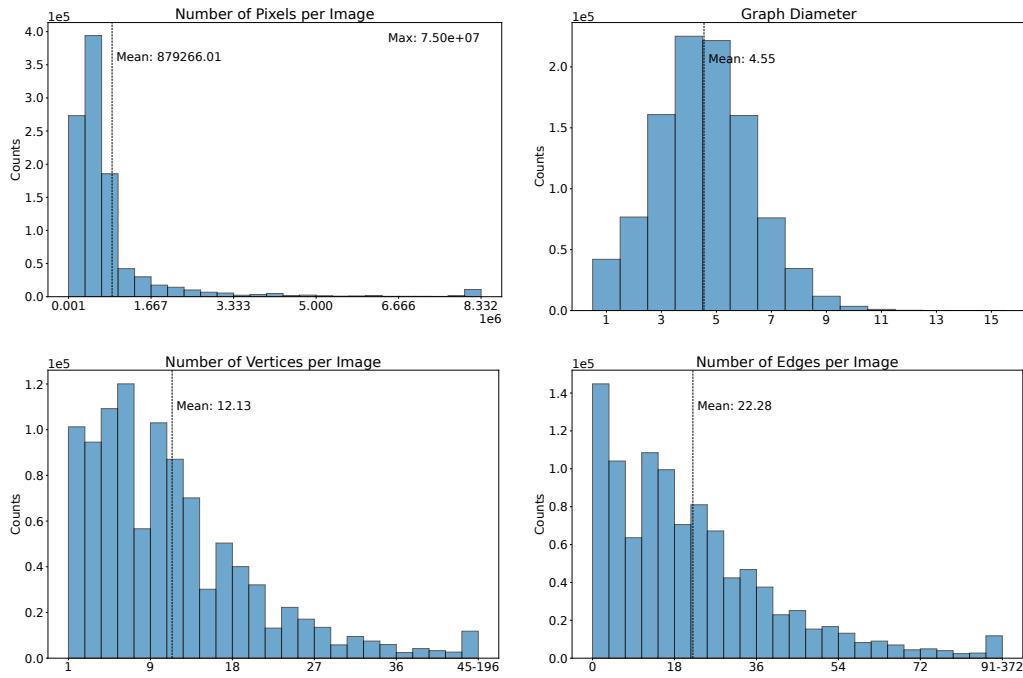


Figure 12: Distributions of metrics at image and graph level in the GBC1M Dataset.

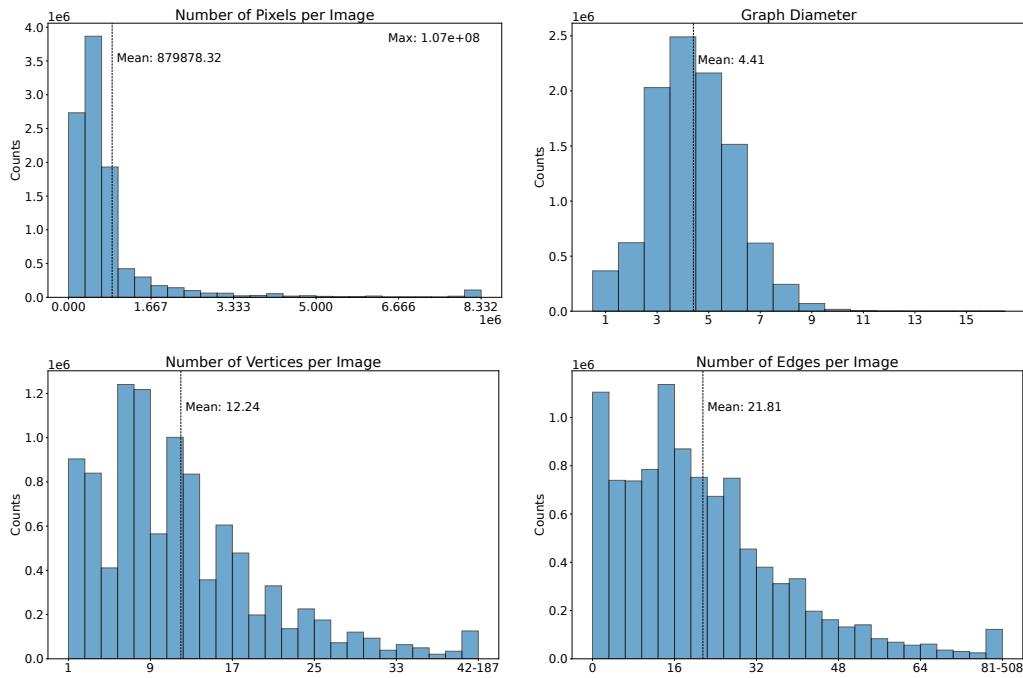


Figure 13: Distributions of metrics at image and graph level in the GBC10M Dataset.

Dataset	# Images	# Regions / Image
COCO [41]	123,000	7
Visual Genome [33]	108,249	42
Objects365 [55]	638,000	16
Open Images [34]	1.7M	8
BigDetection [5]	3.5M	10
SA-1B [32]	11M	100
AS-1B [65]	11M	110
DCI [14]	7,805	40
GBC1M (ours)	1.1M	11
GBC10M (ours)	10.1M	11

Table 6: Comparison of number of regions per image among several vision-language datasets with region-based annotations. We use the statistics reported in the original paper although some datasets, such as COCO and Open Images have been updated after their initial release. Moreover, for Open Images we report the number for the training set with bounding box annotation [34, Tab. 5]. For DCI, we compute the average number of regions per image ourselves using their open-sourced dataset with 7,805 images as this number is not reported in [14].

trained on SA-1B, which has much denser annotations compared to the object detection datasets used for training Yolo-World.

We next examine how this number is distributed across the different types of nodes that are present in our graphs. For this, we plot the distributions of the numbers of composition nodes, relation nodes, entity nodes, and leaves (i.e., the nodes without any children) in Figures 14 and 15. As seen in the figures, a large number of vertices are entity nodes, which focus on describing a single object. In spite of this, we still have an average number of 4 vertices per graph that are dedicated to describing the composition or relationships between multiple objects.

To complete our investigation, we visualize the distributions of the sizes of the vertices’ bounding boxes in Figures 16 and 17. We note that most of the regions have small relative size (smaller than 0.1). This is also observed in other datasets such as Visual Genome [33, Fig. 15] and Open Images [34, Fig. 20]. Relation nodes, whose bounding boxes are defined as the minimum bounding box containing the union of all the involved objects’ bounding boxes, have sizes that spread more uniformly across different ratios. We also observe a large number of entity nodes with bounding boxes that have a relative size close to 1. This likely corresponds to background objects that spans across the entire image, such as “sky” or “grass”.

C.2.3 Edge statistics

Our datasets feature an average of 22 edges per graph. We analyze the origins of these edges in Figure 18, which shows their distributions across different types of source vertices. The figure indicates that the image node is responsible for a large proportion of these edges, suggesting that many of the entities that we identify directly come from the image caption. This is natural provided that an image often contains many objects, while it is less common to need further decomposition of a single object for detailed description. Besides this, these figures also indicate the number of entities that are involved in our composition and relation descriptions. Notably, we see that most of these descriptions only contain 2 or 3 objects, with few of them involving more than 4 objects. In contrast, we observe a relatively large number of entity nodes with 4 outgoing edges, and we believe this can be attributed to the bias caused by the few-shot examples provided in our query template.

We also provide analysis for the edge labels. These edge labels should represent the objects that are associated to their respective target vertices. In particular, during our annotation process, we use these labels as input of the detection model to obtain the bounding boxes of the entity nodes. In Figures 21 and 23, we plot the distributions of the numbers of words and tokens contained in the edge labels. As expected, most of the time we use only 1 or 2 words to represent the entities.

We next study the content of these labels. To this end, we plot the distribution of (*i*) the 20 most common edge labels at the in-edges of the entity nodes, reflecting the content of these entity nodes,

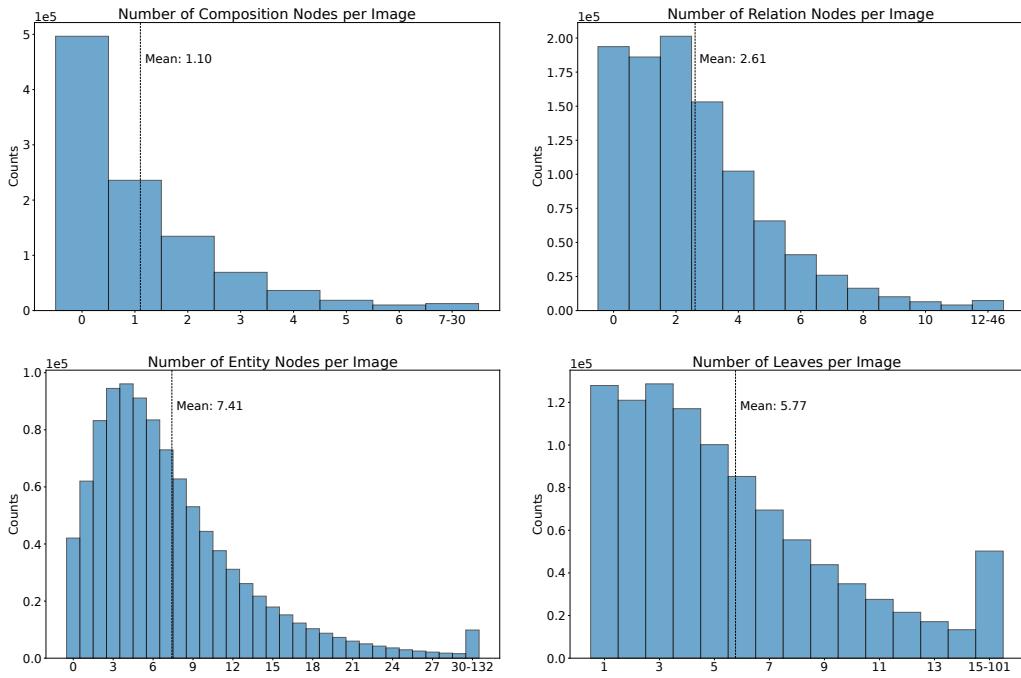


Figure 14: Distributions of vertex numbers across different types of vertices in the GBC1M Dataset.

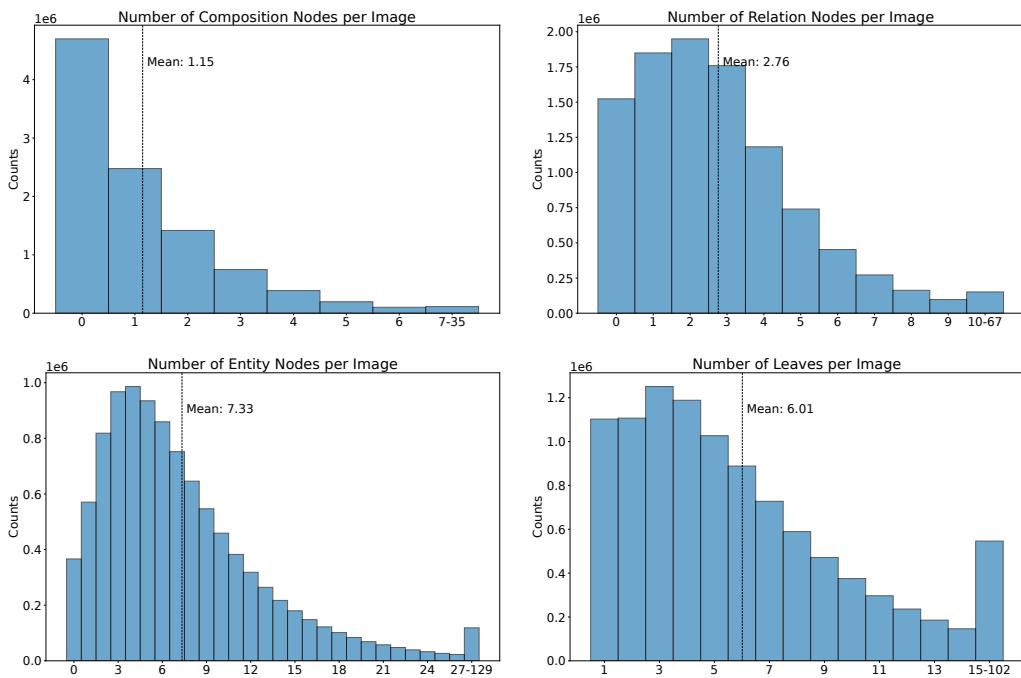


Figure 15: Distributions of vertex numbers across different types of vertices in the GBC10M Dataset.

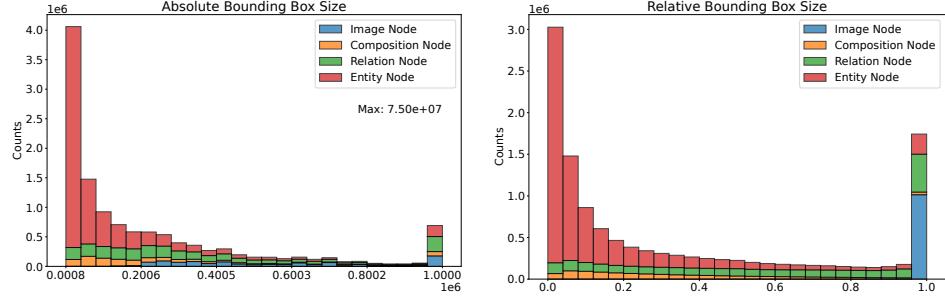


Figure 16: Distribution of bounding box sizes in the GBC1M Dataset. We show both the absolute size (number of pixels) and the relative size (normalized by image size).

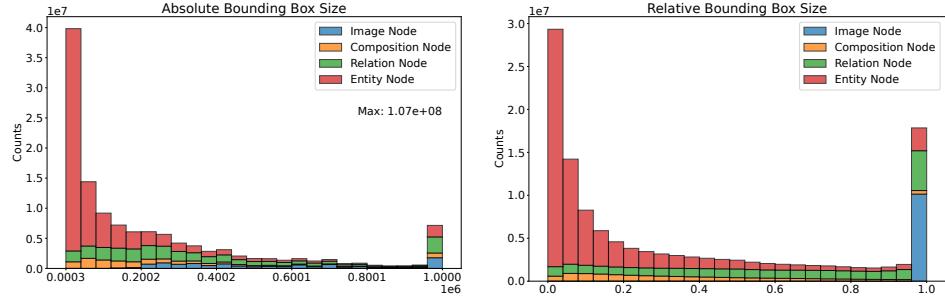


Figure 17: Distribution of bounding box sizes in the GBC10M Dataset. We show both the absolute size (number of pixels) and the relative size (normalized by image size).

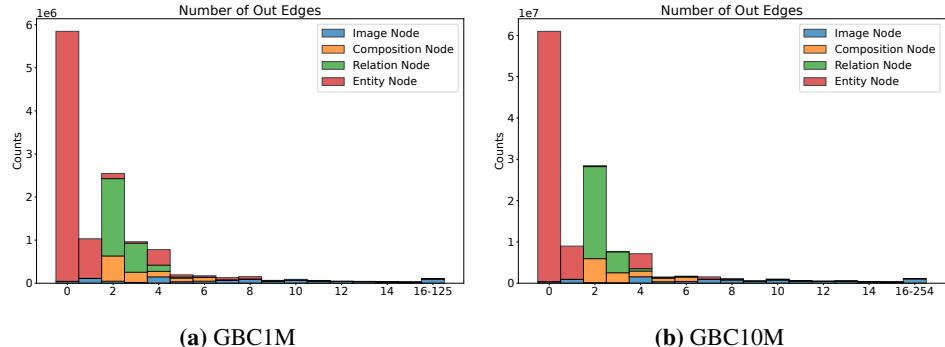


Figure 18: Distributions of outgoing edges across different types of vertices in the GBC1M (left) and GBC10M (right) datasets.

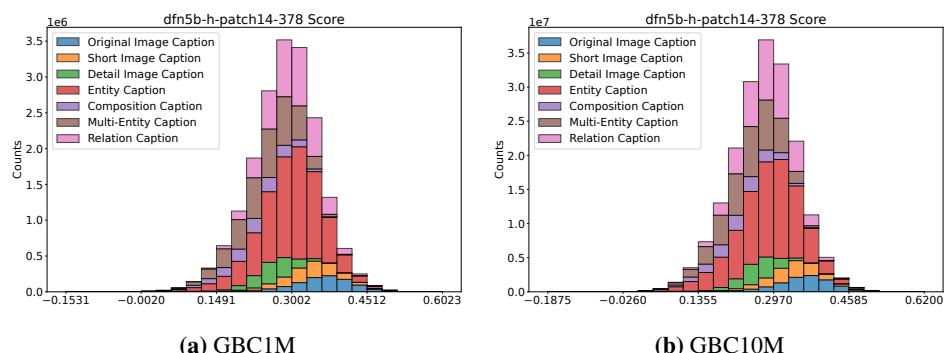


Figure 19: Distributions of DFN-5B CLIP scores across different types of captions in the GBC1M (left) and GBC10M (right) datasets.

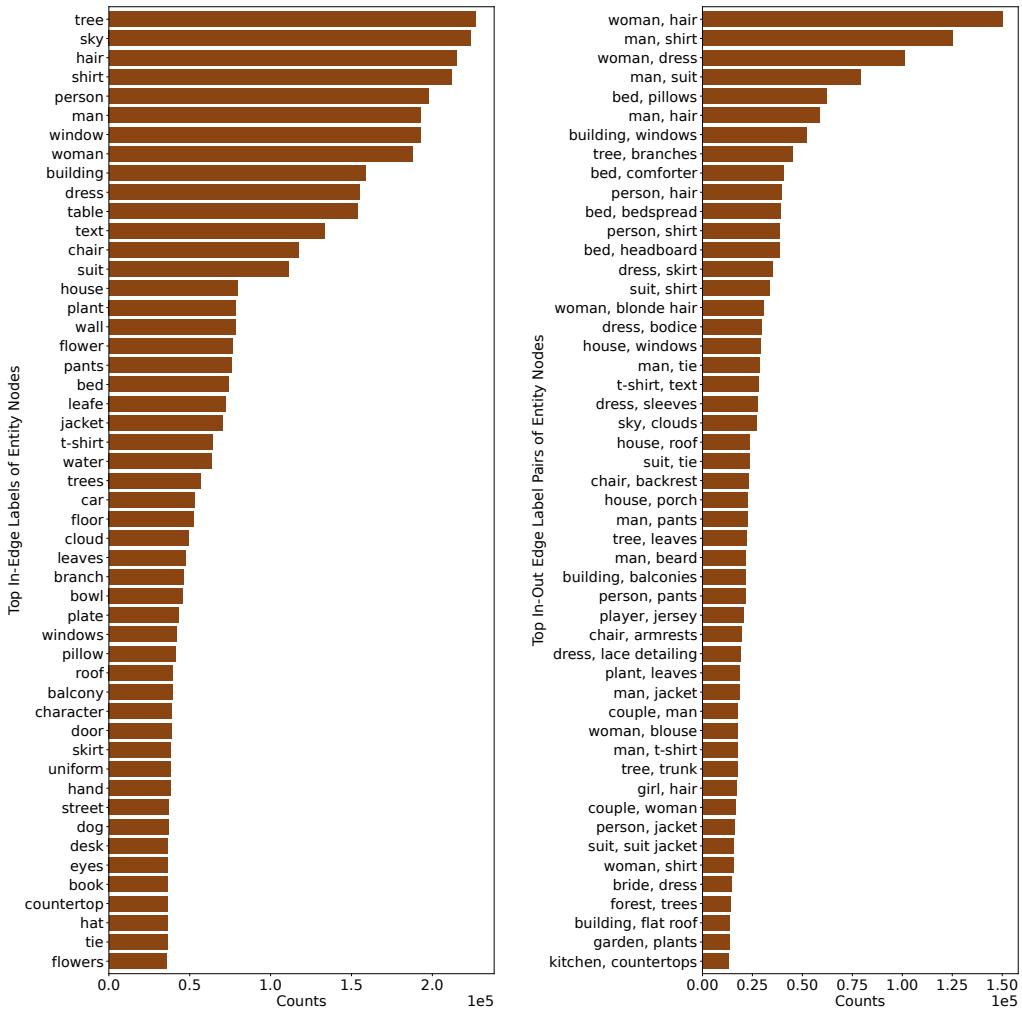


Figure 20: Distributions of the 20 most common in-edge labels and in-/out-edge label pairs at entity nodes in the GBC1M dataset. We remove numbers from the edge labels for the computation of their occurrences in these plots.

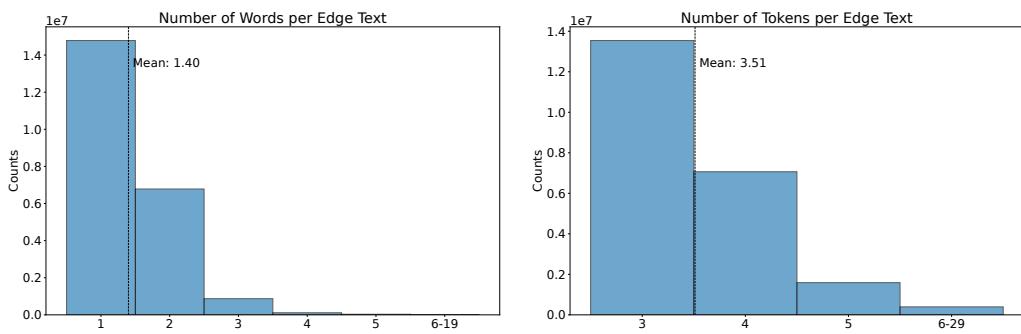


Figure 21: Distributions of numbers of words/tokens in each edge label in the GBC1M dataset. To compute the number of tokens we use the standard CLIP tokenizer.

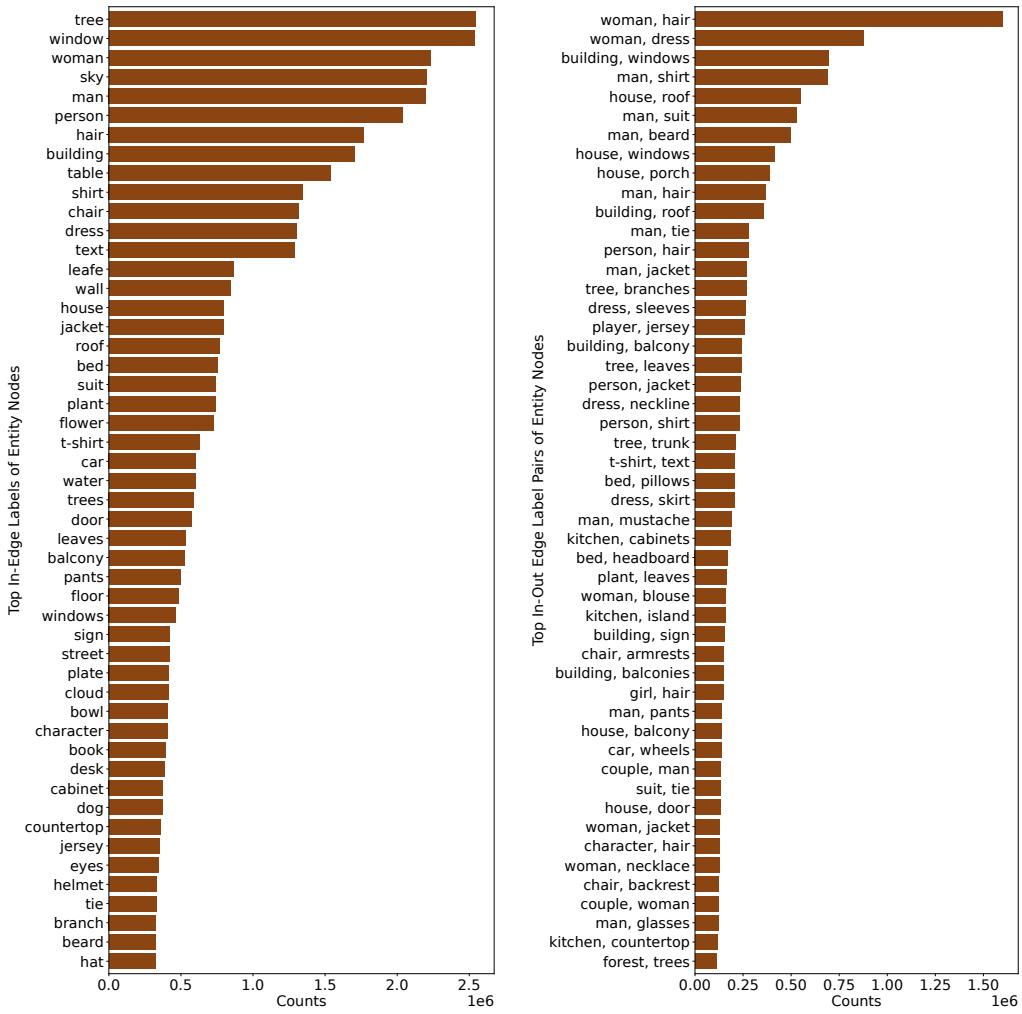


Figure 22: Distributions of the 20 most common in-edge labels and in-/out-edge label pairs at entity nodes in the GBC10M dataset. We remove numbers from the edge labels for the computation of their occurrences in these plots.

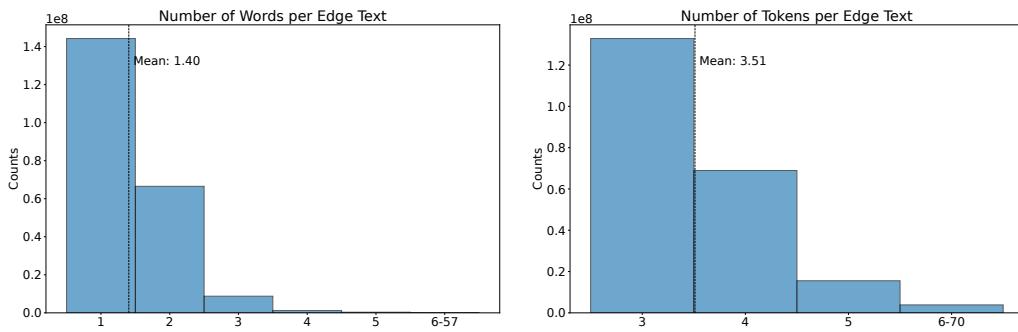


Figure 23: Distributions of numbers of words/tokens in each edge label in the GBC1M dataset. To compute the number of tokens we use the standard CLIP tokenizer.

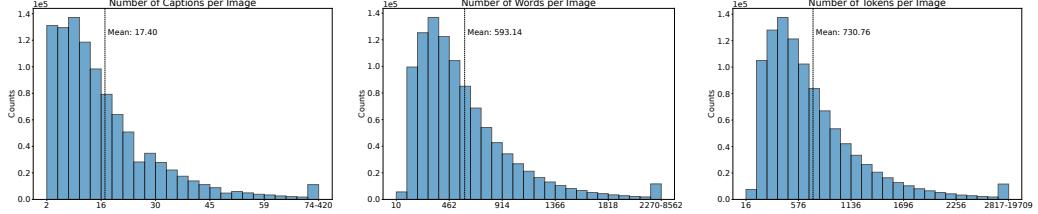


Figure 24: Distributions of numbers of captions, words, and tokens per image in the GBC1M Dataset.

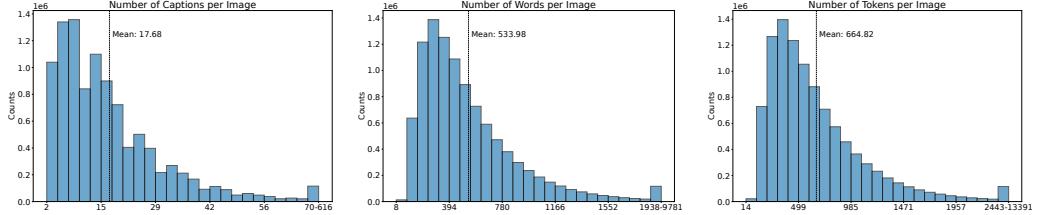


Figure 25: Distributions of numbers of captions, words, and tokens per image in the GBC10M Dataset.

and (ii) the 20 most common edge label pairs when pairing the in- and out-edges of the entity node, reflecting the situation where we zoom in on an object to further describe a part of it. The corresponding histograms are presented in Figures 20 and 22. From these plots, we see that the most common objects from our datasets are “tree”, “sky”, “man”, “woman”, “table”, and “building”, among others. This distribution aligns well with the ones reported for existing datasets, cf. [33, Fig. 22] and [34, Tab. 11]. Furthermore, while the occurrence of certain labels and label pairs, such as (“woman”, “hair”), may be influenced by our system prompts, others like (“bed”, “pillows”) are widely present despite not being included in our prompts. This suggests potential biases in either the model or the dataset itself.

C.2.4 Caption statistics

For statistics at the caption level, we first complete Table 1 and Figure 3 by providing distribution of CLIP scores on the two datasets in Figure 19, and distribution of number of captions, words, and tokens per image in Figures 24 and 25. In particular, the significant variation in CLIP score distributions across different caption types motivates our decision to perform CLIP-filtering independently for each type, as mentioned in Section 5.2.

Going further, we report the average number of words and tokens per caption across different types of captions in Figure 27, 29, and Table 7. We can see that except for the detailed image captions, most captions indeed contain fewer than 77 tokens. Table 7 additionally reveals that we have near 2.5 times more region captions (i.e., entity and multi-entity captions) than the total of relation and composition captions. However, as we have seen in Section 5.3 and will further ablate in Appendix F.3, these relation and composition captions, unique to our dataset, are crucial for the performance improvement that we observe across different evaluations.

We conclude this part by showing the distribution of the 20 most common words and trigrams that appear in our captions, with stop words removed when considering the word distributions. The frequent appearances of colors among the top words again align with the distribution reported in Visual Genome [33, Fig. 24]. In addition, phrases like “appears to be”, “possibly”, and “the image captures” that commonly appear in our data, reflect LLaVA’s use of GPT-generated data during instruction tuning.

C.3 Examples from GBC10M

As a complement to the dataset statistics presented in the previous section, we showcase a few illustrative examples from GBC10M in Figures 30 and 31. These examples demonstrate the varying levels of graph complexity across our dataset. The number of nodes varies from just a few (first

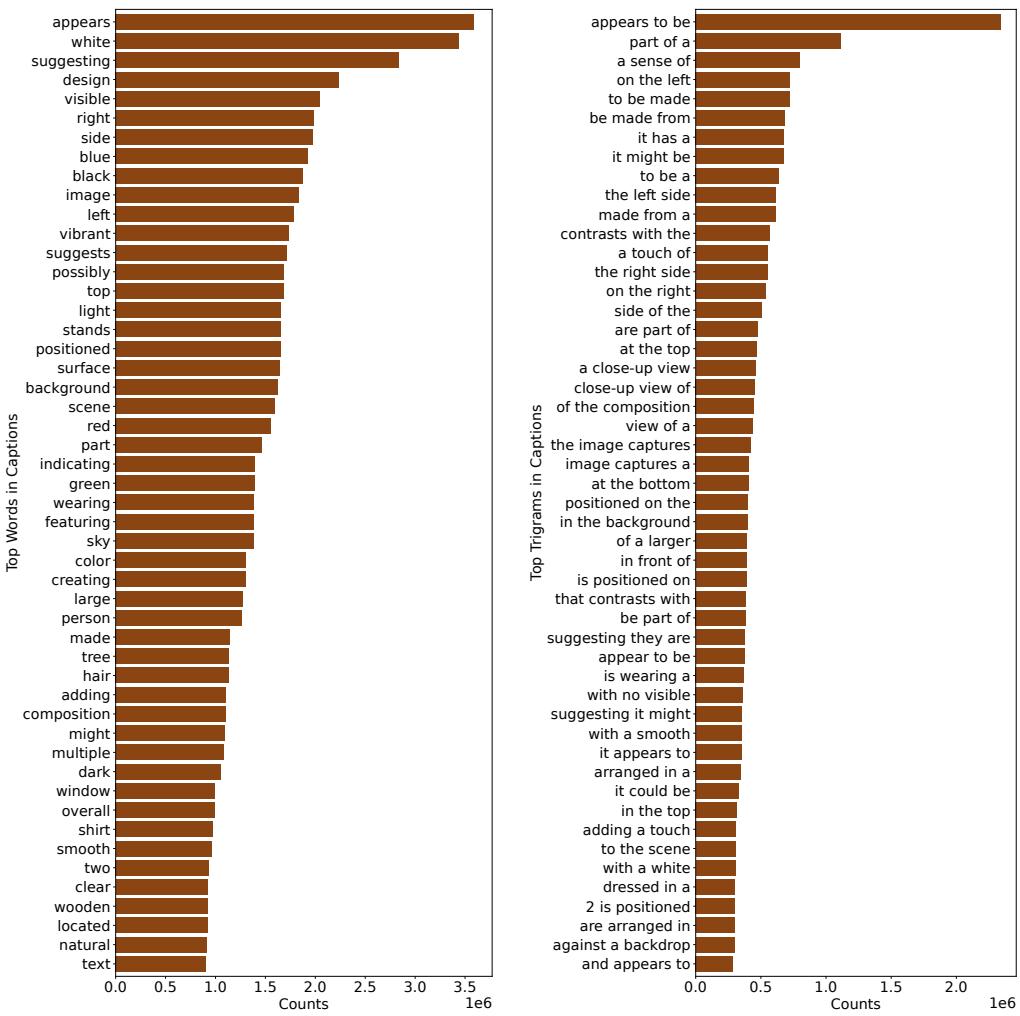
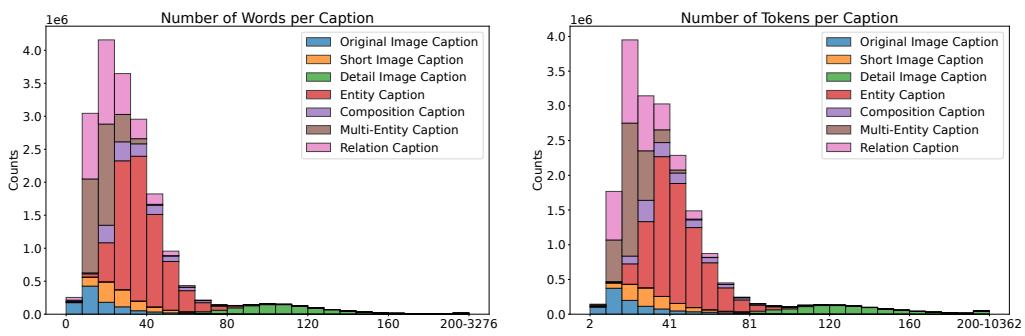


Figure 26: Distributions of the 20 most common words and trigrams that appear in the captions of the GBC1M dataset.



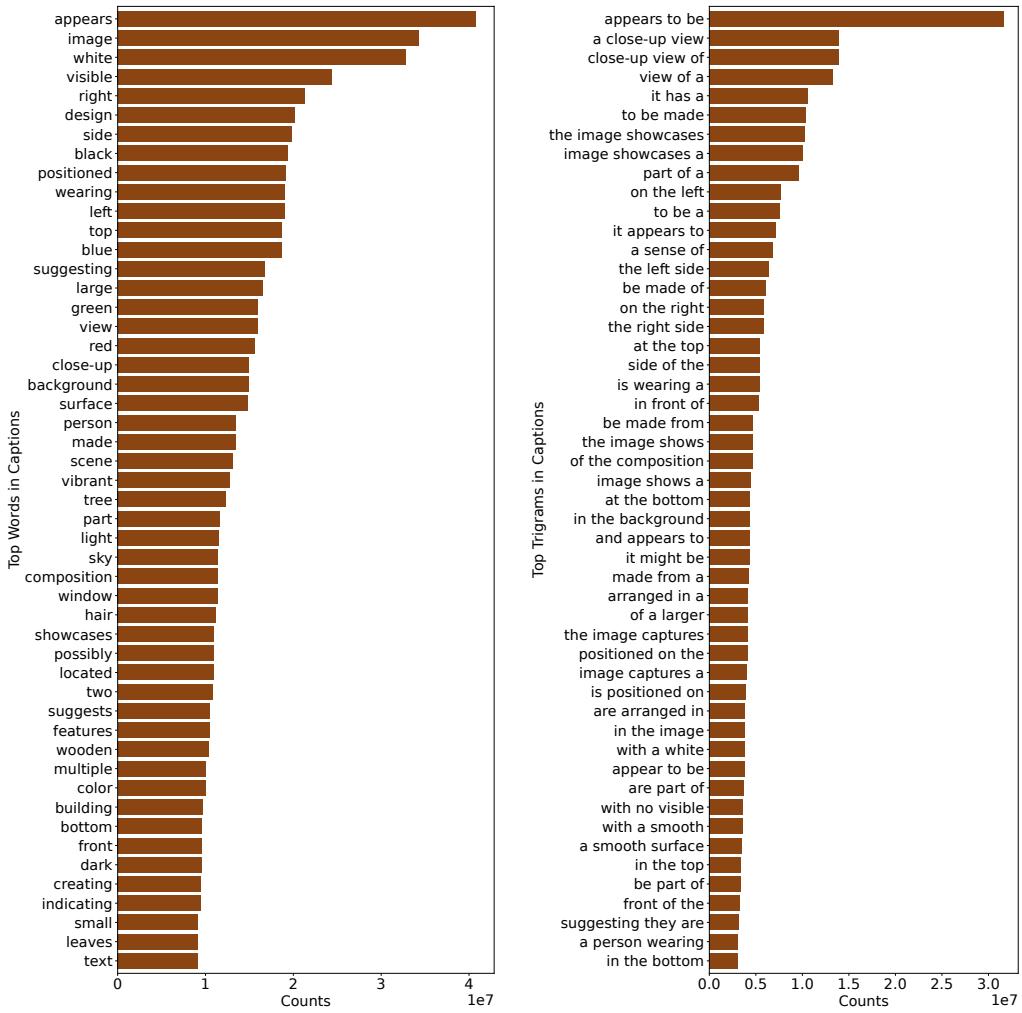


Figure 28: Distributions of the 20 most common words and trigrams that appear in the captions of the GBC10M dataset.

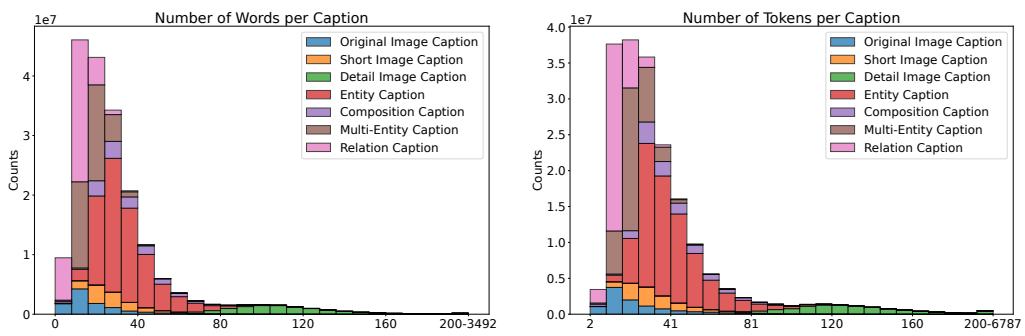


Figure 29: Distributions of numbers of words/tokens across different types of captions in the GBC10M dataset. To compute the number of tokens we use the standard CLIP tokenizer.

Caption Type	# Captions	# Words / Caption	# Tokens / Caption	CLIP score
Image Original	1,013,592	17.4	24.5	0.36
Image Short		28.1	35.3	0.33
Image Detail		110.3	130.9	0.26
GBC1M	Entity	7,512,638	37.5	0.29
	Composition	1,117,935	35.8	0.23
	Multi-Entity	3,487,562	17.8	0.25
	Relation	3,493,543	22.0	0.30
Image Original	10,138,757	17.4	24.6	0.36
Image Short		28.1	35.3	0.33
Image Detail		110.3	130.9	0.26
GBC10M	Entity	74,354,424	33.9	42.1
	Composition	11,621,125	36.2	44.5
	Multi-Entity	36,359,826	17.9	23.2
	Relation	36,606,028	11.5	15.3

Table 7: Key caption statistics of the GBC1M and GBC10M datasets across different types of captions. We use the DFN-5B CLIP model to compute the CLIP scores.

example in Figure 30) to over 10 (third example in Figure 30 and the example in Figure 31). In most cases, this complexity aligns with the visual complexity of the corresponding image.

On the other hand, these examples also reveal limitations arising from the object detection models used. For instance, in the Messe example from Figure 30, the detection model incorrectly identifies a standing priest as a “kneeling figure”. Similarly, in Figure 31, two of the three nodes labeled “trunk” are derived from tree nodes and erroneously associated with the elephant’s trunk or other non-trunk objects on the elephant. These limitations become particularly severe in the Regalia example of Figure 30, where the presence of more specific objects like crowns, scepters, bracelets, and earrings leads to frequent confusion by the object detection model.

Next, we focus on the captions associated with these images. A subset of these captions is presented in Tables 8 and 9. We observe that hallucination is particularly important for detailed captions. These erroneous descriptions can then be inherited by the shorter captions derived from them. We also note there are situations where the model describes an object that actually does not exist in the corresponding region of the image, such as the caption for “scepter 1” in the Regalia example. As we can see from the figure, in the corresponding bounding box, there is no scepter visible but only a crown on a wooden base.

In spite of these inaccuracies in object detection and captioning, the overall graph structure and captions still align well with the images. On top of this, the granularity of our descriptions significantly enhances the descriptive power of our dataset, allowing for a more nuanced understanding of the visual content.

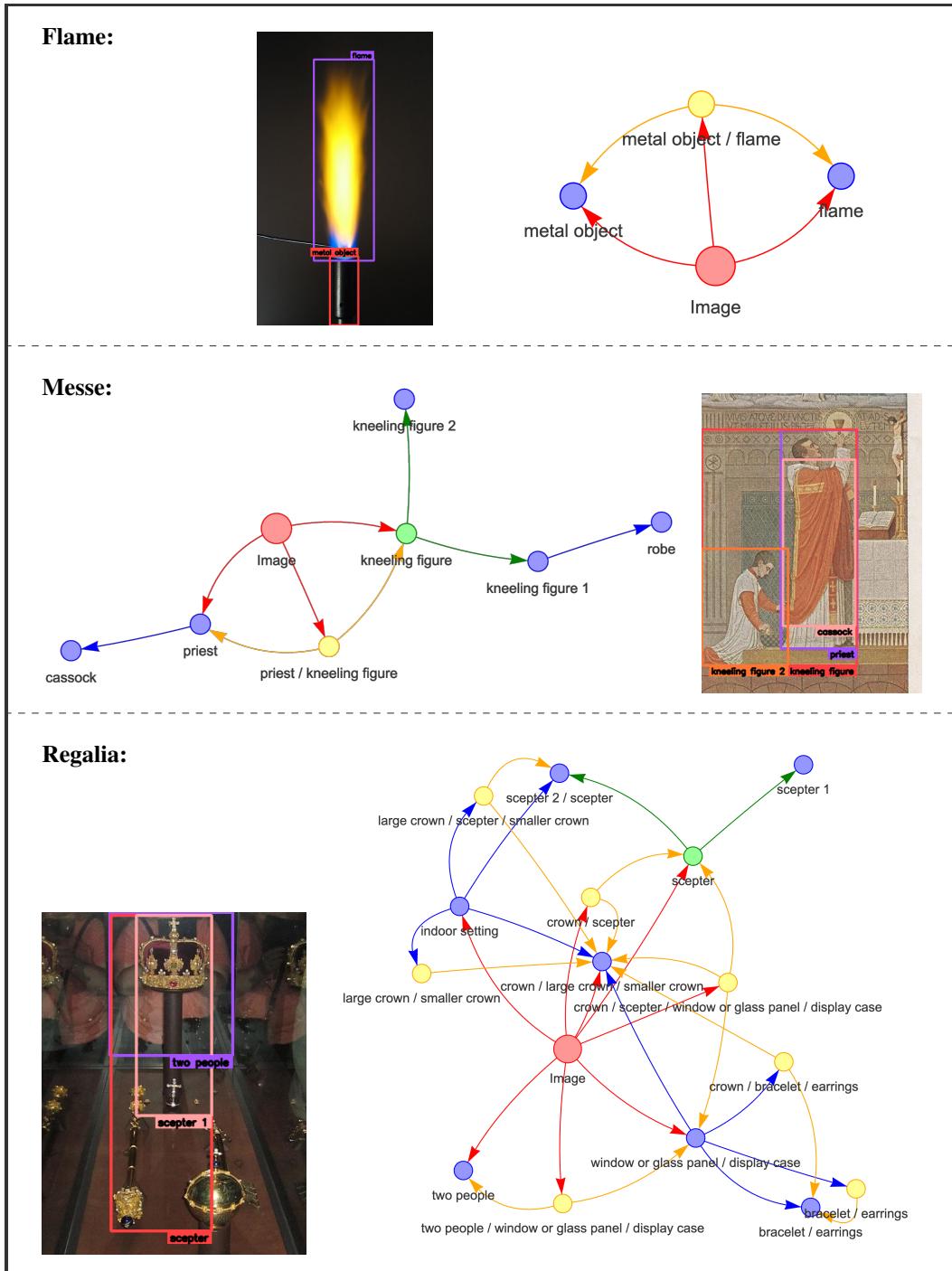


Figure 30: Example images and graphs from the GBC10M dataset. For ease of visualization we only show the bounding boxes of a few nodes.

Elephant:

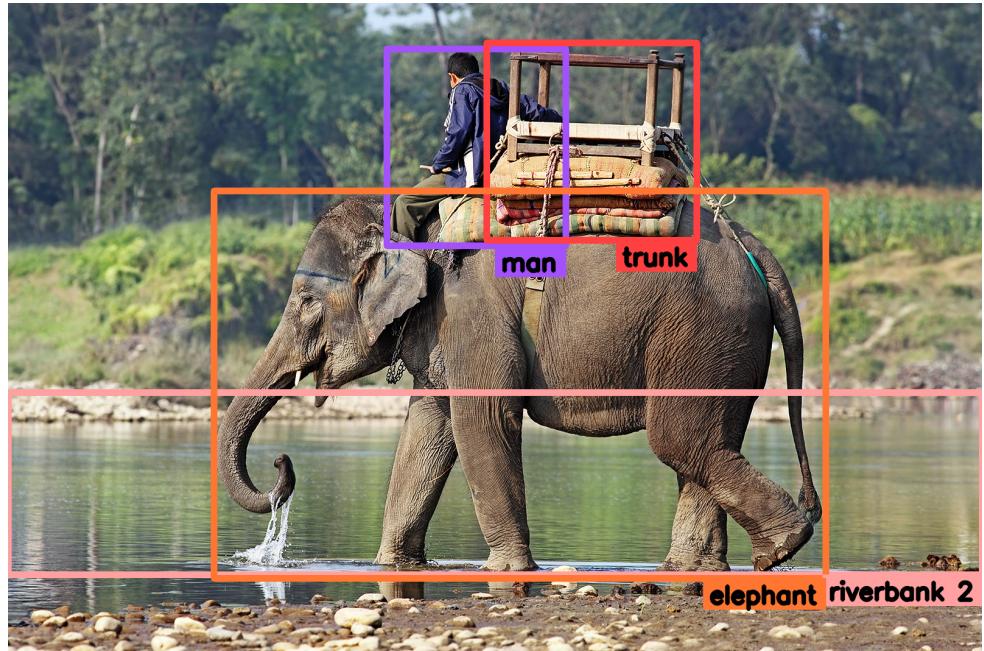
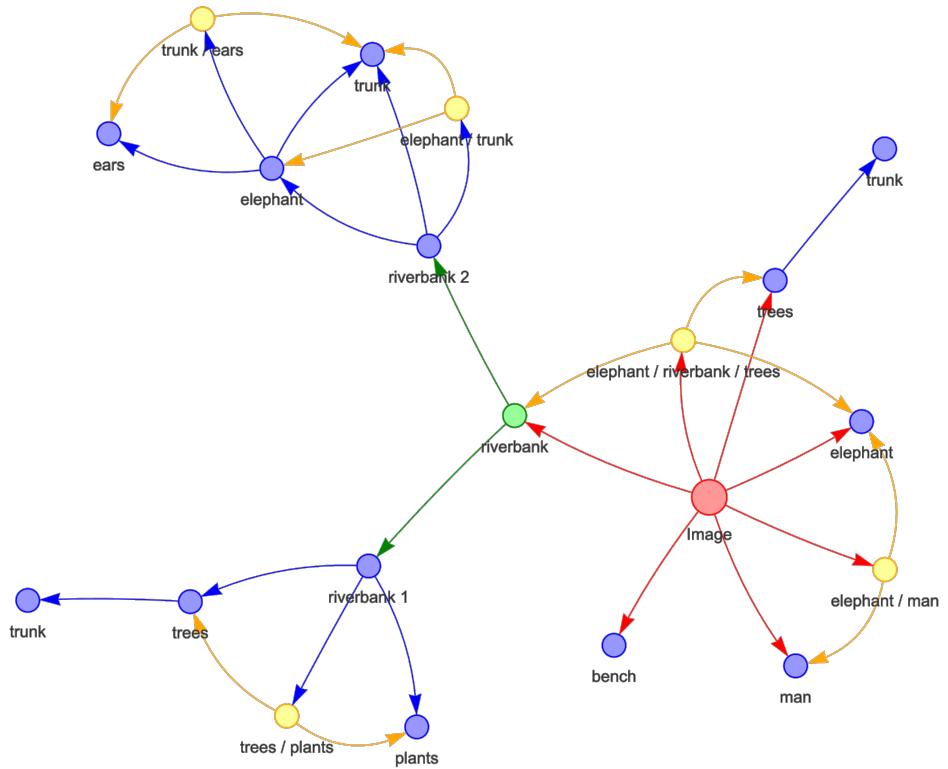


Figure 31: An example image with its associated graph from the GBC10M dataset. For ease of visualization we only show the bounding boxes of a few nodes.

Short Captions	<p>Flame [Figure 30]. A flame with yellow base and <i>blue peak</i> emerges from a metal object against a dark background.</p> <p>Messe [Figure 30]. A priest holds a chalice aloft while another figure kneeling figure kneels on the floor, set against a backdrop of architectural details and ornamentation within what appears to be a religious setting.</p> <p>Regalia [Figure 30]. <i>Two people</i> stand behind a display case containing a crown, scepter with a blue gem, and a golden orb with <i>a red gem</i>, all <i>under natural light from a window or glass panel</i> within an indoor setting.</p> <p>Elephant [Figure 31]. A man rides <i>atop a bench</i> strapped on a elephant drinking from a riverbank, surrounded by trees under a clear sky.</p>
Detailed Captions	<p>Flame [Figure 30]. The image captures a close-up view of a <i>blue flame</i> emanating from a small metal object, which appears to be a lighter or torch. The flame has a vibrant yellow hue at its base, transitioning to <i>a bright blue at its peak</i>. The flame's shape is irregular with wisps extending outward from its core, suggesting it's in motion or has been recently ignited. The metal object has a cylindrical shape with a pointed tip from which the flame emerges. The background is dark, providing a stark contrast that accentuates the flame's colors and form.</p> <p>Messe [Figure 30]. The image portrays a religious scene set within what appears to be a church or chapel. At the center of the composition stands a priest, dressed in traditional religious attire with a red robe and a white cowl. He holds a golden chalice aloft with both hands, suggesting he may be performing a sacrament or ritual. <i>To his right</i>, another figure, possibly another priest or religious figure, kneels on the floor, seemingly in prayer or reverence. The background features ornate architectural details, including arches and intricate patterns on the walls, indicative of Gothic or similar architectural styles. The overall atmosphere suggests a solemn or sacred moment within a religious ceremony or service.</p> <p>Regalia [Figure 30]. The image captures a scene where <i>two individuals</i> are standing behind a display case containing various items. The display case houses a collection of ornate jewelry pieces, including a crown with intricate detailing, a scepter with a blue gem at its top, and a golden orb with <i>a red gem</i>. The individuals are dressed in pink shirts and are positioned behind the display case, which has a reflective surface. The background suggests an indoor setting with <i>a window or glass panel allowing natural light to illuminate the scene</i>.</p> <p>Elephant [Figure 31]. The image captures a serene scene at a riverbank where a man is riding on the back of a large elephant. The elephant, with its majestic gray skin, is partially submerged in the water, drinking from it. The man, dressed in casual attire, <i>sits comfortably on a wooden bench</i> strapped securely on the elephant's back. The bench is adorned with colorful cushions for added comfort during the ride. The backdrop features lush greenery with trees lining the riverbank, adding to the tranquil atmosphere of the scene. The overall setting suggests a peaceful interaction between humans and nature.</p>

Table 8: Corresponding synthetic detailed and short captions for the examples shown in Figures 30 and 31. We highlight the objects described in the children nodes in dark blue and mark some erroneous descriptions in italic.

Entity Captions	<p>Flame—flame [Figure 30]. A bright yellow flame emanates from a blue lighter. The flame is intense and vibrant, with a distinct orange hue at the center. It appears to be burning steadily, casting a warm glow.</p> <p>Messe—priest [Figure 30]. A priest dressed in traditional religious attire, which includes a red robe with gold trim, a white cassock, and a red sash. <i>He holds a cross in his right hand.</i></p> <p>Regalia—scepter 1 [Figure 30]. <i>A gold scepter with a cross atop it. The scepter has intricate designs and patterns throughout its length.</i></p> <p>Elephant—elephant [Figure 31]. A large elephant with a long trunk is seen walking through a body of water. Its skin appears rough and wrinkled, typical of elephants. The elephant has small tusks and large ears that are characteristic of this species.</p>
Relation Captions	<p>Flame—flame/metal object [Figure 30]. The flame is positioned above the metal object.</p> <p>Messe—priest/kneeling figure [Figure 30]. The priest is standing in front of the kneeling figure.</p> <p>Regalia—crown/scepter [Figure 30]. The scepter is positioned <i>next to</i> the crown.</p> <p>Elephant—elephant/riverbank/trees [Figure 31]. The elephant is standing near the riverbank with trees in the background.</p>
Composition Captions	<p>Messe—kneeling figure [Figure 30]. Kneeling figure 1, positioned at the top right, appears to be in a state of prayer or reverence, while kneeling figure 2, located at the bottom left, seems to be in a similar posture, possibly indicating a shared moment of devotion or reflection.</p> <p>Regalia—scepter [Figure 30]. Scepter 2, which is in the bottom left corner, has a golden handle with a blue gemstone at its center, while scepter 1, positioned above scepter 2, features a <i>golden handle</i> with a red gemstone at its center. Both scepters are ornate, with intricate designs and a regal appearance.</p> <p>Elephant—riverbank [Figure 31]. Riverbank 1 is positioned above Riverbank 2, with Riverbank 2 located at the bottom of the composition.</p>
Multi-Entity Captions	<p>Messe—kneeling figure [Figure 30]. Both figures are depicted in a posture commonly associated with prayer or worship, suggesting a religious or spiritual context for their actions.</p> <p>Regalia—scepter [Figure 30]. The gemstones in their handles add a touch of elegance and value to each scepter.</p> <p>Elephant—riverbank [Figure 31]. They are situated near a body of water, which suggests a peaceful, natural setting.</p>

Table 9: Some example relational and region captions for the examples shown in Figures 30 and 31. We highlight the objects described in the children nodes in dark blue and mark some erroneous descriptions in italic.

D Algorithm details

This appendix provides missing details about our architecture and training objective.

D.1 Structure-aware cross attention

We define below mathematically the SACA layer. For this, we denote by \mathcal{N}^C the children of caption C in caption graph \mathfrak{G}^C and write the features of C in the input of our SACA layer as $X^C = [x_1^C, \dots, x_{n^C}^C]$. Recall also that \mathcal{P}^e with $e = (C, C')$ represents the set of token positions in the source caption C that we map the edge label L^e to. Then, the SACA layer maps each feature vector x_i^C to

$$\text{SACA}(x_i^C) = \frac{\sum_{C' \in \mathcal{N}^C} \mathbb{1}_{i \in \mathcal{P}(C, C')} \text{MHA}(x_i^C, X^{C'}, X^{C'})}{\min(1, \sum_{C' \in \mathcal{N}^C} \mathbb{1}_{i \in \mathcal{P}(C, C')})}, \quad (1)$$

where MHA implements the standard multi-head attention mechanism. Note that we average across the results from all the relevant captions that describe this token, as we show in [Figure 4](#).

As a side note, we highlight that with SAHA, information is only propagated from each node to its direct parent within a block. Consequently, the number of blocks must exceed the depth of the \mathfrak{G}^C to ensure that information reaches the root node from all levels of the graph.

D.2 Multi-positive contrastive loss

To pair multiple positive captions to an image, we extend standard contrastive loss [50] into multiple-positive contrastive loss, as also considered in prior studies [14, 16]. Specifically, consider a batch of N images $\{I_i\}_{i=1}^N$, where each image I_i is associated with M_i captions $\{T_{i,j}\}_{j=1}^{M_i}$, we utilize the following loss function to account for multiple positive texts per image:

$$\mathcal{L}_I = -\frac{1}{Z} \sum_{i=1}^N \sum_{j=1}^{M_i} \log \frac{S(I_i, T_{i,j})}{S(I_i, T_{i,j}) + \sum_{k=1, k \neq i}^N \sum_{l=1}^{M_k} S(I_i, T_{k,l})}, \quad (2)$$

where $S(I, T) = \exp(\cos(I, T)/\tau)$, τ is a learnable temperature parameter, and $Z = \sum_{i=1}^N M_i$ is a normalizer. On the other hand, each caption still only has one paired image. Therefore, we use the standard contrastive loss on for text-to-image alignment:

$$\mathcal{L}_T = -\frac{1}{Z} \sum_{i=1}^N \sum_{j=1}^{M_i} \log \frac{S(I_i, T_{i,j})}{\sum_{k=1}^N S(I_k, T_{i,j})}. \quad (3)$$

E Experimental details

This appendix presents further details about our experiments that are omitted in [Section 5](#).

E.1 Data filtering

For the computation of CLIP score, we split any caption that contains more than 77 tokens into individual sentences, compute the score for each of these sentences, and compute the average of these scores. Then, we start by filtering out images whose short synthetic captions have CLIP scores that are lower than the 5% quantile. After this, we consider three filtering strategies depending on the annotation formats.

Long caption. In this case, we just further filter out a portion of original captions and long captions with the lowest CLIP scores (by considering the 5% quantiles from the non-filtered dataset).

GBC. Naive CLIP filtering and tokenizer truncation could break the graph structure as some of the edge labels would not appear in the captions of its source node anymore after these operations. We address this issue by filtering out the captions following the reverse of a topological ordering of the graph, drop a node along with its edges when all its captions and children get filtered, and otherwise, if necessary, add *bag-of-words* captions that collects edge labels from the remaining out edges of a node to ensure all these labels still appear in some captions of this node. Moreover, we split the captions whose length are longer than 77 tokens into concatenations of sentences that fit within this limit, and drop any caption which contains sentences that are of more than 77 tokens.

Hyperparameters	Values
Data augmentation	RRC
Crop size	224×224
Train iterations	45k
Global batch size	4,096
Optimizer	AdamW
Min / max learning rate	{1e-6, 1e-3}
LR. decay schedule type	Cosine
Warmup iterations	1,000
Weight decay rate	0.05
EMA factor	0.9995

Table 10: Hyperparameters for CLIP model training. RRC stands for RandomResizedCrop

Hyperparameters	Values
Data augmentation	RRC
Crop size	512×512
Train iterations	160k
Global batch size	16
Optimizer	AdamW
Peak learning rate	[5e-4, 2e-4, 1e-4, 7e-5, 5e-5]
LR. decay schedule type	Polynomial
Warmup iterations	1,500
Weight decay rate	0.01

Table 11: Training hyperparameters for semantic segmentation experiments on ADE20k. RRC stands for RandomResizedCrop.

	Short	Long	Region	GBC-captions	GBC-concat	GBC-graph
Training time (hr)	22.7	19.7	29.9	38.9	20.3	43.0

Table 12: CLIP model training time for 45,000 iterations with different annotation formats.

Short and Region. We follow the strategy mentioned in GBC, but use selected types of captions. Moreover, bag-of-words captions are not used.

We remark that the filtering procedure is only applied to the training set, and *not* the GBC test set.

E.2 Dynamic batch size

Given the varying sizes of our graph, setting a fixed number of images per batch could result in out-of-memory errors unless we opt for a conservatively small batch size. To overcome this challenge, we implement a dynamic batching strategy for the setups where the number of captions per image is in principle unbounded. This encompasses notably region, GBC-captions, and GBC-graph. With this strategy, we ensure that the number of captions, and, in the case of GBC-graph, the number of edges, that are included in each batch do not exceed a certain limit. In this regard, the batch size that we report in Section 5 is actually just an upper bound on the number of images included in each batch. More specifically, we set this limit based on the number of average captions/edges per image in the filtered dataset. For example, for GBC-captions and GBC-graph we have in average 17.61 captions per graph. We thus set the limit on caption number to $18 \times 64 = 1152$ on each GPU (as mentioned in Appx. E.4, we use 64 GPUs for most of our experiments, which gives a batch size of 64 per GPU).

E.3 Hyperparameters for CLIP training

We used a consistent set of hyperparameters for all model training runs, as detailed in Table 10. The sole exception is training with original CC12M captions, where we used a larger batch size of 8,192 to ensure the model sees a comparable number of texts as during training with both short synthetic and original captions. For this specific setup with the larger batch size, we reported evaluation results from the EMA checkpoint at the end of epoch 15, for it achieving the best performance among the evaluated checkpoints. For GBC-graph, we drop the edges with probability 0.5 so that the model also learns how to match images with short captions.

E.4 Computation cost

We train all CLIP models on A100-80G GPUs. As training with different annotation formats requires varying size of GPU memory, we use different total numbers of GPUs to ensure the same batch size. Specifically, we utilize 16 GPUs for training with *Short* captions, and utilize 64 GPUs for training with all other annotation formats. We list the corresponding time required for training with different annotation formats in Table 12.

Annotation	Hyperparameter			Evaluation results					
	Epoch	Batch size	# Tokens	ImageNet	Flickr	COCO	Share-GPT4V	DCI-concat	GBC test
Short	10	4,096	77	38.8	64.8	38.7	79.1	57.5	87.8
			512	39.0	64.7	39.3	86.7	56.4	89.7
	10	16,384	77	33.2	59.1	34.7	74.0	52.3	83.4
	28			40.0	67.4	38.7	80.6	58.1	88.6
	40			39.0	65.7	37.3	79.9	57.8	88.6
GBC-captions	10	4,096	77	40.8	70.0	43.0	80.4	64.1	91.2
GBC-concat	10	4,096	77	39.0	66.1	40.0	90.5	69.9	94.8
GBC-graph				38.2	67.1	40.8	77.1	61.5	95.9

Table 13: Comparative performance across various benchmarks when we perform CLIP training on short captions with different hyperparameters. For ease of reference, we also include the results from the methods that use GBC annotations. We report the average image and text Recall@1 for all retrieval benchmarks. Specifically, as explained in Section 5.4, we perform retrieval using various annotation formats for GBC test. We thus report here the average of the *highest* image and text Recall@1 scores. The number of iterations is consistently set at 45,000, corresponding to 20 epochs with a batch size of 4,096 and 76 epochs with a batch size of 16,384.

E.5 Evaluation details

Our evaluation uses the validation set of ImageNet-1k [51] and the test sets of Flickr30k [47] and MS-COCO [41]. For SugarCrepe [27] we report the average performance across all variants. As for ShareGPT4V [76], we use a subset of size 15,295 from ShareGPT4V-cap100k. These images were also used for LLaVA training.¹¹ When each image is paired with multiple captions, we only select one of them. The evaluation setups with ADE20K and DCI are more involved, as we explain below.

Evaluation on ADE20K. We evaluate the quality of CLIP models’ image encoder for dense prediction tasks like image segmentation by performing full finetuning on ADE20k [76] dataset. We follow the same setup as described in [62, 63] where we use a ViTDet style feature pyramid network with UperNet [67] head. All models were trained using the MMSegmentation library [11]. We sweep through peak learning rate for all the results reported in the paper and the ranges are listed in Table 11.

Evaluation on DCI. We perform text-to-image and image-to-text evaluations on DCI [61] using either long captions or concatenated captions. The long captions are marked as extra_caption in the released DCI dataset. We filter out samples with empty long captions, resulting in a subset of 7,602 images for evaluation with long captions. Regarding evaluation with concatenated captions, we leverage the full set of 7,805 images. We retain masks containing summary captions (these are masks with bounding boxes larger than 224×224). If the human-annotated caption contains fewer than 77 tokens and is longer than the first summary caption, we use it. Otherwise, we use the first summary caption. For concatenation, we follow the Breadth-First Search (BFS) order based on the provided tree structure between the masks.

F Additional results and experiments

In this appendix, we present additional ablations that we have performed but were not presented in the main paper due to space constraints.

F.1 Matching compute resource for training with short captions

All our models presented in Section 5 used 8 nodes for training, except for the models trained on short captions, which only used 2 nodes. This raises the question of whether the performance gap could be bridged by providing more computational resources to this setup. To address this, we specifically considered two modifications that would naturally necessitate using more nodes for training with short captions: (*i*) extending the context length to 512, as done for training with Long and GBC-concat captions, and (*ii*) using a batch size that is four times larger, i.e., a batch size of

¹¹<https://huggingface.co/datasets/liuhaojian/LLaVA-Pretrain> Accessed: 2024-05-14

Annotation	ImageNet	Flickr-1k	MSCOCO-5k	SugarCrepe	Average Drop
Short	38.8 → 35.2	64.8 → 61.0	38.7 → 36.7	76.0 → 74.4	-2.75
Long	39.6 → 30.5	65.8 → 56.8	40.1 → 33.5	77.0 → 74.0	-6.93
GBC-captions	40.8 → 31.9	70.0 → 58.3	43.0 → 33.2	76.7 → 73.3	-8.45

Table 14: Performance degradation across different annotation types when switching from multi-positive contrastive loss to standard contrastive loss with randomly sampled positive captions. For Flickr-1k and MSCOCO-5k we report the average image and text Recall@1.

16,384 instead of 4,096. All other hyperparameters remained unchanged. We then trained the models on 8 nodes, each with 8 GPUs, as in the other setups, which resulted in training times of 18 and 48 hours for the two modifications respectively. The evaluation results are presented in [Table 13](#).

Training with extended context length. Provided that the models are only trained with short captions, we do not expect any tangible benefit from extending the context length. Yet, surprisingly, while this is indeed the case for classic benchmarks such as ImageNet, Flickr, and COCO, we do observe a significant performance boost on ShareGPT4V retrieval, suggesting that the longer context length is still beneficial for retrieval with long caption even though the model is not explicitly trained for this task. On the other hand, we do not observe any benefit when evaluated using concatenated caption from DCI. Finally, we also get a slight performance improvement on GBC test, and it turns out this improved performance is achieved by performing retrieval using the long caption. This is in line with the performance gain that we observe for the ShareGPT4V benchmark.

Training with larger batch size. More interestingly, CLIP is known to perform better when trained with a large batch size, so we might be able to bridge the performance gap by simply including more images and captions in each batch. To enable a fair comparison for this setup, we report evaluation results from three checkpoints at varying training stages in [Table 13](#). These checkpoints are chosen to align with key training milestones.

- **Number of images seen:** We consider the EMA checkpoint at the end of epoch 10 to align the number of images seen.
- **Number of iterations:** We include the EMA checkpoint at the end of epoch 40 to compare models at a fixed number of training iterations.
- **Best performing checkpoint:** Additionally, we report results for the EMA checkpoint at the end of epoch 28, as it gives the best performance among all evaluated checkpoints (see [Figure 32](#)).

As we can see from the table, while the use of a larger batch size indeed leads to better performance on ImageNet and Flickr, the results still lag behind those achieved with GBC-captions. This discrepancy underscores the **importance of including multiple captions per image to enhance performance**.

F.2 The importance of multi-positive contrastive loss

We next look into the influence of the objective function when an image is paired with multiple captions. Instead of employing the multi-positive contrastive loss introduced in [Appendix D.2](#), we can use a standard contrastive loss with a single randomly sampled caption paired with each image. [Table 14](#) presents the evaluation results for both the models trained with the original objective (left side of the arrow), and this new, sampled, objective (right side of the arrow).

The table clearly shows a performance decline across all the considered annotation formats and benchmarks when sampling is applied, as also observed by Doveh et al. [14] and Fan et al. [16]. The performance drop is particularly important when the captions vary significantly (e.g., long versus short captions, or image versus region captions), and when many captions are involved. More surprisingly, this alternative loss does not lead to improvement but rather to performance degradation when we increase the number of captions paired with each image. We conjecture this is because the additional captions that we consider here are less relevant for these specific benchmarks, leading to a worse performance when they are forced to be treated as positive in the sampled objective.

Overall, these results confirm the **importance of our multi-positive contrastive loss in leveraging the presence of multiple captions for an image**.

Annotation	Flickr-1k		MSCOCO-5k		DCI-concat		ImageNet	SugarCrepe	ADE20K
	T2I	I2T	T2I	I2T	T2I	I2T			
Short	56.3	73.2	30.7	46.7	57.5	57.5	38.8	76.0	42.0
Region	58.3	<u>76.6</u>	31.5	49.1	61.8	61.5	38.5	75.6	43.5
GBC-Relation	60.4	76.5	34.8	52.5	62.0	61.4	41.5	<u>76.4</u>	<u>44.5</u>
GBC-captions	60.6	79.3	<u>34.1</u>	<u>51.9</u>	64.1	63.4	40.8	76.7	45.0

Table 15: Comparative performance on various existing benchmarks when trained using different subsets of GBC-captions.

Annotation	Short		GBC-graph				Star graph		Line graph	
			Groundtruth		Last token					
	T2I	I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I	I2T
GBC-graph	84.3	85.2	95.7	96.1	91.7	92.4	<u>95.1</u>	<u>95.4</u>	<u>94.8</u>	<u>94.9</u>

Table 16: Image and text retrieval performance on the GBC test set when the model is trained using GBC-graph and evaluated across various underlying graph structures.

E.3 Impact of caption type on CLIP training

To further highlight the value of relation and composition captions from GBC, we trained a CLIP model using only these captions alongside short image captions. As shown in the third row of Table 15, these captions, despite being more than twice as scarce as region captions, not only provided a larger performance gain than using only region captions, but sometimes even enabled the model to achieve comparable or better performance than using all captions combined. This underscores **the significant benefit of the relational captions from GBC datasets**.

Looking closely, we note that region captions primarily benefit retrieval and dense prediction tasks, while relation and composition captions improve performance across the board. While using all captions remains the best approach for most benchmarks, the marginal improvement from region captions hints at the potential for more efficient training with these captions through alternative training objectives.

F.4 Impact of the underlying graph on retrieval

In this part, we investigate how much GBC-graph relies on the underlying graph structure for retrieval. For this, we probe the performance of our model when the graph is modified either in the mapped tokens or in the connectivity patterns. In terms of the mapped tokens, we consider

- Last token: For any edge from a caption C to another caption C' , we mapped the information of C' to the last token before the summary token in C .¹²
- Random token: For each edge, we randomly map the information to one token in the source caption.

As for the connectivity pattern, we investigate

- Star graph: All the captions are mapped to the short image synthetic caption.
- Line graph: We map each caption to its next caption in a list (ordered as in GBC-concat following the BFS order), with the short image synthetic caption being the first in the list.

The results are shown in Table 16. Since random-token mapping consistently leads to better result than last-token mapping, we only report results for this in the case of star graph and line graph. First of all, we observe that no matter which graph is given, we always achieve better performance than retrieval with only short caption, suggesting that the model is always able to exploit the additional captions to some extent. Furthermore, employing random-token mapping, whether with the groundtruth graph topology or the star graph, yields performance that closely matches that of using the groundtruth graph with correct mapping (interestingly, when using star graph the performance is also very close to that obtained with GBC-concat, see Table 3). This suggests that the specific retrieval task we are

¹²We also experimented with mapping the information to the summary token but this completely destroys the performance.

Annotation	ImageNet	Flickr-1k	MSCOCO-5k	SugarCrepe	ShareGPT4V-15k	GBC test
CC12M	37.1	50.5	27.9	39.4	47.2	49.5
Short	37.5	62.0	36.1	74.5	77.3	86.9
Long	38.5	64.5	38.4	75.8	93.5	95.5
Region	<u>40.3</u>	<u>68.6</u>	40.8	76.1	78.9	91.8
GBC-captions	41.3	70.6	43.1	76.4	80.1	91.9
GBC-concat	38.2	63.4	37.1	74.9	<u>89.8</u>	96.1
GBC-graph	39.9	68.5	<u>40.9</u>	74.6	77.5	96.2

Table 17: Comparative performance on various existing benchmarks when trained using different annotation schemes. Unlike the other tables that report performance for EMA checkpoints, this table presents the performance at the final non-EMA checkpoints obtained from the end of training.

examining is not highly dependent on the provided mapping and topology. However, we do believe the mapping and topology could play a significant role in other tasks or when more fine-grained distinctions between images are necessary.

E.5 Evaluating at non-EMA checkpoints

For the sake of completeness, we also perform evaluation on the non-EMA checkpoints, with results shown in [Table 17](#) and [Figure 33](#). Comparing [Figure 32](#) with [Figure 33](#), we see that while EMA checkpoints may experience a drop in performance during later training stages, non-EMA checkpoints typically exhibit best performance at the final training checkpoint. Consequently, our evaluations in [Table 17](#) are based on these last checkpoints. From the evaluation results, we observe a similar trend in the performance comparison of annotation formats with non-EMA checkpoints as with EMA ones, confirming the validity of our previous claims. Finally, we also note that the use of larger batch size when training with short captions is only beneficial when we consider EMA checkpoints.

G Image attributions

All the images that we show in this paper come from Wikimedia Commons. We provide in [Table 18](#) the exact source urls and license for each of the images. The urls to the CC BY-SA 3.0 and GFDL 1.2 licenses are respectively <https://creativecommons.org/licenses/by-sa/3.0/> and <https://www.gnu.org/licenses/old-licenses/fdl-1.2.txt>.

Image	Source URL	License
Figure 1	https://commons.wikimedia.org/wiki/File:Tartu_raudteejaama_veetorn,_2010.JPG	CC BY-SA 3.0
Figure 2	https://commons.wikimedia.org/wiki/File:Eiffel_Tower_from_north_Avenue_de_New_York,_Aug_2010.jpg	CC BY-SA 3.0
Figure 30		
Flame	https://commons.wikimedia.org/wiki/File:Flametest--Na.swn.jpg	CC BY-SA 3.0
Messe	https://commons.wikimedia.org/wiki/File:Messe_mit_Wandlungskerze_Beuron.jpg	Public domain
Regalia	https://commons.wikimedia.org/wiki/File:Crown,_sceptre,_orb_%26_key_of_the_King_of_Sweden_2014.jpg	Public domain
Figure 31	https://commons.wikimedia.org/wiki/File:Indian-Elephant-444.jpg	GFDL 1.2

Table 18: Source URLs and licenses of the images shown in this paper.

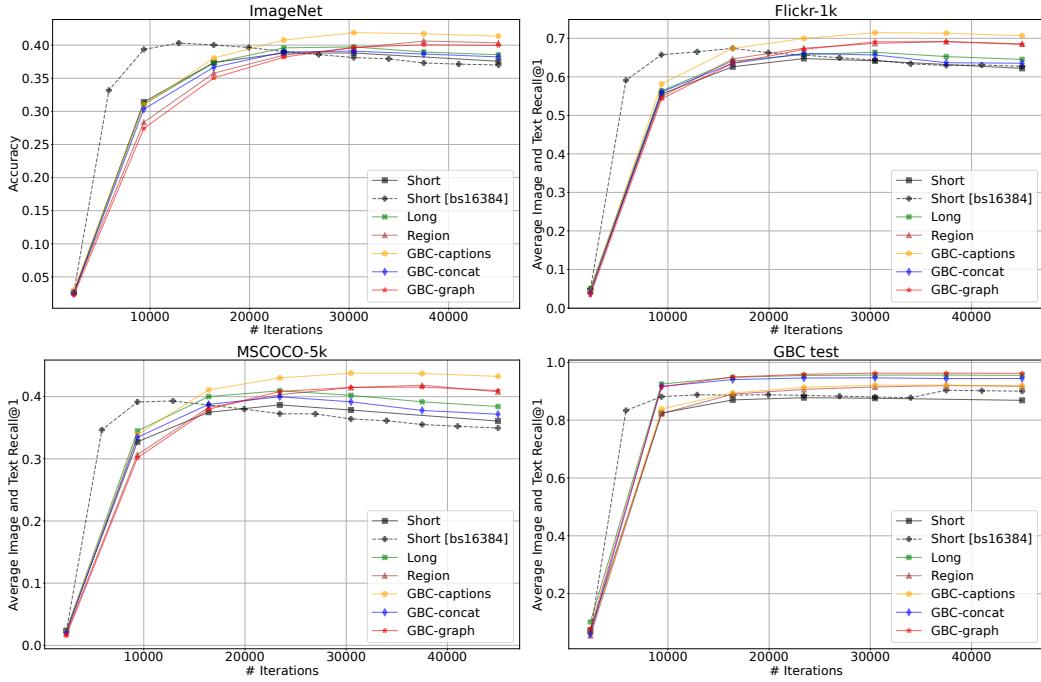


Figure 32: Benchmark performances on ImageNet, Flickr-1k, MSCOCO-5k, and GBC test for EMA checkpoints of models trained with different annotations / hyperparameters. For GBC test we use different formats for retrieval at test time and average the highest scores that are respectively obtained for text-to-image and image-to-text retrievals.

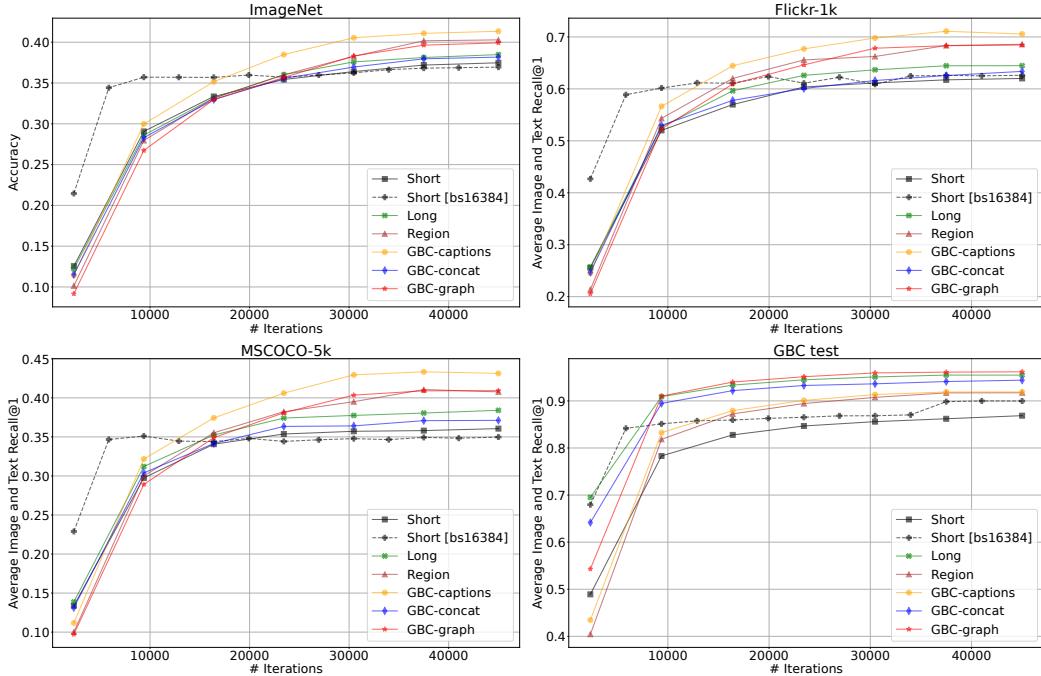


Figure 33: Benchmark performances on ImageNet, Flickr-1k, MSCOCO-5k, and GBC test for non-EMA checkpoints of models trained with different annotations / hyperparameters. For GBC test we use different formats for retrieval at test time and average the highest scores that are respectively obtained for text-to-image and image-to-text retrievals.