Vision-Free Retrieval: Rethinking Multimodal Search with Textual Scene Descriptions

Ioanna Ntinou[1*] **Alexandros Xenos**[1*] **Yassine Ouali**[2]
Adrian Bulat[2,3] **Georgios Tzimiropoulos**[1,2]
[1]Queen Mary University of London, UK
[2]Samsung AI Centre, Cambridge, UK
[3]Technical University of Iași, Romania
{i.ntinou,a.xenos,g.tzimiropoulos}@qmul.ac.uk
y.ouali@samsung.com adrian@adrianbulat.com

Abstract

Contrastively-trained Vision-Language Models (VLMs), such as CLIP, have become the standard approach for learning discriminative vision-language representations. However, these models often exhibit shallow language understanding, manifesting bag-of-words behaviour. These limitations are reinforced by their dual-encoder design, which induces a *modality gap*. Additionally, the reliance on vast web-collected data corpora for training makes the process computationally expensive and introduces significant privacy concerns. To address these limitations, in this work, we challenge the necessity of vision encoders for retrieval tasks by introducing a *vision-free, single-encoder* retrieval pipeline. Departing from the traditional text-to-image retrieval paradigm, we migrate to a text-to-text paradigm with the assistance of VLLM-generated structured image descriptions. We demonstrate that this paradigm shift has significant advantages, including a substantial reduction of the modality gap, improved compositionality, and better performance on short and long caption queries, all attainable with only a few hours of calibration on two GPUs. Additionally, substituting raw images with textual descriptions introduces a more privacy-friendly alternative for retrieval. To further assess generalisation and address some of the shortcomings of prior compositionality benchmarks, we release two benchmarks derived from Flickr30k and COCO, containing diverse compositional queries made of short captions, which we coin subFlickr and subCOCO. Our vision-free retriever matches and often surpasses traditional multimodal models. Importantly, our approach achieves state-of-the-art zero-shot performance on multiple retrieval and compositionality benchmarks, with models as small as 0.3B parameters. Code is available at LexiCLIP.

1 Introduction

Contrastively-trained Vision-Language Models (VLMs) (Radford et al., 2021a) have rapidly become a cornerstone for learning powerful, discriminative vision-language representations. Their success is underscored by remarkable zero-shot transfer abilities across a wide array of tasks (Jia et al., 2021; Li et al., 2022a,c; Radford et al., 2021a; Zhai et al., 2023a). These capabilities are largely attributed to their training on vast quantities of image-text pairs using a simple contrastive objective. However, this scale comes at a significant cost: training such models is computationally expensive, and the reliance on web-collected data introduces notable privacy challenges. Moreover, the prevalent dual-tower architecture that encodes images and text separately induces a *modality gap*, an effect which hinders the model's fairness and compositional abilities (Liang et al., 2022). The latter also stems from the limited understanding of language structure of the CLIP's text encoder, whose representations tend to ignore word order and syntactic relations – effectively treating the caption as an unordered bag-of-words (Yuksekgonul et al., 2023). Due to the above limitations, contrastively trained VLMs (e.g., CLIP, SigLIP (Zhai et al., 2023b)) often exhibit poor compositional generalization and shallow language understanding, yet they achieve strong performance on popular text-image retrieval benchmarks like Flickr30k (Young et al., 2014) and COCO (Lin et al., 2014).

A growing body of research has quantified the limitations of current vision-language models and begun to address them through several approaches. These include constructing or mining hard negative examples (Yuksekgonul et al., 2023), employing shared or partially shared backbones (Likhosherstov et al., 2021), and adding cross-modal fusion modules or adapters to learn fine-grained alignment between image regions and words (Li et al.,

*Equal contribution.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 14046–14062
November 4-9, 2025

2023a). However, the use of hard negatives has been shown to potentially rely on shortcuts or spurious patterns (Hsieh et al., 2023) while approaches based on cross-modal fusion are impractical due to the need for a separate inference pass for each image with every new query.

Departing from previous works, we aim to (1) remove the modality gap by design, (2) reduce the *bag-of-words* behavior, and (3) alleviate the privacy concerns pertaining to the training data; all under a framework that requires limited training and data. Finally, (4) we seek to introduce a new text-image retrieval benchmark that cannot be easily solved by VLMs exhibiting bag-of-words behaviour.

To this end, we propose a paradigm shift by converting images entirely into carefully crafted textual descriptions, thereby enabling language models to reason about visual content purely through text. This strategy offers significant advantages, including leveraging high-capacity pretrained text encoders, significantly narrowing the modality gap through a fully shared encoder, and substantial mitigation of privacy risks as the model avoids direct handling of sensitive image data. However, this approach faces a fundamental challenge: faithfully representing rich visual information solely with text remains an open and under-explored problem. To address this, we first introduce a robust, principled, and carefully designed pipeline for image-to-text conversion that captures the richness of visual information. We then show that the resulting text-based image representation can produce strong, vision-free, zero-shot text-image retrieval models. To further boost the accuracy of the model, we utilize the textual corpus generated by applying the proposed pipeline to 1.5M images from the OpenImages dataset (Kuznetsova et al., 2020a) to fine-tune the model, better aligning it to the input distribution.

Finally, recognizing the aforementioned limitations of existing text-image retrieval benchmarks, we introduce two new datasets, subFlickr and subCOCO (derived from the Flickr and COCO datasets, respectively), specifically designed to assess performance on short compositional tags, an area poorly represented in previous test suites, where, as we show, the standard VLMs appear to struggle. In summary, our main contributions are:

- We introduce LexiCLIP , a novel text-only Vision-Language framework that converts images into textual descriptions, enabling language models to process visual content. This inherently removes the modality gap, reduces the "bag-of-words" effect, and alleviates privacy concerns, all while requiring limited to no training and data.
- A new principled and carefully designed pipeline for accurately converting rich visual information into text, with ample validation on a multitude of benchmarks.
- We introduce two new datasets, subFlickr and subCOCO, specifically curated to evaluate VLMs on short compositional queries, an area previously underrepresented in benchmarks.
- Using solely textual inputs, and no task-specialised data, we set a new state-of-the-art result on image-text composionality and image retrieval with long captions.

2 Related work

2.1 Text-only training

Recent methods propose to drop images from the training pipelines in an attempt to alleviate the modality gap. Knight (Wang et al., 2023) introduces a text-only captioning pipeline where image- or video-derived captions are used to build a text corpus for training a decoder with autoregressive loss. At inference, the k-nearest captions are retrieved and used as embeddings to a decoder. DeCap (Li et al., 2023b) trains a lightweight language decoder purely on a large corpus of text embeddings generated from CLIP's (Radford et al., 2021b) text encoder. CLOSE (Gu et al., 2023) observes the low image-text cosine similarity and proposes a hyper-parameter-scaled noise injection method. IFCap (Lee et al., 2024) similarly injects noise into text embeddings to imitate image embeddings, improving the retrieval of semantically aligned captions. CLIPPO (Tschannen et al., 2023) shifts from the dual encoder paradigm by jointly processing images and text (where alt-text is rendered as an image) using a purely pixel-based model. The resulting image pair is encoded with a shared vision encoder and trained via contrastive loss. Different from previous works, our pipeline is more privacy-friendly, excels at long-form text retrieval—overcoming the fixed sequence-length constraints of pixel-based encoders—and enables extensive linguistic knowledge transfer via strong pre-trained language models.

2.2 Datasets for text-to-image retrieval

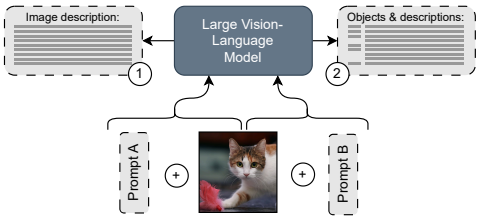
Text-image retrieval datasets typically fall into two categories: long-caption and short-tag datasets. The conventional approach, common in many prior works, uses long captions. Key benchmarks are Flickr30k (Young et al., 2014) (31K images) and MS COCO (Lin et al., 2014) (330K images), each offering five crowdsourced full-sentence descriptions per image, averaging 10–13 tokens and describing the full scene. NoCaps (Agrawal et al., 2019) expands this to 15K OpenImages-derived images with 166K human-written captions covering a broader range of categories. These datasets are characterized by rich, syntactic, sentence-level annotations describing the entire image and often averaging over 10 words per caption.

The newer and second line of research uses short tag datasets with keyword-style annotations. Tag2test (Huang et al., 2023b), RAM (Zhang et al., 2023) and RAM++ (Huang et al., 2023a) automatically extract a set of tags from existing captions or metadata, yielding large-scale image–tag pair corpora without manual labeling. Each image is labeled with a collection of salient keywords (e.g. "dog", "couch", "table" for a living-room scene) rather than a full sentence, enumerating the contents without syntax. Such tag-based datasets are often an order of magnitude larger, on the order of millions of images drawn from web data and covering thousands of distinct tag categories, e.g. 3,400 categories in Tag2Text handles and 4583 in RAM. These tags lack sentence structure but reflect real-world search behavior more closely, where users input short, compositional queries.

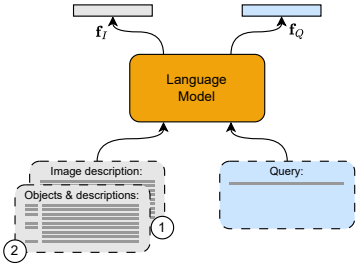
In this work, we introduce *subCOCO* and *subFlickr*, which are positioned between these two extremes regarding annotation granularity. Built from Flickr30K and COCO, they use sub-sentential phrases, shorter than full captions but semantically richer than flat tags. An overview of these datasets and others is presented in Tab. 1.

3 Vision-Free Constrastive Learning

Contrastive language-image pretraining has emerged as a highly effective method for developing vision-language models, leveraging vast amounts of web-collected image-text pairs. Using this data, the prevalent technique trains two independent encoders, one for each modality, using a contrastive loss that aims to map each input to a joint embedding space (Radford et al., 2021b).



(a) The proposed Image to Text conversion pipeline: An image is converted into an equivalent textual description in two steps: one prompted using "Prompt A" to describe the image in detail, and another, using "Prompt B" aimed at capturing each object and corresponding attributes.



(b) LexiCLIP - the proposed Vision-Free image-text retrieval model.

Figure 1: The proposed Vision-Free Retrieval Pipeline.

Despite its success, this approach suffers from a series of drawbacks: (1) Training and finetuning such models is computationally expensive, (2) Using web collected images may result in privacy infringements and (3) The models suffer from a *bag-of-words* behaviour (Yuksekgonul et al., 2022), largely a consequence of the modality gap induced by the two separate towers (Liang et al., 2022). As a solution to these issues, we introduce LexiCLIP , a novel vision-free text-to-text contrastive learning framework that leverages pretrained language models for effective image-text retrieval within a shared single-tower architecture. Our key idea is to bring the images into the language domain via dense captioning, leveraging thereafter the world knowledge of discriminatively pretrained LLMs. The conversion to textual descriptions is also privacy-friendly, as most identity-related information (i.e., faces, private rooms, etc.) is removed. Without any further training, in a zero-shot manner, our solution showcases strong image-text retrieval abilities, which we further boost using a light finetuning on a small dataset.

Representing images using text: *A picture's worth: how many words, and which ones truly matter?* A wide disparity exists in the literature regarding the structure, content, and particularly the length of these textual representations, span-

Table 1: Overview of text-image retrieval evaluation datasets used in this study. * - denotes estimated statistics.

Dataset	# Images	# Queries	Avg. Query Length	Captions per Image	Query Type
Flickr30K (Young et al., 2014)	1 K	5 K	13.4	5	Full Sentence
MS-COCO (Lin et al., 2014)	5 K	25 K	10.4	5	Full Sentence
NOCAPs (Agrawal et al., 2019)	10 K	106 K	9–11	10–11*	Full Sentence
Conceptual Captions (Sharma et al., 2018)	22.5 K	22.5 K	9.7	1	Sentence Caption
Winoground (Thrush et al., 2022)	400	400	8.8	1	Compositional (Paired)
SugarCrepe (Hsieh et al., 2023)	7.5 K	7.5 K	10*	1	Compositional (Paired)
SugarCrepe++ (Dumpala et al., 2024)	4.8 K	9.5 K	10*	2	Compositional (Paired)
ADE20K (Zhou et al., 2017)	2 K	N/A	9.9	N/A	Tag-based Queries
OpenImages (Kuznetsova et al., 2020b)	125 K	N/A	1–5	8	Tag-based Queries
subFlickr	935	280	4.5	6.0	Compositional Caption
subCOCO	4030	256	3.47	4.1	Compositional Caption

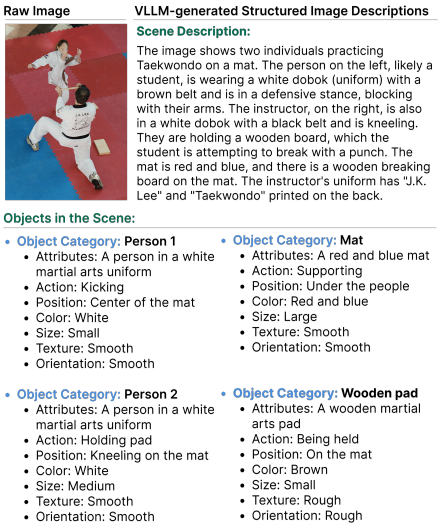


Figure 2: Example of our textual image representations.

ning from brief annotations of one word (Gal et al., 2022) to extensive descriptions comprising over a thousand words (Collell Talleda and Moens, 2016). In the absence of a prevailing standard, we undertake an analysis of this question within the specific domain of image retrieval. We posit that a series of desirable properties characterize an effective textual descriptor: It should capture the (1) high-level scene details, identify (2) all salient objects, and articulate (3) actions, interactions, and object placement. Furthermore, its structure must be (4) coherent and (5) organized.

Thanks to the rapid progress in the area of large vision language modeling (Chen et al., 2024; Bai et al., 2025b), many of these requirements can be readily addressed by providing a Vision LLM (VLLM) with an image and an appropriate prompt that tasks the model to generate highly detailed image descriptions. Generally, we find that longer descriptions are desirable, as they better capture the richness of information present in an image.

Nonetheless, this alone is insufficient. Due to significant variations in image object density, the model may occasionally fail to identify certain objects or inaccurately report or miss their attributes. To mitigate this issue, we propose a supplementary step utilizing the same VLLM: generating a structured list that enumerates the objects present in the image, accompanied by a brief description for each. Our findings indicate that a JSON format is optimal for structuring this output.

We consider a wide variety of prompts and models. Out of the models tested, the best results were obtained using InternVL-2.5-8B-MPO (Wang et al., 2025) VLLM. We provide the final prompts used alongside a set of randomly sampled examples in the supplementary material. The overall image-to-text conversion process is shown in Fig. 1a, while Fig. 2 shows an example of the resulting representation.

Zero-shot vision-free image-text retrieval: Given an image description $\mathcal{T}$ and a query $\mathcal{Q}$, the corresponding image and query embeddings, $\mathbf{f}_T$ and $\mathbf{f}_Q$ are obtained by passing each tokenized input through the shared language model $\Phi(\Theta, .)$. Following best practices for zero-shot evaluations, we also concatenate a handcrafted prompt p_Q for the query. The process is depicted in Fig. 1b.

Vision-Free text-to-text finetuning: While single-tower language models exhibit robust zero-shot performance, finetuning presents opportunities for enhancement, especially in the case of smaller variants. Such improvements are motivated by two key factors: firstly, smaller models tend to have more constrained generalization, and secondly, the statistical distribution of image descriptions can diverge from the data typically used in the pretraining phase of these Language Models.

To this end, we construct a small alignment set by converting 1.5M images from the OpenImages dataset to textual representations using the aforementioned process. Since these images are not paired with short (query) captions, we synthetically generate them using VLLMs. In particular, we run BLIP-2 to generate concise captions of around 10 tokens, and we prompt InternVL-2.5-8B-MPO (Wang et al., 2025) to extract a pool of six 2-3 word

compositional captions.

For fine-tuning, we adopt a two-stage contrastive training paradigm. In the first stage, our model is trained to align longer BLIP-2 captions with images. As a second stage, the model is adapted using a mixture of short compositional captions and BLIP-2 captions, refining its understanding of fine-grained details. More details on the prompt and the training can be found in the appendix.

4 subFlickr and subCOCO

Most widely used image-text retrieval datasets, such as MS COCO and Flickr30K, feature full-sentence captions that average 10–13 words in length. These captions tend to be lengthy, formal, and grammatically complete. Moreover, benchmarks like MS COCO and Flickr30K mostly rely on broad, scene-level descriptions, which do not reflect how people actually search for images. For example, a caption in Flickr30K might be "A man in a green t-shirt and long tan apron hacks apart the carcass of a cow while another man hoses away the blood." but in real-world queries that could possibly be much shorter and more fragmented like "man in a green t-shirt," or "hack the carcass up." These natural queries are typically informal, ungrammatical, and omit function words, resembling spoken language rather than written prose. On the other hand, keyword-style annotations (Huang et al., 2023b; Zhang et al., 2023; Huang et al., 2023a) lack in semantic richness, compositionality, and alignment with real user queries.

To address this limitation, we introduce two retrieval benchmarks, *subFlickr* and *subCOCO*, derived from the test sets of the Flickr30k and MS COCO datasets, respectively. To generate a set of concise queries, we decompose the existing ground-truth captions into shorter, meaningful subcaptions. For this, we employ a pretrained constituency parser (Honnibal et al., 2020) to decompose each caption to its constituent nodes, i.e. complete captions, sentences, sub-phrases, and individual lexical items (nouns, verbs, etc.). From this structure, we extract recurring subphrases that are likely to be visually grounded. We then manually curate a set of queries, choosing compositional expressions, such as "a person with a white shirt. These serve as retrieval queries in our benchmark. To match each image with the relevant queries, we first compute text-to-text similarity scores between the ground truth subcaptions and the curated queries using a text encoder (Bge-large-en-v1.5 (Xiao et al., 2023)). As a second step, we use two VLMs: Qwen2-VL-7B (Bai et al., 2025b) and InternVL2.5-8B (Chen et al., 2024) to verify whether each query is visually present in the image. A query is assigned to an image if both models agree, providing a more reliable and grounded labeling. As a final step, we visually inspect 20% of our dataset. More details about our dataset can be found in the appendix.

5 Experiments

We compare our approach with the current state-of-the-art in four tasks of interest: (1) compositional retrieval using short captions on our newly introduced benchmarks, (2) zero-shot text-to-image retrieval, (3) image-to-text long captions retrieval, and (4) compositional understanding. In each case, we compare our method against a broad set of state-of-the-art two-tower (independent) VLMs. For specialized tasks such as compositional understanding, we include targeted baselines where applicable.

We refer to our models evaluated in a zero-shot setting using structured text-only representations as **LexiCLIP –ZS**, and to their fine-tuned variants as **LexiCLIP –FT**. We focus our evaluation on a 0.3B parameter encoder backbone configuration, denoted as **LexiCLIP (0.3B)**. Specifically, we adopt a robust instruction-tuned, contrastively pretrained text model: BGE-large-en-v1.5 (Xiao et al., 2023), which achieves state-of-the-art performance on the MTE Benchmark (Muennighoff et al., 2022) within its model size.

5.1 Implementation details

We train our model in two stages. For both stages, the model is trained for three epochs using the FlagEmbedding library (Xiao et al., 2023), with a cosine-annealed learning rate schedule and a 5% warm-up phase. The model is trained on two A100 GPUs with an effective batch size of 2,048, a peak learning rate of 1×10^{-4}. In the first stage, the training is done only using BLIP-2 captions. In the second stage, we fine-tune the model using a mixture of BLIP-2 and compositional captions, where each batch of 2,048 samples consists of 200 concise BLIP-2 captions and the remainder compositional ones. The final model is obtained by averaging the checkpoints from the two stages, weighted 0.4 for the first and 0.6 for the second stage.

5.2 Short caption retrieval

We benchmark our datasets, *subFlickr* and *subCOCO*, over a series of models for short-caption text-to-image retrieval. The goal is to recognise all relevant images given a query, which is a short caption. Each query is associated with a binary relevance label over the test set. For evaluation, we opt for mean Average Precision (mAP) and F1-score.

Tab. 2 reports results for both zero-shot and fine-tuned retrieval pipelines. In addition to a range of general two-tower models, we evaluate RAM (Zhang et al., 2023) and RAM++ (Huang et al., 2023a) tagging models. We evaluate our method zero-shot, but also after fine-tuning. The latter, denoted as **LexiCLIP (0.3B)–FT** in our table, and as proposed in this work, is a fine-tuned version of BGE-large-en-v1.5 that is trained exclusively on textual inputs—namely, BLIP-2 captions and corresponding descriptive annotations - without using any image features during training (i.e. the text-only training introduced in Section 3).

We note that both RAM (Zhang et al., 2023) and RAM++ (Huang et al., 2023a) achieve high zero-shot performance on our benchmarks. Their success is attributed to two key factors: first, these models are trained on a large-scale dataset of 14 million image-tag pairs, and second, they leverage explicit tag supervision, which allows them to learn fine-grained object-attribute associations.

5.3 Zero-shot image-text retrieval

We evaluate our approach on the standard Flickr30K (Young et al., 2014) and MSCOCO (Lin et al., 2014) benchmarks. As Tab. 3 shows, without any finetuning, our 300 M-parameter LexiCLIP achieves 69.5 R@1 on Flickr30K and 41.7 R@1 on COCO, comparing favourably with similarly sized CLIP models. Post finetuning, our approach trained only on 1.5M text samples, matches and outperforms OpenCLIP (BiG/14), a 2.5B parameter model trained on 2B image-text pairs.

5.4 Image-text long captions retrieval

The CLIP model's ability to process longer text is greatly restricted by the text encoder, which typically can only process up to 77 (Radford et al., 2021b) tokens. In practice, due to the data distribution of the captions, the effective length is even lower, at around 20-25 tokens. As our approach leverages pretrained language models trained on generic text, we posit that LexiCLIP is well-suited for deployment for retrieval using long text. To test this, in Tab. 4 we evaluate our approach on the Urban1k (Zhang et al., 2024) dataset. As the results show, without any finetuning, we already surpass (1) all other CLIP variants and (2) specialised CLIP models finetuned on long captions (Zhang et al., 2024).

With finetuning using the proposed approach, requiring no specialized data, we outperform prior state-of-the-art results by over 5%.

5.5 Image-text compositionality

We evaluate our LexiCLIP models on compositionality on the SugarCrepe (Hsieh et al., 2023) and SugarCrepe++ (Dumpala et al., 2024) benchmarks. As Tab. 5 shows, even without fine-tuning, LexiCLIP (0.3B) compares favorably against CLIP, outperforming the similarly sized `ViT-L` model. Our proposed finetuning process further improves the results by +7.2% pts on average, surpassing even the much larger 2.5B (BigG/14) model and achieving state-of-the-art performance. We note that the largest gains are the Swap tasks—object up +5.3% pts, attribute up +22.7% pts-which directly probe "bag-of-words" shortcuts.[1] Similar results can be observed on SugarCrepe++ in Tab. 6

6 Ablation studies and analysis

6.1 Bridging The Modality Gap

The modality gap (Liang et al., 2022) is one of the primary factors contributing to the poor compositionality of contrastive vision-language models and can also adversely affect model fairness. We assess below, on the Flickr test set, its evolution from the initial zero-shot configuration through to the post-finetuning stage by analyzing the distribution of pairwise cosine similarities and the inter-modality distance between image and text representations.

Pairwise Cosine Similarity Distributions: Fig. 3 presents the pairwise cosine-similarity distributions for three models: SigLip, LexiCLIP (0.3B)before finetuning, and LexiCLIP (0.3B)after fine-tuning. For both modalities, the pre-distributions are tightly concentrated at high similarity, indicating a partially "collapsed" space where unrelated pairs remain largely aligned. After fine-tuning, the distributions shift toward lower similarities and broaden, demonstrating that the model has learned to *decollapse* its representation space, pushing unrelated instances farther apart. This increased spread

[1] We provide the detailed results in the appendix.

Table 2: Zero-shot text–image retrieval metrics (mAP and F1@K) on SubFlickr and SubCOCO

Method	SubFlickr				SubCOCO			
	mAP	F1@1	F1@5	F1@10	mAP	F1@1	F1@5	F1@10
CLIP (ViT-B) (Radford et al., 2021b)	29.2	10.3	20.0	21.9	33.6	4.5	11.9	17.5
CLIP (ViT-L) (Radford et al., 2021b)	29.7	11.0	20.0	21.7	36.1	4.8	13.4	19.5
OpenCLIP (ViT-G/14) (Schuhmann et al., 2022)	35.3	12.0	24.5	26.6	41.2	6.3	15.9	22.0
OpenCLIP (ViT-BigG/14) (Schuhmann et al., 2022)	36.5	13.3	25.4	26.8	42.1	6.3	16.1	22.6
SigLIP ViT-B/16 (Zhai et al., 2023b)	36.6	13.6	25.6	27.2	43.6	6.5	16.4	23.0
EVA-02-CLIP (ViT-L-336) (Fang et al., 2023)	36.5	12.9	24.9	27.3	41.6	6.1	15.3	21.6
RAM (Zhang et al., 2023)	48.3	14.7	32.2	34.2	50.8	6.1	17.2	26.0
RAM++ (Huang et al., 2023a)	49.1	15.3	32.7	35.5	52.5	6.2	17.4	26.2
LexiCLIP (0.3B)–ZS	45.6	17.1	30.6	31.9	48.3	6.2	17.1	25.4
LexiCLIP (0.3B)–FT	**55.1**	**17.9**	**37.6**	**40.0**	**54.3**	**6.7**	**18.5**	**27.5**

Table 3: Zero-shot text–image retrieval accuracy on Flickr30K and COCO.

Method	Params (B)	Image retrieval				Text retrieval			
		Flickr30K		COCO		Flickr30K		COCO	
		R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10
CLIP (ViT-B) (Radford et al., 2021b)	0.15	58.8	89.8	30.5	66.8	77.8	98.2	51.0	83.5
SigLIP ViT-B/16 (Zhai et al., 2023b)	0.23	74.6	95.6	47.8	81.0	89.1	99.3	65.7	91.3
CLIP (ViT-L) (Radford et al., 2021b)	0.43	67.3	93.3	37.0	71.5	87.2	99.4	58.1	87.8
BLIP (ViT-L) (Li et al., 2022b)	0.23	70.0	95.2	48.4	83.2	75.5	97.7	63.5	92.5
BLIP2 (ViT-L) (Li et al., 2023a)	1.17	74.5	97.0	50.0	86.1	86.1	99.4	63.0	**93.1**
EVA-02-CLIP (ViT-L-336) (Fang et al., 2023)	0.43	78.0	96.8	47.9	80.0	89.6	99.6	64.2	90.9
OpenCLIP (ViT-G/14) (Schuhmann et al., 2022)	1.37	77.8	96.9	48.8	81.5	91.5	99.6	66.3	91.8
OpenCLIP (ViT-BigG/14) (Schuhmann et al., 2022)	2.54	**79.5**	**97.5**	51.3	83.0	**92.9**	97.1	67.3	92.6
LexiCLIP (0.3B)–ZS	0.3	69.5	94.2	41.7	76.7	75.9	97.4	45.4	80.3
LexiCLIP (0.3B)–FT	0.3	79.2	97.4	**52.7**	**84.5**	91.6	**99.7**	**67.4**	92.1

Table 4: Zero-shot text–image retrieval on Urban1k.

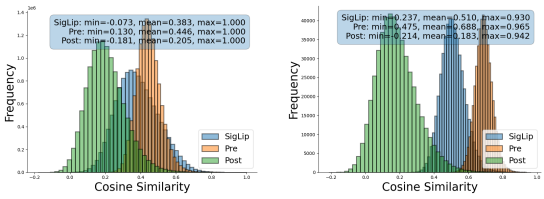
Method	Image retrieval		Text retrieval	
	R@1	R@10	R@1	R@10
CLIP (ViT-B) (Radford et al., 2021b)	46.5	78.7	62.5	90.5
SigLIP ViT-B/16 (Zhai et al., 2023b)	62.1	89.1	62.8	90.7
CLIP (ViT-L) (Radford et al., 2021b)	51.4	82.9	63.5	91.7
EVA-02-CLIP (ViT-L-336) (Fang et al., 2023)	70.1	92.5	76.7	95.5
OpenCLIP (ViT-G/14) (Schuhmann et al., 2022)	76.0	95.1	76.7	96.3
OpenCLIP (ViT-BigG/14) (Schuhmann et al., 2022)	81.9	96.1	82.0	97.6
Long-CLIP (ViT-L) (Zhang et al., 2024)	86.1	96.2	82.7	96.4
TULIP (ViT-L) (Najdenkoska et al., 2025)	91.1	—	90.1	—
LexiCLIP (0.3B)–ZS	86.9	98.2	84.4	97.5
LexiCLIP (0.3B)–FT	**97.1**	**99.9**	**96.7**	**100**

Table 5: Comparison with state-of-the-art on the SugarCrepe compositionality benchmark.

Method	Params (B)	Replace	Swap	Add	Avg.
CLIP (ViT-B) (Radford et al., 2021b)	0.15	80.1	62.7	73.0	71.9
SigLIP ViT-B/16 (Zhai et al., 2023b)	0.23	84.1	65.7	86.4	78.7
CLIP (ViT-L) (Radford et al., 2021b)	0.43	79.5	61.3	74.9	71.9
EVA-02-CLIP (ViT-L-336) (Fang et al., 2023)	0.43	84.2	65.1	89.2	79.5
BLIP (ViT-L) (Li et al., 2022b)	0.23	82.4	71.7	88.6	80.9
BLIP2 (ViT-L) (Li et al., 2023a)	1.17	85.7	63.8	89.9	79.8
OpenCLIP (ViT-G/14) (Schuhmann et al., 2022)	1.37	84.4	67.1	86.8	79.4
OpenCLIP (ViT-BigG/14) (Schuhmann et al., 2022)	2.54	86.5	68.9	88.4	81.3
NegCLIP (Yuksekgonul et al., 2023)	0.15	85.0	**75.3**	85.8	82.0
LexiCLIP (0.3B)–ZS	0.3	85.3	61.6	85.4	77.4
LexiCLIP (0.3B)–FT	0.3	**86.8**	**75.6**	**91.3**	**84.6**

is especially pronounced in the image modality, where the Post histogram has both peaks at lower cosine values and extends over a wider band, suggesting that visual features benefit strongly from fine-tuning in terms of discriminative power.

Modality Gap Comparison: Fig. 4 demonstrates



(a) Text embeddings (b) Image embeddings

Figure 3: Distributions of pairwise cosine similarities for three embedding sets: **SigLip** (blue), ***LexiCLIP (0.3B)–ZS***(orange), and ***LexiCLIP (0.3B)–FT*** (green).

that fine-tuning narrows the distance between text and image embeddings in the shared space, as shown by their projection into two dimensions via PCA. In the SigLip baseline (Fig. 4a), the centroids of text and image representations are separated by ~ 1.008, reflecting a substantial modality gap. With our method, even prior to target-task finetuning (Fig. 4b), this gap is reduced to 0.476, a significant improvement attributable to the unimodal architecture. Crucially, after finetuning (Fig. 4c), the centroid distance further decreases to 0.260. This final gap is nearly half that of our model before finetuning and roughly a quarter of the original SigLip gap. This progressive narrowing demonstrates two key points: (1) our initial zero-shot alignment significantly improves upon the SigLip, and (2) the subsequent finetuning fur-

Table 6: Comparison with state-of-the-art on the SugarCrepe++ compositionality benchmark.

Method	Params (B)	Swap Object ITT	Swap Object TOT	Swap Attribute ITT	Swap Attribute TOT	Replace Object ITT	Replace Object TOT	Replace Attribute ITT	Replace Attribute TOT	Replace Relation ITT	Replace Relation TOT	Avg. ITT	Avg. TOT
Human	–	100.00	96.7	96.7	93.3	100.00	97.0	100.00	98.3	100.00	96.7	99.3	96.4
CLIP (ViT-B) (Radford et al., 2021b)	0.15	45.2	19.7	45.2	33.0	86.8	83.7	65.6	59.1	56.3	38.6	59.8	46.8
SigLIP ViT-B/16 (Zhai et al., 2023b)	0.23	39.5	23.0	56.1	46.4	91.3	79.2	75.2	64.0	54.8	45.0	63.4	51.5
CLIP (ViT-L) (Radford et al., 2021b)	0.43	46.0	14.5	44.5	28.7	92.0	81.3	68.8	56.3	53.4	39.1	60.6	44.0
EVA-02-CLIP (ViT-L-336) (Fang et al., 2023)	0.43	44.1	19.2	47.3	34.4	94.2	91.6	74.5	69.5	59.8	48.9	64.0	52.7
BLIP (ViT-L) (Li et al., 2022b)	0.23	46.8	29.8	60.1	52.5	92.6	89.1	71.7	75.0	56.8	57.7	65.6	60.8
BLIP2 (ViT-L) (Li et al., 2023a)	1.17	37.9	39.5	51.9	55.4	**94.8**	96.9	73.2	86.5	65.1	69.6	64.6	69.6
OpenCLIP (ViT-G/14) (Schuhmann et al., 2022)	1.37	40.7	27.4	54.2	49.6	93.1	89.4	72.5	73.1	57.6	51.4	63.6	58.2
OpenCLIP (ViT-BigG/14) (Schuhmann et al., 2022)	2.54	48.8	28.2	57.7	52.4	94.2	90.5	76.4	72.6	59.4	53.6	67.3	59.5
NegCLIP (Yuksekgonul et al., 2023)	0.15	**55.3**	34.7	58.0	56.5	89.5	94.5	69.4	76.3	52.3	51.6	64.9	62.7
CLIP-SVLC (Doveh et al., 2022)	0.15	43.0	18.9	48.4	34.6	80.9	91.6	57.0	66.9	47.3	51.3	55.3	52.7
BLIP-SGVL (Herzig et al., 2023)	0.15	13.2	–	38.8	–	53.8	–	34.4	–	30.7	–	34.2	–
LexiCLIP (0.3B)–ZS	0.3	48.2	20.8	43.8	28.1	91.2	95.6	75.5	85.8	**72.4**	**77.2**	66.2	61.5
LexiCLIP (0.3B)–FT	0.3	**53.9**	**43.3**	**68.3**	**68.3**	93.8	**97.3**	**77.2**	**88.7**	65.9	71.6	**71.8**	**73.8**

ther tightens modality alignment, enhancing the cross-modal retrieval performance.

6.2 Impact of the proposed components

To better understand the key components of our data-to-text pipeline, we ablate the impact of a) object-based descriptions and b) length of the image description. Zero-shot image retrieval (on Flickr30k, COCO, Urban1K) and compositional understanding (on SugarCreppe++) are evaluated for the 0.3B **LexiCLIP (0.3B)–ZS** and a much bigger 7B parameter decoder-only model based on BGE-en-ICL-7B (Li et al., 2024), denoted as **LexiCLIP (7B)–ZS**. Based on our experiments, we draw the following conclusions:

Object-based descriptions lead to improved accuracy: As the results from Tab. 7a show, the addition of object-based descriptions consistently enhances performance across all evaluated datasets and tasks, underscoring the importance of dense object-attribute coverage.

Longer descriptions do not bring improvements: in Tab. 7b we report results for a 300M and 7B sized model on two different sequence lengths, 256 tokens and 512/1024 for the 300M/7B model. For the smaller 300M model, increasing the sequence length from 256 to its maximum of 512 tokens yields minimal gains. Similarly, the larger 7B model remains largely stable with only minor gains on Urban1k. In general, 256 tokens suffice, and further increases do not demonstrate accuracy improvements.

6.3 Captioner choices

In Tab. 8, we ablate the impact of the captioner in zero-shot retrieval of Flickr30k.

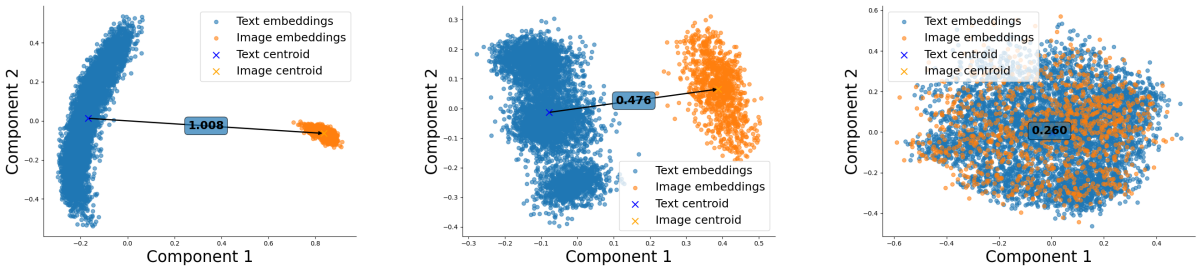
VLLM architecture: In Tab. 8a we evaluate how the choice of the image-to-text VLLM convertor impacts the downstream performance. We evaluate three similarly sized state-of-the-art VLLM, Qwen2.5-VL-7b (Bai et al., 2025a), MiniCPM-V-2-6-8b (Yao et al., 2024) and InternVL-2.5-8b-MPO. We found InternVL to perform better. This highlights that better VLLM results in higher performance and that careful consideration should be made when choosing the VLLM.

Effect of Size of the Image-to-Text VLLM Convertor: In Tab. 8b we ablate models ranging from 1B to 14B parameters as captioners. We note that our approach is robust to the size of the image captioner. Also, we observe that the 1B model performs nearly as well as the 8B and 14B models, showcasing efficient performance even with smaller models.

Captioner-Agnostic Inference: In Tab. 8c, we investigate the transferability of LexiCLIP (0.3B)-FT across different caption models. Our model, fine-tuned once on captions from the 8B InternVL model, maintains its performance even when switching to different image description sources (e.g., using a 1B model instead of 8B) without further fine-tuning. This indicates that we can fine-tune the model only once and then change the image captioner freely.

6.4 Information loss when converting Images to Text

Departing from raw pixels to text descriptions, even if they are detailed, inherently risks losing subtle visual information. To estimate how well our generated image descriptions capture visual content, we measure the overlap with the ground-truth object classes in MS-COCO (80 classes). With exact class-name matching, we get 69% match score



(a) SigLip (zero-shot) (b) Our model, pre–finetuning (c) Our model, post–finetuning

Figure 4: 2D (PCA) projections of text vs. image embeddings, with centroids (×) and the modality-gap arrow annotated by its Euclidean length. (a) SigLip exhibits a large gap of ≈ 1.008, (b) our model before fine-tuning has a gap of ≈ 0.476, (c) after fine-tuning the gap shrinks to ≈ 0.260, indicating improved alignment between modalities.

Table 7: Zero-shot retrieval accuracy and compositional understanding under two different ablations.

(a) With vs. without object-based descriptions

Model	Obj. Desc.	Flickr30K R@1	Flickr30K R@10	COCO R@1	COCO R@10	Urban1k R@1	Urban1k R@10	SugarCrepe++ (ITT) Swap Obj	Swap Attr	Repl Obj	Repl Attr	Repl Rel
LexiCLIP (0.3B)–ZS	✓	69.5	94.2	41.7	76.6	86.9	98.2	48.2	43.8	91.2	75.5	72.4
LexiCLIP (7B)–ZS	✓	74.4	95.1	46.4	80.2	91.8	99.2	49.8	57.2	91.5	77.9	77.7
LexiCLIP (0.3B)–ZS		66.4	92.8	38.7	73.8	83.7	96.5	47.3	42.5	88.6	76.6	71.9
LexiCLIP (7B)–ZS		72.1	94.5	43.4	78.0	89.2	98.7	49.0	56.2	89.0	78.9	77.2

(b) Max sequence-length ablation

Model	Max Seq Len	Flickr30K R@1	Flickr30K R@10	COCO R@1	COCO R@10	Urban1k R@1	Urban1k R@10	SugarCrepe++ (ITT) Swap Obj	Swap Attr	Repl Obj	Repl Attr	Repl Rel
LexiCLIP (0.3B)–ZS	256	69.5	94.2	41.7	76.6	86.9	98.2	48.2	43.8	91.2	75.5	72.4
LexiCLIP (7B)-ZS	256	74.4	95.1	46.4	80.2	91.8	99.2	49.8	57.2	91.5	77.9	77.7
LexiCLIP (0.3B)–ZS	512	69.1	94.0	41.1	76.3	86.9	98.2	47.8	44.4	91.3	75.0	71.2
LexiCLIP (7B)-ZS	1024	74.2	95.0	46.1	80.2	92.7	99.3	49.4	57.8	91.1	78.9	77.8

Table 8: Zero-shot retrieval on Flickr30K: Impact of the VLLM captioner.

(a) Effect of VLLM architecture.

Method	VLLM	Image retrieval R@1	Image retrieval R@10	Text retrieval R@1	Text retrieval R@10
LexiCLIP (0.3B)–ZS	InternVL2.5-8B-MPO	69.5	**94.2**	75.9	**97.4**
LexiCLIP (0.3B)–ZS	Qwen2.5-VL-7B	65.8	92.6	67.8	96.0
LexiCLIP (0.3B)–ZS	MiniCPM-V-2_6-8B	67.2	92.9	**75.3**	96.6

(b) Effect of VLLM size.

Method	VLLM	Size (B)	Image retrieval R@1	Image retrieval R@10	Text retrieval R@1	Text retrieval R@10
LexiCLIP (0.3B)–ZS	InternVL2.5-MPO	1	69.4	94.0	75.8	97.1
LexiCLIP (0.3B)–ZS	InternVL2.5-MPO	2	68.2	94.0	74.3	97.2
LexiCLIP (0.3B)–ZS	InternVL2.5-MPO	4	**70.9**	93.9	**78.6**	96.9
LexiCLIP (0.3B)–ZS	InternVL2.5-MPO	8	69.5	94.2	75.9	97.4
LexiCLIP (0.3B)–ZS	InternVL3	9	70.7	**94.6**	77.9	**97.7**
LexiCLIP (0.3B)–ZS	InternVL3	14	70.3	94.3	74.2	97.2

(c) Effect of cross-VLLM inference (different captioners at training vs. inference).

Method	VLLM	Size (B)	Image retrieval R@1	Image retrieval R@10	Text retrieval R@1	Text retrieval R@10
LexiCLIP (0.3B)–FT	InternVL2.5-MPO	1	78.7	97.1	90.8	99.5
LexiCLIP (0.3B)–FT	InternVL2.5-MPO	2	77.9	96.7	**92.0**	99.6
LexiCLIP (0.3B)–FT	InternVL2.5-MPO	4	**79.3**	**97.6**	**92.0**	99.4
LexiCLIP (0.3B)–FT	InternVL2.5-MPO	8	79.2	97.4	91.6	**99.7**
LexiCLIP (0.3B)–FT	InternVL3	9	**79.5**	97.2	91.9	**99.9**
LexiCLIP (0.3B)–FT	InternVL3	14	78.7	97.0	**92.0**	99.5

between objects in our descriptions and those in ground truth annotations. However, we note that possibly there is a higher overlap, as our estimation does not account for synonyms or paraphrasing.

7 Conclusions

We introduce a text-to-text paradigm for training a vision-free single-encoder CLIP model, challenging the conventional two-tower paradigm. Our framework uses VLLMs to generate structured image descriptions and omits images during training. This reduces the modality gap and improves compositional generalisation while achieving better performance on short caption queries. Unlike traditional two-tower architectures, our LexiCLIP is able to effectively model the full spectrum of query distributions, from brief user-centric queries to long, paragraph-level descriptions - all using the same single encoder. To further assess compositional generalisation, we release subFlickr and subCOCO, two curated benchmarks with diverse compositional queries made of short captions. Finally, we show that models with 0.3B parameters can match or even surpass traditional multimodal architectures, achieving SOTA results across multiple compositionality benchmarks and retrieval tasks.

8 Limitations

The limitations of this work are mostly related to its strong dependence on VLLMs. As we depart from raw pixels to text-based image descriptions generated by VLLMs, certain visual details will probably be lost. In particular, descriptions of crowded scenes or those containing many small objects are likely to omit a significant amount of information. Moreover, since VLLMs are often biased or have hallucinations, generated descriptions often inherit them. Such errors can potentially propagate into the retrieval process. As future work, to mitigate these issues, we will try to adopt filtering methods, ensemble captioners, or even some kind of human-in-the-loop verification. Then, another limitation of using VLLMs as an image captioner is the extra computational overhead introduced. We calculate that for an A100 GPU, an unoptimized implementation requires approximately 0.2 seconds per image. This cost can be significantly reduced through optimized implementations (e.g., the `vllm` project) and techniques like quantization. In our work, this step is performed once, offline, ahead of evaluation, while retrieval itself remains efficient. On an A100 GPU, our method requires only 1.6 ms per image, compared to 7.4 ms for OpenCLIP-2.54B. Finally, we note that similar retrieval performance can be obtained with smaller generators (e.g., a 2B InternVL), which can lower the preprocessing cost.

9 Broader Impact

We also reflect on the broader impact and ethics of our work. Given that the main body of our retrieval pipeline is based on text rather than images, we consider that LexiCLIP allows for a more interpretable and transparent retrieval. Additionally, LexiCLIP is a more inclusive approach for users with visual impairments. However, we acknowledge that the heavy reliance of our work, in VLLMs to generate image descriptions, is accompanied by inherited biases that may reinforce stereotypes or amplify unfair associations.

10 Acknowledgments

This work was partially funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (grant No. 10099264) and by the European Union (under EC Horizon Europe grant agreement No. 101135800 (RAIDO)).

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8948–8957. 3, 4, 16

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025a. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923. 8

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*. 4, 5

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*. 4, 5

Guillem Collell Talleda and Marie-Francine Moens. 2016. Is an image worth more than a thousand words? on the fine-grain semantic differences between visual and linguistic representations. In *Proceedings of the 26th international conference on computational linguistics*, pages 2807–2817. ACL. 4

Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. 2022. Why is winoground hard? investigating failures in visuolinguistic compositionality. *Preprint*, arXiv:2211.00768. 16

Sivan Doveh, Assaf Arbelle, Sivan Harary, Rameswar Panda, Roei Herzig, Eli Schwartz, Donghyun Kim, Raja Giryes, Rogério Schmidt Feris, Shimon Ullman, and Leonid Karlinsky. 2022. Teaching structured vision & language concepts to vision & language models. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2657–2668. 8

Sri Harsha Dumpala, Aman Jaiswal, Chandramouli Sastry, Evangelos Milios, Sageev Oore, and Hassan Sajjad. 2024. Sugarcrepe++ dataset: Vision-language model sensitivity to semantic and lexical alterations. In *Advances in Neural Information Processing Systems*, volume 37, pages 17972–18018. Curran Associates, Inc. 4, 6

Yuxin Fang, Hao Zhang, Wen Wang, Zhiyu Yu, Xiaojie Zhu, and Cewu Lu. 2023. Eva-02: A visual representation for neon genesis. *arXiv preprint arXiv:2303.11331*. 7, 8, 16

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*. 4

Sophia Gu, Christopher Clark, and Aniruddha Kembhavi. 2023. I can't believe there's no images! learning visual tasks using only language supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2672–2683. 2

Roei Herzig, Alon Mendelson, Leonid Karlinsky, Assaf Arbelle, Rogerio Feris, Trevor Darrell, and Amir Globerson. 2023. Incorporating structured representations into pretrained vision & language models using scene graphs. *Preprint*, arXiv:2305.06343. 8

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. 5

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. 2023. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems*, volume 36, pages 31096–31116. Curran Associates, Inc. 2, 4, 6

Xinyu Huang, Yi-Jie Huang, Youcai Zhang, Weiwei Tian, Rui Feng, Yuejie Zhang, Yanchun Xie, Yaqian Li, and Lei Zhang. 2023a. Open-set image tagging with multi-grained text supervision. *arXiv e-prints*, pages arXiv–2310. 3, 5, 6, 7

Xinyu Huang, Youcai Zhang, Jinyu Ma, Weiwei Tian, Rui Feng, Yuejie Zhang, Yaqian Li, Yandong Guo, and Lei Zhang. 2023b. Tag2text: Guiding vision-language model via image tagging. *arXiv preprint arXiv:2303.05657*. 3, 5

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR. 1

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020a. 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. 2

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020b. The open images dataset v6: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision (IJCV)*, 128(7):1956–1981. 4

Soeun Lee, Si-Woo Kim, Taewhan Kim, and Dong-Jin Kim. 2024. Ifcap: Image-like retrieval and frequency-based entity filtering for zero-shot captioning. *Preprint*, arXiv:2409.18046. 2

Chaofan Li, MingHao Qin, Shitao Xiao, Jianlyu Chen, Kun Luo, Yingxia Shao, Defu Lian, and Zheng Liu. 2024. Making text embedders few-shot learners. *Preprint*, arXiv:2409.15700. 8

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR. 1, 7, 8, 16

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022a. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR. 1

Junnan Li, Dongxu Li, Caiming Xiong, and Steven CH Hoi. 2022b. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*. 7, 8, 16

Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. 2023b. Decap: Decoding clip latents for zero-shot captioning via text-only training. *Preprint*, arXiv:2303.03032. 2

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. 2022c. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *Preprint*, arXiv:2110.05208. 1

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625. 1, 3, 6

Valerii Likhosherstov, Anurag Arnab, Krzysztof Choromanski, Mario Lucic, Yi Tay, Adrian Weller, and Mostafa Dehghani. 2021. Polyvit: Co-training vision transformers on images, videos and audio. *arXiv preprint arXiv:2111.12993*. 1

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing. 1, 3, 4, 6

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*. 5

Ivona Najdenkoska, Mohammad Mahdi Derakhshani, Yuki M. Asano, Nanne van Noord, Marcel Worring, and Cees G. M. Snoek. 2025. Tulip: Token-length upgraded clip. *Preprint*, arXiv:2410.10034. 7

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021a. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR. 1

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021b. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR. 2, 3, 6, 7, 8, 16

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, volume 35, pages 25278–25294. Curran Associates, Inc. 7, 8, 16

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*. 4

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5238–5248. 4

Michael Tschannen, Basil Mustafa, and Neil Houlsby. 2023. Clippo: Image-and-language understanding from pixels only. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11006–11017. 2

Junyang Wang, Ming Yan, Yi Zhang, and Jitao Sang. 2023. From association to generation: Text-only captioning by unsupervised cross-modal mapping. *Preprint*, arXiv:2304.13273. 2

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. 2025. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *Preprint*, arXiv:2411.10442. 4

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597. 5, 13

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*. 8

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. 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. 1, 3, 4, 6

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2022. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*. 3

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. When and why vision-language models behave like bags-of-words, and what to do about it? *Preprint*, arXiv:2210.01936. 1, 7, 8, 16

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023a. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986. 1

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023b. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*. 1, 7, 8, 16

Beichen Zhang, Pan Zhang, Xiao wen Dong, Yuhang Zang, and Jiaqi Wang. 2024. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*. 6, 7

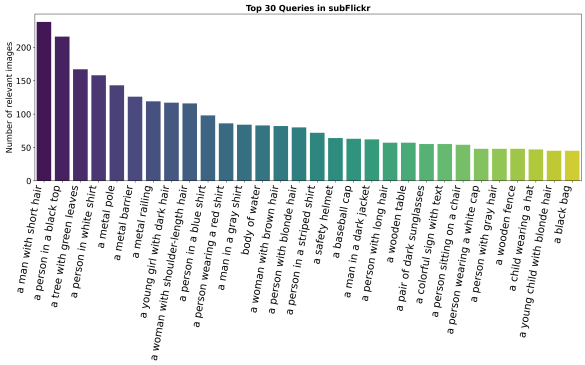
Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, and 1 others. 2023. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*. 3, 5, 6, 7

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 633–641. 4

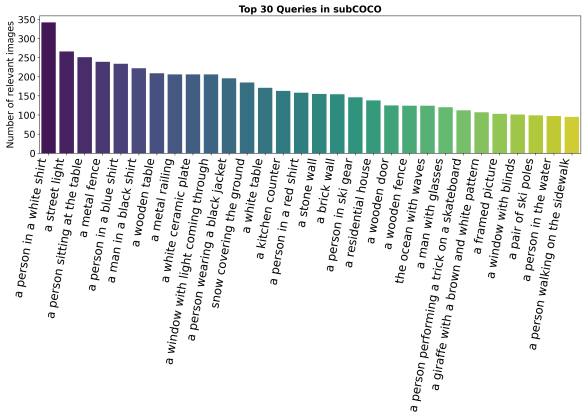
A Technical Appendices and Supplementary Material

A.1 subFlickr and subCoco benchmarks

Fig. 5 shows the distribution of the top 30 most frequent queries in the *subFlickr* and *subCOCO* datasets. Both datasets have long-tail distribution with *subCOCO* showing a slightly sharper drop in frequency.



(a) Top 30 most frequent queries in the *subFlickr* dataset.



(b) Top 30 most frequent queries in the *subCOCO* dataset.

Figure 5: Query frequency distributions in the *subFlickr* (a) and *subCOCO* (b) datasets. The y-axis indicates how many images are relevant to each query.Looks better zoomed in.

A.2 Representing images using text

To convert images into rich, structured text, we employ the OpenGVLab/InternVL2_5-8B-MPO model[2]. We extract two complementary views of each image:

- A *detailed scene description*, obtained by prompting the model with:

  ```
  Please describe the image in
  detail.
  ```

- *Object annotations*, generated using the prompt shown in Fig. 6.

Fig. 7 presents several examples of this text-based representation—on the left, the raw image; on the right, our model's concatenated scene summary and per-object attribute list.

A.3 Generation of compositional captions

As stated in 5.1, we train our pipeline in two stages. Given that OpenImages is not paired with longer or compositional captions, we synthetically generate them. For concise captions, we run BLIP-2. For smaller compositional captions, we ask InternVL-2.5-8B-MPO to generate a pool of 6 compositional captions for each image based on its generated image description in order to accelerate the data preparation process. Below, we provide the prompt we have used.

> You are given a scene caption and a list of structured object descriptions extracted from an image. Your task is to generate 6 short compositional search queries (2 to 4 words) that someone might use to find this image. Each query should refer to an object, attribute, or visual element described in the input.
>
> Use combinations like: - object + color (e.g., "red bird") - object + position (e.g., "bird on branch") - object + action (e.g., "bird flying") - or noun phrases (e.g., "bird cage")
>
> Return only a raw JSON list of 6 strings like: ["black bird", "bird feeder", "bird in cage", "wooden birdhouse", "birds perched", "red flowers"]
>
> Do not include markdown or code formatting such as triple backticks or `json` labels.

A.4 Training details

We train our model in two stages. For both stages, the model is trained for three epochs using the FlagEmbedding library (Xiao et al., 2023), with a cosine annealed learning rate schedule and a warm-up phase 5%. The model is trained on two A100 GPUs with an effective batch size of 2,048, a peak learning rate of 1×10^{-4} and a weight decay of 0.1. Training employs mixed precision (FP16), gradient checkpoint, and DeepSpeed.

[2] https://huggingface.co/OpenGVLab/InternVL2_5-8B-MPO

"List and describe each main subject or object in the image. For every item, include details about what it is doing, its spatial relationship to other elements, and a brief description of its appearance or notable features. In addition, provide specific attributes such as:"

-Color: The primary color or color scheme.
-Size: The relative or absolute size."
-Texture/Material: The texture or material quality (e.g., smooth, rough, metallic).
-Orientation/Direction: The way the object is facing or its directional alignment.
-Additional Context: Any other relevant details that help describe the object.

If any detail is unavailable, set its value to null. Please output your answer as a JSON array, where each element is an object containing the keys 'object', 'object_description', 'action', 'position', 'color', 'size', 'texture', and 'orientation'. Optionally, you may include other keys if applicable. Return only the JSON array without any additional text or formatting. Limit the output to at most 7 objects.

For example:

```
[
 {
  "object": "cat",
  "object_description": "a small tabby with green eyes",
  "action": "sitting",
  "position": "to the left of the couch",
  "color": "brown and white",
  "size": "small",
  "texture": "fluffy",
  "orientation": "facing forward"
 },
 {
  "object": "lamp",
  "object_description": "a tall, modern lamp",
  "action": "standing",
  "position": "near the window",
  "color": "black",
  "size": "medium",
  "texture": "smooth",
  "orientation": "angled slightly to the right"
 }
]
```

Ensure your response is valid JSON and that all objects adhere exactly to this structure."

Figure 6: The prompt that was given to the model for extracting the object annotations.

During the first stage, the training is done only using BLIP-2 captions using a symmetric contrastive loss. In the second stage, we fine-tune the model using a mixture of BLIP-2 and compositional captions, where each batch of 2,048 samples consists of 200 concise BLIP-2 captions and the remainder compositional ones. In this stag,e the loss is only text-to-image. The final model is obtained by averaging the checkpoints from the two stages, weighted 0.4 for the first and 0.6 for the second stage.

B Additional ablation studies

We extend our ablation studies of Sect. 6 to investigate two aspects. First, we assess the effect of the maximum sequence length on downstream retrieval in Tab. 9. We note a clear performance drop for descriptions shorter than 128 and diminishing returns above 256. Second, in Tab. 10 we check the zero-shot retrieval performance on Flickr30K by varying the size of the VLLM used to generate image and object-level descriptions, while keeping the 7B model as the retriever.

Table 9: Zero-shot retrieval on Flickr30K, ablating the max seq. len used for encoding the image and object descriptions.

Method	Max Seq Len	Image retrieval		Text retrieval	
		R@1	R@10	R@1	R@10
LexiCLIP (0.3B)–ZS	64	63.3	90.1	71.5	94.6
LexiCLIP (0.3B)–ZS	128	67.5	93.2	73.5	95.8
LexiCLIP (0.3B)–ZS	256	**69.5**	**94.0**	**75.9**	**97.4**
LexiCLIP (0.3B)–ZS	512	69.1	94.0	74.8	97.1

Table 10: Zero-shot retrieval on Flickr30K, ablating the size of the VLLM used for extracting the image and object descriptions when using the 7B model as the retriever.

Method	VLLM	Size (B)	Image retrieval		Text retrieval	
			R@1	R@10	R@1	R@10
LexiCLIP (7B)–ZS	InternVL2.5-MPO	4	75.2	95.7	81.0	97.4
LexiCLIP (7B)–ZS	InternVL2.5-MPO	8	74.4	95.1	**82.6**	**98.0**
LexiCLIP (7B)–ZS	InternVL3	9	**75.6**	95.4	82.3	97.8
LexiCLIP (7B)–ZS	InternVL3	14	74.3	**95.8**	80.6	96.8

B.1 SugarCrepe Detailed Results

Table 11 shows the detailed results of our method in all seven categories. Our finetuned model achieves state-of-the-art performance.

Raw Image

VLLM-generated Structured Image Descriptions

Scene Description:

The image shows a woman and a young child outdoors. The woman has short dark hair and is smiling as she looks at the child. The child, who appears to be a toddler, is sitting on the woman's shoulders. The child has light hair and is wearing a striped navy and white shirt with khaki shorts. They are both in a sunny area with trees and buildings in the background. The buildings have balconies and windows, suggesting an urban setting. The scene is bright and cheerful, capturing a moment of joy and connection.

Objects in the Scene:

- **Object Category:** Woman
  - Attributes: A woman with short dark hair
  - Action: Holding a child
  - Position: In the foreground
  - Color: Dark hair, light skin
  - Size: Adult
  - Texture: Smooth
  - Orientation: Smooth
- **Object Category:** Child
  - Attributes: A young child with light hair
  - Action: Being held
  - Position: On the woman's shoulders
  - Color: Light hair, striped shirt
  - Size: Small
  - Texture: Smooth
  - Orientation: Smooth
- **Object Category:** Building
  - Attributes: A multi-story building with balconies
  - Action: Standing
  - Position: Background
  - Color: Beige
  - Size: Large
  - Texture: Smooth
  - Orientation: Smooth
- **Object Category:** Tree
  - Attributes: A tree with green leaves
  - Action: Standing
  - Position: Background, top
  - Color: Green
  - Size: Large
  - Texture: Rough
  - Orientation: Rough
- **Object Category:** Car
  - Attributes: A parked car
  - Action: Parked
  - Position: Background, left side
  - Color: Dark
  - Size: Medium
  - Texture: Smooth
  - Orientation: Smooth

Raw Image

VLLM-generated Structured Image Descriptions

Scene Description:

The image shows a young girl balancing on a wooden beam in what appears to be a gymnasium. She is wearing a navy blue dress with yellow polka dots and an orange long-sleeve shirt underneath. Her hair is tied back with a pink clip. The gym has a wooden floor with mats and a blue balance beam. In the background, there are other children and adults, some sitting and some standing. The setting suggests a gymnastics or physical activity class.

Objects in the Scene:

- **Object Category:** Girl
  - Attributes: A young girl in a dress
  - Action: Balancing
  - Position: On a balance beam
  - Color: Blue and orange
  - Size: Small
  - Texture: Smooth
  - Orientation: Smooth
- **Object Category:** Person
  - Attributes: A person in pink
  - Action: Walking
  - Position: To the left
  - Color: Pink
  - Size: Medium
  - Texture: Smooth
  - Orientation: Smooth
- **Object Category:** Backpack
  - Attributes: A black backpack
  - Action: On the floor
  - Position: Behind the girl
  - Color: Black
  - Size: Medium
  - Texture: Rough
  - Orientation: Rough
- **Object Category:** Child
  - Attributes: A child in red
  - Action: Sitting
  - Position: In the background
  - Color: Red
  - Size: Small
  - Texture: Smooth
  - Orientation: Smooth
- **Object Category:** Person
  - Attributes: A person in black
  - Action: Bending
  - Position: In the background
  - Color: Black
  - Size: Medium
  - Texture: Smooth
  - Orientation: Smooth
- **Object Category:** Wall
  - Attributes: A beige brick wall
  - Action: Standing
  - Position: Background
  - Color: Beige
  - Size: Large
  - Texture: Brick
  - Orientation: Brick

Raw Image

VLLM-generated Structured Image Descriptions

Scene Description:

The image shows two men engaged in what appears to be a martial arts or self-defense training session. The man on the left, wearing a red shirt, is demonstrating a technique by placing his hand on the neck of the man on the right, who is wearing a gray sweatshirt. The man in red has a focused expression and is flexing his arm, possibly to show strength or control. The background features a brick wall and some gym equipment, suggesting this is taking place in a training facility. The gray sweatshirt has some circular designs on the sleeve.

Objects in the Scene:

- **Object Category:** Man 2
  - Attributes: A man in a gray shirt
  - Action: Leaning forward
  - Position: Leaning over
  - Color: Gray
  - Size: Medium
  - Texture: Smooth
  - Orientation: Smooth
- **Object Category:** Man 1
  - Attributes: A man in a red shirt
  - Action: Demonstrating a technique
  - Position: Standing
  - Color: Red
  - Size: Medium
  - Texture: Smooth
  - Orientation: Smooth

Raw Image

VLLM-generated Structured Image Descriptions

Scene Description:

The image shows four children gathered around a barbecue grill at night. They appear to be enjoying themselves, with smiles on their faces. The child on the left is wearing a striped shirt and a cap, leaning back slightly with a playful expression. Next to him, a child in a white hoodie is holding a bowl and smiling. The third child, wearing a dark blue shirt, is also holding a bowl and smiling. The fourth child, on the right, is wearing a white shirt and is holding a bowl as well. The grill is filled with pieces of meat, and there are various bottles and containers on a nearby table. The scene is illuminated by a lantern, creating a warm and casual atmosphere.

Objects in the Scene:

- **Object Category:** Child
  - Attributes: A child in a white shirt
  - Action: Holding a bowl
  - Position: Standing to the right
  - Color: White
  - Size: Small
  - Texture: Smooth
  - Orientation: Smooth
- **Object Category:** Child
  - Attributes: A child wearing a striped shirt and cap
  - Action: Leaning back slightly
  - Position: In front of the grill, left side
  - Color: Multicolored
  - Size: Small
  - Texture: Smooth
  - Orientation: Smooth
- **Object Category:** Child
  - Attributes: A child in a blue shirt
  - Action: Holding a bowl and smiling
  - Position: Standing near the grill
  - Color: Blue
  - Size: Small
  - Texture: Smooth
  - Orientation: Smooth
- **Object Category:** Grill
  - Attributes: A barbecue grill with food on it
  - Action: Cooking
  - Position: Center of the image
  - Color: Metallic
  - Size: Medium
  - Texture: Metal
  - Orientation: Metal
- **Object Category:** Child
  - Attributes: A child in a white hoodie
  - Action: Holding a bowl and smiling
  - Position: Standing next to the first child
  - Color: White
  - Size: Small
  - Texture: Smooth
  - Orientation: Smooth
- **Object Category:** Bowl
  - Attributes: A white bowl
  - Action: Held by the child in the white hoodie
  - Position: In front of the child in the white hoodie
  - Color: White
  - Size: Small
  - Texture: Smooth
  - Orientation: Smooth
- **Object Category:** Bowl
  - Attributes: A white bowl
  - Action: Held by the child in the white shirt
  - Position: In front of the child in the white shirt
  - Color: White
  - Size: Small
  - Texture: Smooth
  - Orientation: Smooth

Figure 7: Examples of our image representation in text: on the left, the raw image; on the right, the corresponding structured description generated by our model.

Table 11: Comparison with state-of-the-art on the SugarCrepe compositionality benchmark.

Method	Params (B)	Replace			Swap		Add		Avg.
		Object	Attribute	Relation	Object	Attribute	Object	Attribute	
CLIP (ViT-B) (Radford et al., 2021b)	0.15	90.9	80.1	69.2	61.4	64.0	77.2	68.8	73.1
SigLIP ViT-B/16 (Zhai et al., 2023b)	0.15	95.3	86.7	70.3	60.0	71.5	89.1	83.8	79.5
CLIP (ViT-L) (Radford et al., 2021b)	0.43	94.1	79.2	65.2	60.2	62.3	78.3	71.5	73.0
EVA-02-CLIP (ViT-L-336) (Fang et al., 2023)	0.43	96.6	85.1	70.9	64.9	65.3	**92.9**	82.1	80.2
BLIP (ViT-L) (Li et al., 2022b)	0.23	96.5	81.7	69.1	66.6	76.8	92.0	85.1	81.1
BLIP2 (ViT-L) (Li et al., 2023a)	1.17	**97.6**	81.7	77.8	62.1	65.5	92.4	87.4	80.6
OpenCLIP (ViT-G/14) (Schuhmann et al., 2022)	1.37	95.8	85.0	72.4	63.0	71.2	91.5	82.1	80.1
OpenCLIP (ViT-BigG/14) (Schuhmann et al., 2022)	2.54	96.6	**87.9**	74.9	62.5	75.2	92.2	84.5	81.9
NegCLIP (Yuksekgonul et al., 2023)	0.15	92.7	85.9	76.5	**75.2**	75.4	88.8	82.8	82.5
LexiCLIP (0.3B)–ZS	0.3	94.0	82.5	**79.3**	63.7	59.6	84.1	86.8	78.6
LexiCLIP (0.3B)–FT	0.3	96.7	86.3	77.5	69.0	**82.3**	90.7	**91.9**	**84.9**

Table 12: Zero-shot text–image and image–text retrieval on **NoCaps** in the Out-of-Domain partition.

Method	Params (B)	Image retrieval			Text retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10
CLIP (ViT-B/16) (Radford et al., 2021b)	0.15	53.5	85.7	93.3	69.4	93.6	97.8
CLIP (ViT-L/14) (Radford et al., 2021b)	0.43	56.8	87.0	93.2	74.9	95.8	98.4
OpenCLIP (ViT-G/14) (Schuhmann et al., 2022)	1.37	70.8	93.5	97.3	85.6	98.2	**99.8**
OpenCLIP (ViT-bigG/14) (Schuhmann et al., 2022)	2.54	**72.2**	93.9	97.3	85.4	**98.8**	99.5
SigLIP ViT-B/16 (Zhai et al., 2023b)	0.23	71.5	93.6	97.3	85.3	98.6	99.8
EVA-02-CLIP (ViT-L-336) (Fang et al., 2023)	0.43	66.6	91.2	96.0	81.4	97.4	98.9
LexiCLIP (0.3B)–ZS	0.30	67.8	91.4	96.5	79.4	96.4	98.5
LexiCLIP (0.3B)–FT	0.30	71.0	**93.7**	97.3	**85.9**	97.3	99.3

Table 13: Zero-shot compositional retrieval on Winoground (Diwan et al., 2022) across Group, Image, and Text scores.

Model	Params (B)	Image	Text	Group
CLIP (ViT-B/16) (Radford et al., 2021b)	0.15	10.5	25.0	7.3
CLIP (ViT-L/14) (Radford et al., 2021b)	0.43	12.3	27.5	8.3
OpenCLIP (ViT-G/14) (Schuhmann et al., 2022)	1.37	12.8	32.0	9.3
OpenCLIP (ViT-bigG/14) (Schuhmann et al., 2022)	2.54	15.5	35.5	12.0
SigLIP ViT-B/16 (Zhai et al., 2023b)	0.23	13.0	33.0	10.5
LexiCLIP (0.3B)–ZS	0.33	6.7	24.7	3.7
LexiCLIP (0.3B)–FT	0.33	13.3	35.8	10.8

B.2 Generalisation to out-of-domain data

To further validate our approach, we evaluated it on two additional datasets: (1) Out-of-Domain Subset NoCaps (Agrawal et al., 2019), a subset specifically designed to assess retrieval performance on images that are out-of-domain relative to the COCO and Flickr datasets. (2) Winoground (Diwan et al., 2022), a challenging dataset exhibiting a combination of unusual, adversarial images, sketches, etc.

As the results below demonstrate, our approach achieves performance comparable to significantly larger models like OpenCLIP 2.54B. This is noteworthy given that our model was trained on only 1.5 million textual samples, whereas OpenCLIP utilized a massive 5 billion image-text pairs.

C Ethics Statement and Artifacts

Licenses. We follow the original licenses of all datasets and models used in this work. Our released artifacts (code, subFlickr, subCOCO) will be distributed under the MIT license.

Intended Use. All datasets and models used in this work were employed strictly for research purposes, in accordance with their original intended use. Our derived benchmarks (subFlickr and subCOCO) are released exclusively for research use, consistent with the original licensing and access conditions of COCO and Flickr30k.

Documentation. All datasets and models used in this work are well-documented in their original publications. Our derived benchmarks (subFlickr and subCOCO) contain short, compositional En-

glish queries paired with corresponding images, designed to evaluate fine-grained retrieval. We release the benchmarks with accompanying documentation to ensure transparency and reproducibility.