Discovering Meaningful Units with Visually Grounded Semantics from Image Captions

Anonymous ACL submission

Abstract

Fine-grained knowledge is crucial for vision-language models to obtain a better understanding of the real world. While there has been work trying to acquire this kind of knowledge in the space of vision and language, it has mostly focused on aligning the image patches with the tokens on the language side. However, image patches do not have any meaning to the human eye, and individual tokens do not necessarily carry groundable information in the image. It is groups of tokens which describe different aspects of the scene. In this work, we propose a model which groups the caption tokens as part of its architecture in order to capture a fine-grained representation of the language. We expect our representations to be at the level of objects present in the image, and therefore align our representations with the output of an image encoder trained to discover objects. We show that by learning to group the tokens, the vision-language model has a better fine-grained understanding of vision and language. In addition, the token groups that our model discovers are highly similar to groundable phrases in text, both qualitatively and quantitatively.

1 Introduction

Vision-language models have been shown to be less effective at capturing fine-grained information about the images described by the captions (Bugliarello et al., 2023; Kamath et al., 2023; Yuksekgonul et al., 2022). This information is crucial for the models to obtain a better understanding of the real world. While there has been work trying to acquire this kind of knowledge in the space of vision and language, it has mostly focused on aligning the image patches with the tokens on the language side (Yao et al., 2022; Wang et al., 2022; Zeng et al., 2022a; Mukhoti et al., 2023). However, image patches do not have any meaning to the human eye, and individual tokens often do not carry information groundable in the image. Minimally, it is groups of image patches which represent objects and the group of tokens in the text that refer to those objects. For this reason, there has been an active line of research in vision investigating the unsupervised discovery of objects by learning to assign image patches to their representative object slots (Locatello et al., 2020; Sajjadi et al., 2022; Wu et al., 2024). Recently, Xu et al. (2022) integrated an object discovery module into their vision-language model to learn the object entities. They showed that representing the image at the level of its constituent objects improves the performance of their model in downstream tasks. In this paper, we investigate the unsupervised discovery of groundable phrases on the language side to get better correspondence with objects on the vision side. We hypothesise that finding these meaningful units in language representations will improve the fine-grained understanding of image-caption semantic relationships. As far as we are aware, we are the first to investigate this possibility.

We base our model on the recent model of visual object discovery using image caption pairs proposed by Xu et al. (2022). We freeze the image side of the model, and introduce analogous deep learning mechanisms to discover *objects*[1] on the language side. We investigate two types of losses, one which promotes the correspondence between representations of the language side and representations on the vision side, and one which promotes the ability to reconstruct the text from the language representations. We find that training with both these losses leads to better fine-grained understanding of the image-text relationship, and discovers units which are highly similar to groundable phrases in text, both qualitatively and quantitatively. Further analysis finds that optimising the image-text correspondence alone does not lead to the discovery of

[1]We use the terms *objects*, *entities*, *groups* and *units* interchangeably.

meaningful units on the language side, and while this model does learn a good fine-grained understanding of the image-text relationship, it does not represent the semantics of objects as well as the model which does represent groundable phrases. We also find that optimising the reconstruction loss alone does lead to the discovery of meaningful units on the language side, but they have a slightly worse similarity to groundable phrases than the model which includes grounding information, and do not capture image-text relationships.

Our contributions are as follows,

- We develop a novel model to discover meaningful units from the image captions in the vision language setup (Section 2.1).
- We improve the fine-grained vision and language understanding of our model compared to a single-vector representation of text, under two different benchmarks (Section 4.2).
- We show that the segments that our model discovers are meaningful both qualitatively, and in terms of accordance with groundable phrases (Section 4.3).

2 Method

To facilitate learning the fine-grained semantics of image-text relationships, we propose a model for learning text representations whose granularity matches the granularity of objects in the image, meaning that it is neither as course-grained as having a single vector for embedding the entire text[2] nor as fine-grained as having a different vector for every token. Given a dataset of image-caption pairs, $D = \{(I_i, T_i)\}_{i=1,\dots,N}$, we want to learn a representation of each caption in the form of groups of tokens which are aligned with the semantic space of objects in its image. To do so, we freeze the image encoder which has been trained to output the objects in the image and only train the text encoder and the projection heads. In particular, if the input representation of language is at the level of subwords, we aim to find a higher level representation of them which would approximately represent groundable phrases. More specifically, let $T_i = [t_{i1}, \dots, t_{iM}]$, where t_{ij} is a subword of T_i and M is the total number of subwords in T_i. We would like to group the subword tokens t_{ij}s into non-overlapping groups $T_i = \{g^T_{i1}, \dots, g^T_{iK}\}$ where $K < M$. This would lead to a more compact abstract representation of T_i.

[2]This is the common way of representing text in dual-stream vision-language models like CLIP (Radford et al., 2021).

2.1 Model

We illustrate an overview of our model in Figure 1 and describe each of its components in the following sections.

2.1.1 Text Encoder: Text Group Transformer

We design our text encoder to learn semantic units of language. The key idea is to have shared learnable group vectors which can bind to different tokens of input (Xu et al., 2022). At each stage the groups carry the information from the previous layer to the next layer. To initiate the binding, the groups are appended to the input tokens they need to bind, and they all interact via several Transformer encoder layers to allow the groups and tokens to exchange information. Then, by performing a top-down attention mechanism shown as the Grouping block, the groups bind to different parts of the input.

More specifically, we first embed the input tokens and add learned positional encodings to them. Then, we append the learnable group vectors, $[g^T_{ik}]_{k=1\dots K}$, to these embedded inputs, $[t_{ij}]_{j=1\dots M}$, and pass the resulting vectors through some Transformer encoder layers, allowing them to interact with each other. We denote the encoded tokens and groups as $\hat{t}_{ij}$ and $\hat{g}_{ik}$. Then the grouping happens in a grouping block. In this block, the groups act as the queries and the encoded inputs as keys and values through a top-down attention mechanism. As with standard attention, the raw attention scores are computed as

$$A^{\text{raw}}_{kj} = \frac{Q(\hat{g}^T_{ik})K^\intercal(\hat{t}_{ij})}{\sqrt{d}} \tag{1}$$

where d is the dimension of the model and Q and K are linear query and key projections. In order to have discrete assignments of inputs to the groups, GroupViT actually performs a hard assignment over A^{raw} by utilizing Gumble softmax (Jang et al., 2017; Maddison et al., 2017). Namely,

$$A' = \text{Gumble Softmax}(A^{\text{raw}}). \tag{2}$$

In top-down attention, instead of normalizing over the keys in the softmax function, the A' weights are first normalized over the queries, which are the groups. This will make the groups compete

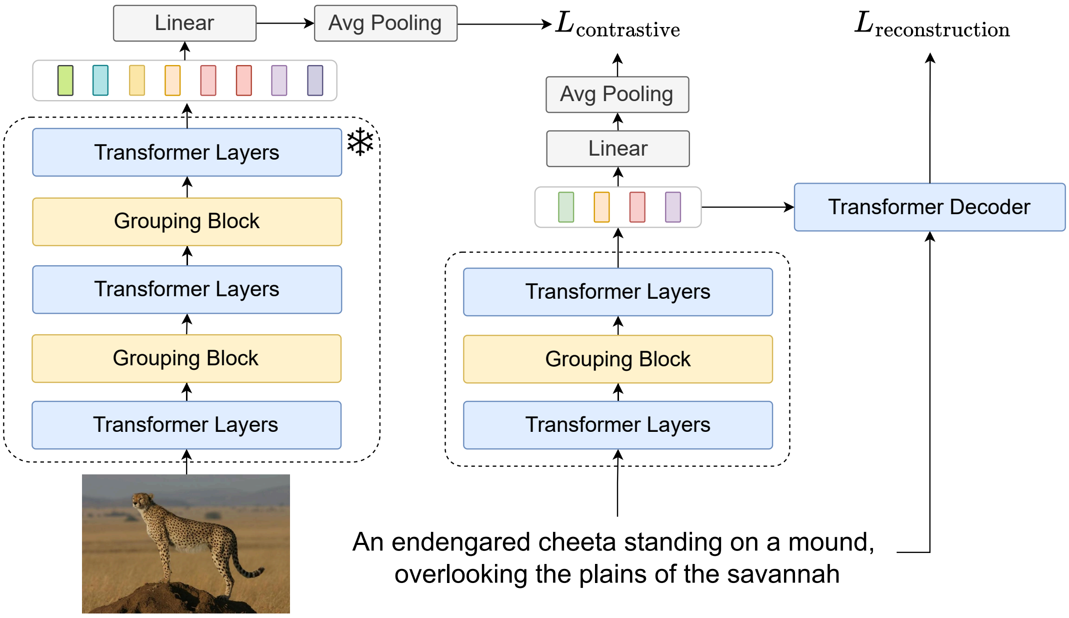

Figure 1: Overview of the model. We freeze the image encoder and only train the text encoder, decoder and the linear projection heads. The image passes through Transformer layers followed by the grouping blocks. The output of the image encoder is a set of groups which are approximately representing the objects. The caption also passes through the same set of blocks and the output of the text encoder is a set of groups representing units in language. The two modalities interact via a contrastive loss. There is also a reconstruction loss where the decoder decodes the text groups into the original input.

for representing different inputs (Locatello et al., 2020) and has been shown to be the most important component in discovering objects (Wu et al., 2023). After the normalization, the hard assignment happens and the gradient is backpropagated with the straight through trick (Van Den Oord et al., 2017), that is:

$$A = \text{one-hot}(\text{argmax}_{\text{groups}}(A')) - \text{sg}(A') + A' \tag{3}$$

where sg is the stop gradient operator. Finally, the group vectors get updated as

$$\bar{g}_{ik}^T = \hat{g}_{ik}^T + W(\sum_j \frac{A_{kj}}{\sum_j A_{kj}} V(t_{ij})) \tag{4}$$

where V and W are the linear projections for values and outputs respectively.

After the grouping block, the updated group vectors serve as inputs to subsequent Transformer encoder layers. Finally, these refined groups represent the fine-grained semantics of the text in our model.

2.1.2 Image Encoder

We use the image encoder of Xu et al. (2022), which follows the same architecture as the text encoder, but with two stacked levels of transformer encoder layers and grouping blocks. As its input, the images are first divided into patches and then linearly projected. The encoder then extracts the set of image groups denoted as $\{\bar{g}_{ik}^I\}$. Due to the computational cost, we freeze the image encoder and assume that the image groups are representing objects in the image.

2.2 Training Objectives

Our model is trained with two different losses, i.e., a contrastive loss and a reconstruction loss, which we will explain in the following. The two losses are combined with a hyperparameter λ which controls the ratio between the two terms.

$$L_{\text{total}} = L_{\text{contrastive}} + \lambda L_{\text{recontruction}} \tag{5}$$

2.2.1 Contrastive Loss

The image and text modalities interact via a contrastive loss. First, the final groups for each modality are mapped into a common space with a Linear projector (Φ^T), i.e., $z_{ij}^T = \Phi^T(\bar{g}_{ij}^T)$. Then, we average pool over them to obtain the global features for each modality ($\hat{z}_i^T$). We compute the InfoNCE loss (Oord et al., 2018) for every modality separately. Given a batch size of B and a similarity function (sim), the infoNCE loss for the image to text is

$$L_{\text{I-T}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{e^{\text{sim}(\hat{z}_i^T,\hat{z}_i^I)/\tau}}{\sum_{j=1}^{B} e^{\text{sim}(\hat{z}_j^T,\hat{z}_i^I)/\tau}}, \tag{6}$$

and respectively for the text to image is

$$L_{\text{T-I}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{e^{\text{sim}(\hat{z}_i^T,\hat{z}_i^I)/\tau}}{\sum_{j=1}^{B} e^{\text{sim}(\hat{z}_i^T,\hat{z}_j^I)/\tau}}. \tag{7}$$

The final contrastive loss is calculated by averaging the two losses,

$$L_{\text{contrastive}} = \frac{1}{2}(L_{\text{I-T}} + L_{\text{T-I}}). \tag{8}$$

As for the similarity function sim(a,b), we consider the cosine similarity between the vectors.

2.2.2 Reconstruction Loss

In order to encourage the model to group the tokens into meaningful units, we incorporate a reconstruction loss from a text decoder. This loss encourages the model to assign tokens to different groups in order to spread information about the text across multiple vectors, and thus make better use of the available vectors.

We employ a simple shallow Transformer decoder to reconstruct the original input conditioned on the text groups. The shallow decoder has to rely on the information in the groups for decoding. Thus, it enforces the encoder to better encode the information into the groups (Bowman et al., 2015).

The output of this layer is

$$\overline{T}_i = \text{TransformerDecoder}(T_i|\{g_{i1}^T \dots g_{iK}^T\}). \tag{9}$$

The probabilities from these predictions are then used to define the reconstruction loss:

$$L_{\text{reconstruction}} = \sum_{i=1}^{B}\text{CE}(\overline{T}_i, T_i|\{g_{i1}^T, \dots, g_{iK}^T\}) \tag{10}$$

where CE is the cross entropy between the output probabilities of the decoder and the original input given the discovered groups.

3 Related Work

Our work is related to different tasks in vision and language, which we will explain in this section.

Object discovery. Here the task is to discover the objects in an image or video without any supervision. Slot-based object discovery (Locatello et al., 2020) has become popular due to the simplicity of the method (Singh et al., 2022; Sajjadi et al., 2022; Singh et al., 2023a; Seitzer et al., 2023; Singh et al., 2023b; Wu et al., 2023, 2024). We have a novel adaptation of this method in discovering units similar to phrases in language with visually grounded semantics.

Weakly supervised visual grounding. Visual grounding refers to the tasks where a phrase or expression is grounded in the image. In the weakly supervised setup, the only information used is the pairing of the image with its caption. In weakly supervised phrase grounding, the phrases are predetermined and no discovery happens on the language side (Datta et al., 2019; Gupta et al., 2020; Wang et al., 2020; Chen et al., 2022). In referring expression comprehension and referring image segmentation, the model must identify a specific part of the image described in a single expression. Kim et al. (2023) addressed the task of referring image segmentation by employing a slot-based object discovery module and merging relevant slots by cross attending over them with the textual query to build the final segmentation.

Vision language models with vision and language alignments. While many large-scale vision language models have been developed, it has been shown that they fall short in understanding fine-grained details in the image. This is especially more pronounced in the dual-stream Vision Language Models (VLMs), where the modalities interact only via a single-vector representation. Therefore, there has been efforts to align language and vision at the level of patches and tokens (Yao et al., 2022; Wang et al., 2022; Mukhoti et al., 2023). Zeng et al. (2022b) use additional supervision from the phrase grounding annotations to help the model learn the alignments. (Bica et al., 2024) aligns tokens and patch embeddings at different levels of granularity simultaneously. (Li et al., 2022) learns the semantic alignment from the perspective of game-theoretic interactions.

Object detection. The objective of this task is to detect the object boundaries in an image. Our work is related to query-based object detection, such as the approach in (Carion et al., 2020; Kamath et al., 2021), where, at decoding time, learnable object queries attend to the input features and encode an object. Liu et al. (2023) extend this approach by proposing a dual query model, demonstrating that simultaneously learning phrases and their corresponding objects improves the module's groundable understanding. The main difference between our model and this line of work lies in the weakly supervised nature of our approach.

Zero-shot open-vocabulary semantic segmentation. Semantic segmentation is a well-established task in computer vision. Recently, with the rise of VLMs, these models have demonstrated promising zero-shot capabilities in the semantic segmentation task as well. (Xu et al., 2022) propose a hierarchical grouping architecture that learns to group image regions without pixel-level annotations, relying solely on paired image and text data. Patel et al. (2023) expanded on image-text alignment, suggesting to not only align an image to the corresponding text but also to the text from visually similar samples. Additionally, Mukhoti et al. (2023) propose aligning patch tokens from a vision encoder with the <cls> token from a text encoder to enhance the model's performance.

Unit discovery in language. Lately, discovering language units as part of the model architecture has been explored. These models operate on top of characters, where the units are usually at the level of subwords or words. The purpose is to optimize model efficency (Dai et al., 2020; Nawrot et al., 2022, 2023; Sun et al., 2023) or to skip the tokenization step of preprocessing and develop an end-to-end model (Clark et al., 2022; Tay et al., 2022; Cao, 2023; Behjati and Henderson, 2023; Behjati et al., 2023). Our research aligns with these developments by also focusing on language unit discovery. However, it differs in that these units are semantically grounded to vision.

4 Experiments

In this section, empirically evaluate our proposed model. We will first evaluate the quality of the discovered segments quantitatively by their accordance with the groundable phrases in Section 4.4, and probe the fine-grained vision-language understanding of our proposed text encoder under two benchmarks in Section 4.2. Then, we show the effectiveness of our model in finding meaningful units by visualizing the attention maps in Section 4.3. We also analyse the contributions of different aspects of our model with a series of ablation studies in Section 4.5.

4.1 Experimental Setup

Datasets: We trained our models on the training split of GCC3M dataset which consists of around 3 million image-caption pairs collected from the web (Sharma et al., 2018). The average caption length in this dataset is 10.5 tokens. We will explain the datasets we used for evaluation in their corresponding sections.

Parameters: We first resize the images to 224×224 and then divide them into patches of size 16×16. The image encoder has 12 Transformer encoder layers with the hidden dimension of 384 and two grouping blocks at the 6th and 9th layers. The number of groups in the first block is 64 and 8 in the second block. We load the weights from the GroupViT released checkpoint[3] (Xu et al., 2022) and keep it frozen during training.

For the text encoder, we have 6 Transformer encoder layers followed by a grouping block[4] and then another 3 Transformer encoder layers. Each self-attention layer has 4 heads. We experiment with $K = 1, 2, 4, 8, 16$ as the number of groups. We report the performance and results of the model trained with 4 groups as it has the best performance, and study the effect of having different numbers of groups in our ablations. The text decoder has only 1 Transformer decoder layer consisting of one self-attention and one cross attention layer, each with 1 attention head. We tie the weights between the token embeddings in the encoder and the decoder. Both the encoder and the decoder have a model dimension of 128. The linear projection heads map each modality's feature vector to 256 dimension. We fix the τ to 0.07 in our contrastive losses and λ equals to 1. We use Byte Pair Encodings (Sennrich et al., 2016) as our tokenizer with a vocabulary size of around 50k tokens and the maximum number of tokens is set to $M = 77$ following previous work (Radford et al., 2021; Xu et al., 2022). We train our models with a batch-size of 4096 for 25 epochs and use the GradeCache library (Luyu Gao, 2021) to obtain this batch size on a single RTX3090 GPU[5]. We trained our models

[3] We take the checkpoint trained on GCC3M (Sharma et al., 2018), GCC12M (Changpinyo et al., 2021) and YFCC14M (Thomee et al., 2016) datasets.

[4] Our prilimenary experiments with two blocks did not lead to reasonable results.

[5] It takes around GPU 48 hours for every model to train.

Model	subj	verb	object	overall
random	50	50	50	50
groupvit	81.6	77.3	91.7	81.0
transformer	80.5	69.5	89.0	75.3
ours (4 groups)	80.3	70.1	90.4	76.0

Table 1: The zero-shot pairwise ranking accuracy of different models under SVO probes.

Model	accuracy
random	50
groupvit	82.5
transformer	80.91
ours (4 groups)	81.68

Table 2: The zero-shot performance of different models under the FOIL-COCO benchmark.

with AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 0.0016 with linear warmup for 2 epochs and cosine annealing decay.

Baselines: We compared our model against a text encoder with 9 Transformer layers, where the final text representation is taken from the <eos> token. This is the architecture used in GroupViT and other dual-stream vision-language models (Radford et al., 2021) and has approximately the same number of parameters as our proposed model. We train this model under the same training setup as our own model.

In addition, we report the results of the trained GroupViT model with its own text encoder and 2 layer projection heads. Note that this model has many more parameters and has been trained on 10x more data.

4.2 Fine-grained Vision-Language Understanding Probes

We evaluate the fine-grained vision and language understanding of our model by employing different benchmarks which are specifically designed for this purpose. We will explain each of these benchmarks and the zero-shot performance of our models in the following sections. In each case, the zero-shot classifier ranks the image-text pairs by their similarity scores $\mathrm{sim}(\hat{z}_j^T, \hat{z}_i^I)$, which is the cosine between the pooled embeddings on the image and text sides. We refer to the score obtained from this zero-shot classifier as pair-wise ranking accuracy.

4.2.1 SVO Probes

Hendricks and Nematzadeh (2021) designed a benchmark where they pair every sentence with two images, one positive and one negative. The negative images are selected in a controlled fashion where only either subject, verb or the object of the image is different from the original one. The test split of this dataset contains around 30k examples.

Table 1 shows the results of the zero-shot performance of different models under this benchmark. We observe that our model has a better overall performance compared to the Transformer baseline, which verifies our hypothesis that representing the language in a fine-grained and meaningful manner helps the fine-grained vision and language understanding of the model. The Transformer's single-vector representation succeeds in capturing information about subjects, but our multi-vector representation does a much better job of representing objects, and to a lesser extent verbs. Both of these models are well above the random baseline. The results for GroupViT's Transformer model are not comparable because it is trained on much more data, but we see that the resulting increase is much higher on verbs than on the the groundable phrases (subjects and objects) that our model is designed to represent as separate vectors.

4.2.2 FOIL-COCO

Shekhar et al. (2017) propose FOIL-COCO dataset where for every image there is a correct caption and a "foil" one. The foil caption is different from the original caption by altering one of the nouns in the original caption into a foil one. We evaluate the zero-shot performance of our model with pairwise ranking accuracy in Table 2 on the test split of this benchmark which has around 99k examples. We observe that our model demonstrates a remarkably good performance, outperforming the transformer model. This indicates that the noun understanding of our model has improved by learning fine-grained representations. Additionally, despite being trained on substantially less data than the GroupViT text encoder, our model performs nearly as well.

4.3 Attention Visualization

In order to understand what each group is representing, we visualize the soft attention weights of the groups over the input subwords in Figure 2. Interestingly, we can observe that contiguous segments

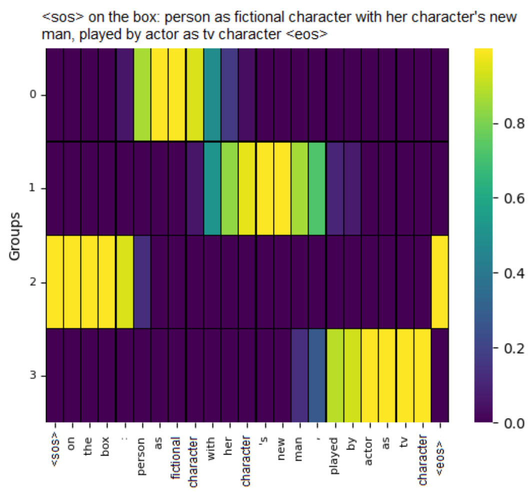

Figure 2: Soft attention of the groups over the input tokens. It shows that contiguous segments have emerged which capture phrase-like units.

have emerged, without imposing any contiguity constraints in the groupings. We believe that this is due to the fact that usually in language the contiguous tokens capture highly correlated information and that's why our model is grouping them together as part of its compression. Moreover, we can see that the emerging segments are meaningful in that they capture phrase-like units. We quantitatively evaluate the phrase discovery performance of our model in the following section (Section 4.4). In our examination of a sample of attention maps, we observe that a given group tends to bind to similar positions in the text, but that the boundaries between groups vary.

4.4 Zero-shot Segmentation Evaluation

In order to evaluate the emerging segments in the attention maps quantitatively, we propose a metric similar to Intersection-over-Union (IoU) in the visual object detection literature which we call "tIoU". We first compute the soft attention weights of the groups over the input tokens. Then, by taking the argmax over the inputs, we have an assignment matrix of every input to a group. Given a gold segmentation, we can compute the IoU for each discovered group of tokens and each gold segment. For the computation of IoU, the intersection is equal to the number of overlapping tokens. For the union, we do not count the tokens which were not annotated in the dataset, as the annotators did not have the constraint to include all the tokens in their annotation. This gives us a matrix where by

Model	tIoU	P	R	F1
random	42.15	61.51	60.03	54.54
k-means	52.77	61.82	64.87	59.55
spectral-clustering	38.88	49.81	52.82	45.52
mean shift	50.38	99.64	51.73	65.13
ours (4 groups)	**76.42**	**87.25**	**85.83**	**83.72**

Table 3: Phrase segmentation performance of different models under different evaluation metrics.

applying the Hungarian matching algorithm (Kuhn, 1955) maximizing this metric, we can obtain a 1-1 mapping between the discovered groupings and the gold segments. By having the mappings, we can compute precision, recall and F1 as well as IoU for each paired group and gold segment. In reporting the results, we first average every metric for the text input and then report the average over all examples.

For the gold segmentation, we use the annotations in Flickr30k Entities (Plummer et al., 2015) where groundable phrases are human-annotated. We report the results on the validation set of this dataset which has around 5000 examples. The number of annotated phrases in this dataset is on average 3.5.

In Table 3, we report the results of our evaluation. We compare our model against multiple baselines, including an untrained, randomly initialized model. We also report the performance of applying different clustering methods over the encoded features of our transformer baseline. In particular, we apply k-means, spectral clustering (Shi and Malik, 2000) and mean shift (Comaniciu and Meer, 2002) with 4 clusters. We observe that our model surpasses all the baselines by a large margin in all the metrics. Specifically, the high tIoU indicates that our model is indeed very good at discovering groundable phrases in the captions.

4.5 Ablation Study

In this section, we study the effect of different design choices on the performance of our models both in terms of groundable phrase discovery and fine-grained vision and language understanding.

4.5.1 Training Losses

In Table 4 we see the different effects of the two types of loss on our multi-vector model. Without the contrastive loss, the model has no training on the image-text relationship, so it is not surprising that the image-text semantic evaluations are very

model	tIoU	SVO subject	verb	object	overall	FOIL-COCO	Noun Understanding
ours	**76.42**	**80.3**	70.1	**90.4**	76.0	**81.68**	**84.12**
w/o contrastive loss	76.18	51.4	49.7	50.8	50.2	42.59	48.26
w/o reconstruction loss	40.80	78.2	**72.6**	89.1	**76.9**	78.66	81.98

Table 4: The performance of our model compared to the ablated ones on multiple datasets. Noun understanding refers to the average of performance on noun phrases (i.e. subjects, objects and FOIL-COCO).

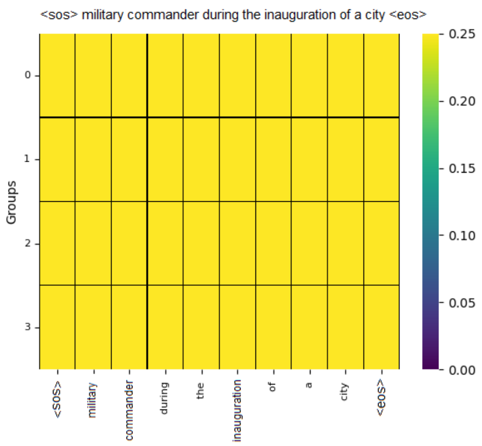

Figure 3: Soft attention of the groups over the input tokens for a model trained without the reconstruction loss. It shows a uniform attention map and lack of segmentation.

# of groups	tIoU	SVO	Foil
1	43.55	74.99	80.23
2	53.12	75.30	80.01
4	**76.42**	**76.0**	**81.68**
8	63.93	74.8	80.56
16	52.54	72.4	79.43

Table 5: The performance of our model trained with different number of groups.

low. More surprisingly, although it still segments in a meaningful way, without contrastive loss, the segmentation corresponds slightly less well to groundable phrases. This suggests that semantic grounding in images actually helps the model discover meaningful units of text.

Interestingly, without the reconstruction loss, the model fails to segment in a meaningful way. We can see this both in the tIoU score and in the uniform attention pattern shown in Figure 3. This lack of segmentation in turn affects the fine-grained understanding of the image-text relationship. The holistic representations indicated by Figure 3 are relatively good at representing verbs, because verb understanding combines information across multiple objects. But if we only consider the noun phrases (i.e. subjects, objects categories from SVO probes and FOIL-COCO), averaged in the last column, then segmenting the representation according to semantic objects, as indicated in Figure 2, results in much better understanding of the image-text relationship.

4.5.2 Number of Groups

In Table 5, we report the performance of our model trained with different numbers of groups. We can see that the model trained with 4 groups achieves the best results in all our evaluations. This implies that having too many or too few groups hurts the performance of our model.

5 Conclusions

In this work, we developed a novel model for discovering meaningful units that are semantically aligned to the objects in the image. We freezed an image encoder which outputs groups that approximately represent objects and employ an analogous architecture on the text side to discover units that are at the level of phrases. While many dual-stream VLMs represent text as a single vector, we hypothesise that learning to represent language at a finer granularity will improve their fine-grained vision and language understanding.

We verified our hypothesis by employing two specifically designed probing benchmarks, namely, SVO probes and FOIL COCO. In addition, we showed that the segments that appear in the attention maps of groups attending to tokens are meaningful both qualitatively and quantitatively, in term of overlapping with groundable phrases. Moreover, we ablated the effect of our losses on learning these units and concluded that both are necessary for having meaningful and semantically aligned units.

Limitations

We have performed our experiments on the datasets and benchmarks in English. However, we do not make any language dependent assumptions in developing our model. Therefore, we believe that our method is generalizable across other languages as long as enough data for training is available.

We were not able to perform our experiments at scale due to the computational limitations. We expect that training the image and text encoder simultaneously from scratch would lead to better alignment between the two modalities, which should in turn improve our results.

References

Melika Behjati, Fabio Fehr, and James Henderson. 2023. Learning to abstract with nonparametric variational information bottleneck. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1576–1586, Singapore. Association for Computational Linguistics.

Melika Behjati and James Henderson. 2023. Inducing meaningful units from character sequences with dynamic capacity slot attention. *Transactions on Machine Learning Research*.

Ioana Bica, Anastasija Ilić, Matthias Bauer, Goker Erdogan, Matko Bošnjak, Christos Kaplanis, Alexey A Gritsenko, Matthias Minderer, Charles Blundell, Razvan Pascanu, et al. 2024. Improving fine-grained understanding in image-text pre-training.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2015. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*.

Emanuele Bugliarello, Laurent Sartran, Aishwarya Agrawal, Lisa Anne Hendricks, and Aida Nematzadeh. 2023. Measuring progress in fine-grained vision-and-language understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1559–1582, Toronto, Canada. Association for Computational Linguistics.

Kris Cao. 2023. What is the best recipe for character-level encoder-only modelling? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5924–5938, Toronto, Canada. Association for Computational Linguistics.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*.

Keqin Chen, Richong Zhang, Samuel Mensah, and Yongyi Mao. 2022. Contrastive learning with expectation-maximization for weakly supervised phrase grounding. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8549–8559, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

Dorin Comaniciu and Peter Meer. 2002. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5):603–619.

Zihang Dai, Guokun Lai, Yiming Yang, and Quoc Le. 2020. Funnel-transformer: Filtering out sequential redundancy for efficient language processing. *Advances in neural information processing systems*, 33:4271–4282.

Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. 2019. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2601–2610.

Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. 2020. Contrastive learning for weakly supervised phrase grounding. In *European Conference on Computer Vision*, pages 752–768. Springer.

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing image-language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644, Online. Association for Computational Linguistics.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. *International Conference on Learning Representations (ICLR)*.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. Text encoders bottleneck compositionality in contrastive vision-language models. In *Proceedings of* the 2023 Conference on Empirical Methods in Natural Language Processing, pages 4933–4944, Singapore. Association for Computational Linguistics.

Dongwon Kim, Namyup Kim, Cuiling Lan, and Suha Kwak. 2023. Shatter and gather: Learning referring image segmentation with text supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15547–15557.

Harold W Kuhn. 1955. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Juncheng Li, Xin He, Longhui Wei, Long Qian, Linchao Zhu, Lingxi Xie, Yueting Zhuang, Qi Tian, and Siliang Tang. 2022. Fine-grained semantically aligned vision-language pre-training. *Advances in neural information processing systems*, 35:7290–7303.

Shilong Liu, Shijia Huang, Feng Li, Hao Zhang, Yaoyuan Liang, Hang Su, Jun Zhu, and Lei Zhang. 2023. Dq-detr: Dual query detection transformer for phrase extraction and grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1728–1736.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. 2020. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33:11525–11538.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR)*.

Jiawei Han Jamie Callan Luyu Gao, Yunyi Zhang. 2021. Scaling deep contrastive learning batch size under memory limited setup. In *Proceedings of the 6th Workshop on Representation Learning for NLP*.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. 2017. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations (ICLR)*.

Jishnu Mukhoti, Tsung-Yu Lin, Omid Poursaeed, Rui Wang, Ashish Shah, Philip HS Torr, and Ser-Nam Lim. 2023. Open vocabulary semantic segmentation with patch aligned contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19413–19423.

Piotr Nawrot, Jan Chorowski, Adrian Lancucki, and Edoardo Maria Ponti. 2023. Efficient transformers with dynamic token pooling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6403–6417, Toronto, Canada. Association for Computational Linguistics.

Piotr Nawrot, Szymon Tworkowski, Michał Tyrolski, Lukasz Kaiser, Yuhuai Wu, Christian Szegedy, and Henryk Michalewski. 2022. Hierarchical transformers are more efficient language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1559–1571, Seattle, United States. Association for Computational Linguistics.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Yash Patel, Yusheng Xie, Yi Zhu, Srikar Appalaraju, and R Manmatha. 2023. Simcon loss with multiple views for text supervised semantic segmentation. *arXiv preprint arXiv:2302.03432*.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. 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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Mehdi SM Sajjadi, Daniel Duckworth, Aravindh Mahendran, Sjoerd van Steenkiste, Filip Pavetic, Mario Lucic, Leonidas J Guibas, Klaus Greff, and Thomas Kipf. 2022. Object scene representation transformer. *Advances in Neural Information Processing Systems*, 35:9512–9524.

Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, and Francesco Locatello. 2023. Bridging the gap to real-world object-centric learning. In *The Eleventh International Conference on Learning Representations*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

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

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. FOIL it! find one mismatch between image and language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265, Vancouver, Canada. Association for Computational Linguistics.

Jianbo Shi and Jitendra Malik. 2000. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905.

Gautam Singh, Fei Deng, and Sungjin Ahn. 2022. Illiterate dall-e learns to compose. In *International Conference on Learning Representations*.

Gautam Singh, Yeongbin Kim, and Sungjin Ahn. 2023a. Neural systematic binder. In *The Eleventh International Conference on Learning Representations*.

Gautam Singh, Yeongbin Kim, and Sungjin Ahn. 2023b. Neural systematic binder. In *The Eleventh International Conference on Learning Representations*.

Li Sun, Florian Luisier, Kayhan Batmanghelich, Dinei Florencio, and Cha Zhang. 2023. From characters to words: Hierarchical pre-trained language model for open-vocabulary language understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3605–3620, Toronto, Canada. Association for Computational Linguistics.

Yi Tay, Vinh Q. Tran, Sebastian Ruder, Jai Gupta, Hyung Won Chung, Dara Bahri, Zhen Qin, Simon Baumgartner, Cong Yu, and Donald Metzler. 2022. Charformer: Fast character transformers via gradient-based subword tokenization. In *International Conference on Learning Representations*.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73.

Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. 2022. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in Neural Information Processing Systems*, 35:33536–33549.

Qinxin Wang, Hao Tan, Sheng Shen, Michael Mahoney, and Zhewei Yao. 2020. MAF: Multimodal alignment framework for weakly-supervised phrase grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2030–2038.

Yi-Fu Wu, Klaus Greff, Gamaleldin Fathy Elsayed, Michael Curtis Mozer, Thomas Kipf, and Sjoerd van Steenkiste. 2023. Inverted-attention transformers can learn object representations: Insights from slot attention. In *Causal Representation Learning Workshop at NeurIPS 2023*.

Yi-Fu Wu, Minseung Lee, and Sungjin Ahn. 2024. Structured world modeling via semantic vector quantization. *arXiv preprint arXiv:2402.01203*.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. 2022. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2022. Filip: Fine-grained interactive language-image pre-training. In *The Eleventh International Conference on Learning Representations*.

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? In *The Eleventh International Conference on Learning Representations*.

Yan Zeng, Xinsong Zhang, and Hang Li. 2022a. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *Proceedings of the 39th International Conference on Machine Learning*.

Yan Zeng, Xinsong Zhang, and Hang Li. 2022b. Multi-grained vision language pre-training: Aligning texts with visual concepts. *International Conference on Machine Learning*.

A Artifacts statements

The datasets used do not have personally identifying information or offensive content. We provide the list of datasets used and the corresponding licenses in Table 7, which are all consistent with our academic use.

B Descriptive Statistics

Our results are from single runs for all the models trained.

C Packages

We provide a list of packages used in our code in Table 6.

D AI Assistants

We utilized AI assistants for minor text editing and code completion tasks during the development of the model.

Package	version
Python	3.7
PyTorch	1.8
webdataset	0.1.103
mmsegmentation	0.18.0
timm	0.4.12
nltk	3.8.1
ftfy	6.1.1
regex	2023.6.3

Table 6: The packages used in our code development

.

Dataset	License
GCC3M	Google license (link)
SVO-Probes	Creative Commons Attribution 4.0 International Public License (CC BY 4.0)
FOIL-COCO	Creative Commons Attribution 4.0 License
Flikr	Creative Commons Attribution 0: Public Domain

Table 7: Datasets and their licenses.