VladVA: Discriminative Fine-tuning of LVLMs

Yassine Ouali*[1] Adrian Bulat*[1,2] Alexandros Xenos[1,3] Anestis Zaganidis[1]
Ioannis Maniadis Metaxas[1] Brais Martinez[1] Georgios Tzimiropoulos[1,3]
[1]Samsung AI Cambridge [2]Technical University of Iasi [3]Queen Mary University of London

Abstract

Contrastively-trained Vision-Language Models (VLMs) like CLIP have become the de facto approach for discriminative vision-language representation learning. However, these models have limited language understanding, often exhibiting a "bag of words" behavior. At the same time, Large Vision-Language Models (LVLMs), which combine vision encoders with LLMs, have been shown to be capable of detailed vision-language reasoning, yet their autoregressive nature renders them less suitable for discriminative tasks.

In this work, we propose to combine "the best of both worlds": a new training approach for discriminative fine-tuning of LVLMs that results in strong discriminative and compositional capabilities. Essentially, our approach converts a generative LVLM into a discriminative one, unlocking its capability for powerful image-text discrimination combined with enhanced language understanding.

Our contributions include (1) A carefully designed training/optimization framework that utilizes image-text pairs of variable length and granularity for training the model with both contrastive and next-token prediction losses. This is accompanied by ablation studies that justify the necessity of our framework's components. (2) A parameter-efficient adaptation method using a combination of soft prompting and LoRA adapters. (3) Significant improvements over state-of-the-art CLIP-like models of similar size, including standard image-text retrieval benchmarks and notable gains in compositionality.

1. Introduction

Contrastively-trained Vision Language Models (VLMs) (e.g. CLIP [40]) have become the predominant direction for vision-language representation learning, exhibiting remarkable zero-shot abilities [20, 29, 32, 40, 53]. However, the great success of these models in many vision-language and vision tasks, even in a zero-shot manner, "sweeps under the rug" some of their important limitations. Specifically, such models struggle to exhibit advanced language understanding capabilities, suffer from a limited understanding of compositionality, and manifest a *bag of words* behavior [26, 52]. For example, even with *bag of words* behavior, VLMs have shown remarkable zero-shot retrieval accuracy on the Flickr [51] and COCO [34] datasets. Still, they perform poorly on a simple word order permutation task on the same datasets [52]. Unfortunately, these issues persist even when the model and the dataset size increase [18].

Concomitantly, inspired by the success of LLMs [5, 48] in acting as generalist assistants [14], a series of works combine pretrained vision encoders and LLMs [27, 28, 55] to construct Large Vision-Language Models (LVLMs) capable of performing interactive multi-modal conversations. Among others, these models have been shown capable of *exhibiting strong reasoning and vision-language understanding capabilities*, offering fine-grained and detailed responses [10, 12, 27, 28]. However, they are trained with a next-token prediction loss in an autoregressive manner, which appears less suitable for direct utilization in discriminative image-text tasks (*e.g.* image-text retrieval).

To our knowledge, the very recent (concurrent) work [22] is the first one to show that, with appropriate prompting, LVLMs can serve as zero-shot discriminative models. Importantly, [22] advocates for a text-text optimization approach, stating that contrastive image-text fine-tuning has a detrimental effect on the model's performance. In contrast to [22], we propose a new training framework for discriminative image-text fine-tuning of LVLMs, aiming to convert the original *generative* LVLM into a *discriminative* one, thereby significantly enhancing its capability for image-text discrimination while preserving the compositional strengths of the original model.

In our approach, following the (independent) two-towers paradigm, the vision embeddings are produced by passing the image through the entire LVLM, and the text embeddings by passing the text through the LLM of the LVLM. Intuitively, for the vision embedding, the LLM acts as an information processor that refines the visual information while simultaneously aligning it with the textual representations. We coin our approach **VladVA**: **V**ision-**L**anguage **A**daptation for **D**iscriminative **V**isual **A**ssistant. Our **main contributions** are:

- We devise a carefully designed optimization framework that utilizes image-text pairs of variable length and granularity for model training (*i.e*. both short and long captions). Using this data, the model is trained with both contrastive and next-token prediction losses, which are both shown to be necessary for unlocking strong discrimination and compositionality capabilities. Our design choices are accompanied by ablation studies, which justify the necessity of our framework's components.
- To facilitate efficient training, we show how the model can be fine-tuned using a parameter-efficient adaptation method based on a combination of soft prompting [31] and LoRA adapters [19]. We show the positive impact of both components.
- We report significant improvements over state-of-the-art two-tower models (e.g. CLIP-like models) of similar size on standard image-text retrieval benchmarks (+4.7-7.0% gains in absolute terms). Moreover, we report notable gains on several vision-language understanding and compositionality benchmarks (up to +15%).

2. Related work

2.1. Large Vision Language Models (LVLMs)

Inspired by breakthrough research in language modeling [5, 21, 46, 48], a series of methods seek to combine pretrained LLMs and vision encoders to construct Large Vision Language Models (LVLMs) capable of processing image-text data jointly [3, 12, 33, 35, 36, 49, 50, 55]. The prevalent strategy consists in aligning the features produced by a pretrained vision encoder to the textual space assumed by a pretrained LLM using a projection module, *e.g*. LLaVA [36], following a two-stage alignment procedure. Follow-up works expand this to interleaved image-text data [1, 28] and multiple input crops [1] while seeking to improve the model's efficiency [11].

Despite their strong generative and comprehension abilities [35], current LVLMs are primarily restricted to generative tasks. Only very recently, Jiang *et al*. [22], inspired by the recent progress in NLP [4, 25] adapted a LLaVA-NeXT [27] model to discriminative tasks using a contrastive-like loss and text data only. We note that unlike [22], we introduce a training framework that learns from multi-turn image-text pairs (as opposed to text only) using a novel formulation that jointly combines a contrastive loss with a next-token prediction, reflecting the data characteristics and inducing a gradual representation buildup. Concurrently, VLM2Vec [23] adapts an LVLM for multi-modal retrieval. However, it uses a different loss and training strategy (no generative loss, no short-long captions training, no soft prompting). We compare our approach with both E5-V and VLM2Vec, significantly improving upon their results despite using smaller/lighter models.

2.2. Discriminative Vision-Language Models

The prevalent approach for training Discriminative VLMs follows the two-tower contrastive approach pioneered by CLIP [40], whereby an image and text encoder are trained on web-collected image-text pairs to learn a joint multimodal (*i.e*. vision and language) space. Subsequent works build upon CLIP by scaling the data [7, 42, 53], improving the architecture using late/early interactions [30] or improving the training loss [8, 53]. Despite their remarkable zero-shot and representation learning abilities [40] such models were shown to have significant shortcomings related to limited language understanding capabilities, including: lack of compositionality understanding [26], manifesting *bag of words* behavior [52], struggling with spatial relations [26], being susceptible to typographical attacks [16], *etc*. Recent works aim to address these shortcomings by constructing synthetic hard negatives [52] or performing cross-modality attention [30]. However, the former does not inherently change the model's behaviors and has been shown to potentially learn a series of shortcuts/artifacts [18]. Meanwhile, the latter is impractical for deployment at scale, as, due to the interactions between the encoders, each new query incurs an additional inference for every image within the set.

To alleviate these shortcomings and improve the overall capabilities of such models, we depart from the prevalent approach of training VLMs using a contrastive loss and, instead, propose a new approach that seeks to convert generative LVLMs into discriminative models by adapting them using a newly proposed framework that combines generative and discriminative objectives.

3. Method

Herein, we present VladVA (**V**ision-**L**anguage **A**daptation for **D**iscriminative **V**isual **A**ssistant), our novel approach for discriminative fine-tuning of LVLMs that results in strong discriminative and compositional capabilities. This section is structured as follows: Sec. 3.1 briefly introduces the architecture, detailing how LVLMs can be used as discriminators in a zero-shot manner. Sec. 3.2 details the core component of our approach: a carefully designed optimization framework that utilizes image-text pairs of variable length and granularity for training the model with both contrastive (Sec. 3.2.1) and next-token prediction (Sec. 3.2.2) losses, showcasing that contrastive is best with short captions and autoregressive with long captions. In Sec. 3.3, we present our parameter-efficient adaptation, while Sec. 3.4 analyzes how the model's behavior changes after training.

3.1. LVLMs as zero-shot discriminative models

LVLMs consist of an LLM Φ_t, a vision encoder Φ_v, and a module g that projects the vision features into the LLM's textual space. Once fine-tuned, such models can produce

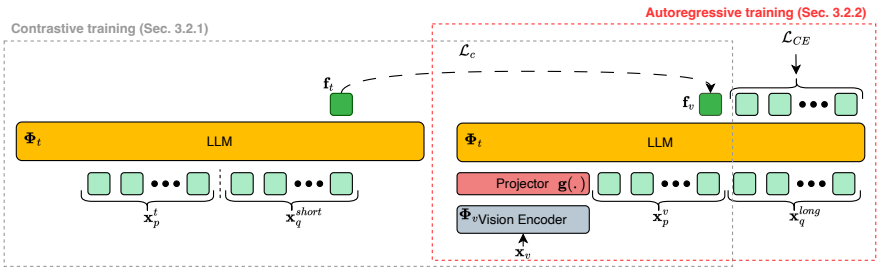

Figure 1. **Overall VladVA framework:** a generative LVLM is adapted into a discriminative model with the help of (1) a contrastive training loss (Sec. 3.2.1), and (2) an autoregressive loss (Sec. 3.2.2). The first one is applied on image-text pairs with short(er) captions, encouraging the last token produced by both modalities to be discriminative. The second one, jointly optimized with the first one, is applied only on longer captions and allows the model to learn fine-grained details.

a textual answer $\mathbf{x}_a = \Phi_t(g(\Phi_v(\mathbf{x}_v)), \mathbf{x}_q)$ when presented with an input image $\mathbf{x}_v$ and a text query (or prompt) $\mathbf{x}_q$.

Despite being solely trained with an autoregressive next-token prediction loss on limited amounts of data (< 5M), such models can act as multi-modal discriminative models in a zero-shot manner [22]. To elicit this capability, the image embedding $\mathbf{f}_v = \Phi_t(g(\Phi_v(\mathbf{x}_v), \mathbf{x}_p^v))[eos]$ is obtained by passing the image alongside a handcrafted image prompt $\mathbf{x}_p^v$ (e.g., "in one word, describe the image") through the LVLM and taking the output representation of the last token. Analogously, the text embedding $\mathbf{f}_t = \Phi_t(\mathbf{x}_p^t, \mathbf{x}_q)[eos]$ is produced by passing the handcrafted text prompt $\mathbf{x}_p^t$ (e.g., "in one word, describe the text") and input query $\mathbf{x}_q$ through the LLM (of the LVLM) and taking again the output representation of the last token. We will refer to these particular tokens as "summary tokens" (summarizing image and text information, respectively). Note that, typically, the respective handcrafted prompts for the image ($\mathbf{x}_p^v$) and text ($\mathbf{x}_p^t$) modalities are different. Finally, the similarity between an image and a text query can be computed by taking the cosine similarity between the two: $s = \texttt{cos\_sim}(\mathbf{f}_v, \mathbf{f}_t)$.

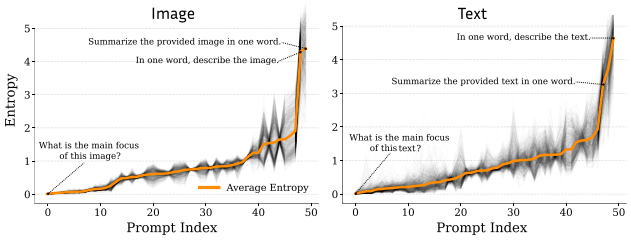

Figure 2. **Entropy of the output probability distribution** at the next-to-be-predicted token location using a LLaVA-1.5-7B for a set of 50 prompts for both images and captions.

What makes a good prompt? Zero-shot adaptation by prompting already provides decent results despite the task changing from generation to discrimination. To shed some light, herein, we study (a) what makes a good prompt and (b) how we can identify it.

To answer these questions, we construct a testbed consisting of 1,000 image-caption pairs from Flickr30k [51], which we then use to evaluate the quality of various prompts. The prompts (50 image-text pairs in total) are constructed using ChatGPT. Each prompt pair is fed, alongside an image and its respective caption, through the LLaVA-1.5-7B model. For each image-prompt pair and caption-prompt pair, we extract the token embedding at the output position and the corresponding output probability distribution over the vocabulary. These are then used to compute two metrics for each prompt: the average entropy of its output distributions and the cumulative variance of its embeddings. As Figs. 2 and 4 show, when the model is prompted with sentences consisting of specific keywords, such as `in a few words` or in `one word`, the model is pushed to condense the information of the image or text in the next token, resulting in an output distribution with high entropy. More importantly, when investigating the generated embeddings, we observe that higher entropy prompts result in embeddings with more spread-out cumulative variance, *i.e.* requiring more principal components to capture the same amount of variance, indicating an embedding matrix with a high rank (see Fig. 3). This translates into discriminative embeddings that can capture more information about the inputs, making them suitable for embedding tasks. The benefit of this behavior is illustrated in Fig. 5, which

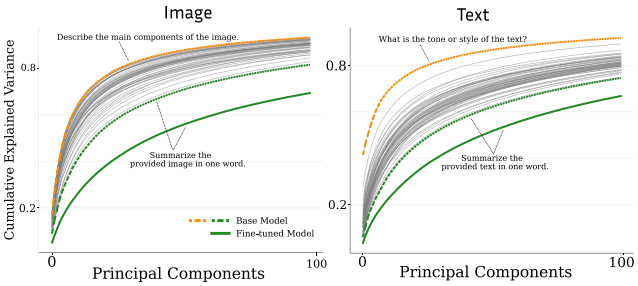

Figure 3. **Cumulative variance of the image and text embedding matrices** over a set of 50 prompts on Flickr30k. Embeddings that capture more information about the input translate into a cumulative variance that requires more principal components to be explained, *i.e.* a higher-rank embedding matrix.

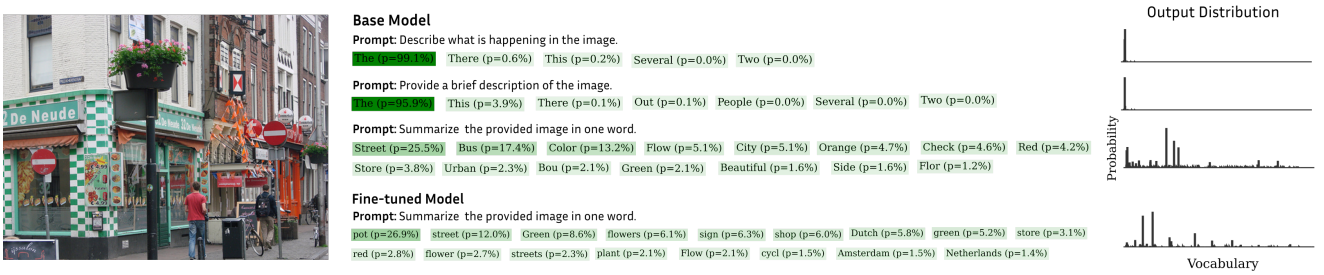

Figure 4. **Top-k next-to-be-predicted tokens** before and after VladVA fine-tuning (our approach). On the right, we show the output probability distribution for each case. When using the best prompt ("Summarize the provided image in one word"), the representations of the next token can encode diverse and more discriminative information, making potentially better-quality embeddings. This behavior is further improved after VladVA fine-tuning.

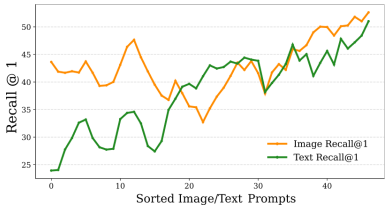

Figure 5. **Image and text retrieval score on Flickr30k** over a set of 50 image-text prompts ordered by their entropy scores (Fig. 2). We can observe that prompts with high average entropy scores correlate positively with the zero-shot retrieval performance.

shows a positive correlation between prompts with high entropy scores and the model's zero-shot retrieval performance. Hence, our approach should seek to produce embeddings with a) spread-out variance and b) probability distributions over the vocabulary with increased entropy.

3.2. Discriminative fine-tuning of LVLMs: from generation to discrimination

Despite exhibiting surprising innate zero-shot abilities, LVLM's direct discriminative performance lags behind that of state-of-the-art contrastively trained VLMs. Hence, carefully designed frameworks are needed to unlock the full potential of such models. This is the very goal of our work: to introduce a well-grounded adaptation/training framework that surfaces the discriminative image-text capabilities of a generative LVLM.

Notably, our findings contradict those of the very recent work of [22], which found that contrastive image-text fine-tuning is detrimental and limits training to text-text contrastive learning alone. This highlights the importance of our proposed approach, which overcomes such impediments and significantly boosts the discriminative performance of the model.

Having established the architecture in the previous section, the two other pillars are the data and training strategy.
Data strategy: We argue for the importance of data diversity in terms of granularity and group captions according to their length: short captions (< 30 tokens) and long captions ($30 - 500$ tokens). The short captions capture coarse details and summarize image content teaching the model to discriminate with regard to high-level image information. Longer captions capture finer image details and promote a better understanding of language concepts such as spatial relationships and compositionality. For a strong discriminative model, both are necessary. Therefore, for images missing either caption type, we use a BLIP2 [30] captioner to generate short captions and ShareGPT-4V [9] to generate long captions. This allows us to leverage both supervisory signals for training.
Training strategy: As we demonstrate in this work, the variable length of the training data poses its own challenges: unlike the case of short captions, where training using the well-studied contrastive loss performs well, it collapses for longer captions. This brings us to the proposed training strategy, whereby, to address this challenge, we propose a hybrid training approach that combines a contrastive loss (see Sec. 3.2.1) and a *next-token prediction loss for discriminative adaptation* (see Sec. 3.2.2). Finally, as full model fine-tuning is computationally expensive, in Sec. 3.3, we detail a fine-tuning strategy that combines adapters with soft prompting.

3.2.1. Image-text contrastive alignment

Under a multi-modal contrastive formulation, the image and text representations, $\mathbf{f}_v$ and $\mathbf{f}_t$ respectively, must be close if they are semantically similar and far apart otherwise, under a specified distance metric. At train time, this is enforced using a symmetric image-text and text-image contrastive loss, which, for a given mini-batch containing b randomly selected samples, can be described as:

$$\mathcal{L}_c = \frac{1}{b}\sum_{k=1}^{b}\left(-\log\frac{\exp(s_v^{k,k})}{\sum_j \exp(s_v^{k,j})} - \log\frac{\exp(s_t^{k,k})}{\sum_j \exp(s_t^{j,k})}\right), \tag{1}$$

where $s_v^{k,j} = \texttt{cos\_sim}(\mathbf{f}_v^k, \mathbf{f}_t^j)$ denotes the cosine similarity between the k-th image and the j-th caption (image-to-text), and similarity, $s_t^{k,j}$ the text-to-image similarity.

During training, the contrastive loss is applied to the very same tokens used for the zero-shot evaluation, as they represent the optimal starting point for further fine-tuning (Sec. 3.1). We note that the contrastive loss is mostly suitable for training using short captions $\mathbf{x}_q^{short}$ (*i.e.* < 30 tokens), like the ones typically used for CLIP pre-training. We found that training the model using a contrastive loss on longer captions proves challenging. Hence, to address this, in the following section, we study and propose a new formulation that enables discriminative training on variable-length data.

3.2.2. Autoregressive training for learning discriminative LVLM representations

Until now, the modality-specific embeddings are obtained by taking the last token, prior to any generation, while the training is largely focused on short (*i.e.* < 30 tokens) captions, mimicking the CLIP-style data used for contrastive training. This contrasts with the LLaVA-style autoregressive training, where long and highly descriptive captions (typically 200–500 tokens) are used to help the LVLM learn strong links between the vision and text domains, pay attention to fine-grained details, and develop strong reasoning and compositionality capabilities.

As noted earlier, directly using the long captions with the contrastive loss is ineffective, as, due to the high specificity of the long captions, the task is easy and nearly trivial to solve, with the loss going to 0 in just a few hundred iterations. To address this, we propose to instead apply the next-token prediction loss over the long captions:

$$\mathcal{L}_{CE} = \sum_{i=1}^{L} \log p_\theta(u_i | \mathbf{x}_v, \mathbf{x}_p^v, \mathbf{x}_{q,<i}^{long}), \tag{2}$$

where L is the length of the long caption $\mathbf{x}_q^{long}$, $\mathbf{x}_v$ the input image, and $\mathbf{x}_p^v$ the prompt which prompts the model to describe the image in detail (e.g., "Describe the image in detail"), and p_θ the next-token probability distribution learned by the model.

Intuitively, this formulation possesses multiple advantages: (1) It allows the model to learn from long captions, as predicting each and every token correctly is a challenging task (as opposed to applying the contrastive loss to long captions); (2) The decoding process encourages the condensation of information into the starting token used as a feature embedding; (3) It offers an avenue for retaining the generative capabilities of the model while strengthening its discriminative abilities.

3.2.3. Overall training loss

As depicted in Fig. 1, we apply the next-token prediction loss over the long captions and the contrastive loss over the short ones in a unified manner. During training, the templates presented to the LVLM for the image and text modality take the following form:

Template for the image modality:
`USER: [Image Prompt] <image>`
`ASSISTANT: <out_token>`
`USER: Describe the image in detail.`
`ASSISTANT: <long_caption>`

Template for the text modality:
`USER: [Text Prompt] <short_caption>`
`ASSISTANT: <out_token>`

with the contrastive loss applied on the output representations `<out_token>` for the image modality and `<out_token>` for the text modality. Concomitantly, the next-token prediction loss is applied on the tokens of the `<long_caption>`. Generally, the short caption must be sufficiently different from the long caption to prevent shortcuts during training, a property that naturally emerges in our case due to the difference in length and annotation procedure. Note that the distinction between long and short captions is made only during training. At test time, the model is used in discriminative mode as detailed in Sec. 3.1.

3.3. Parameter-efficient adaptation

As direct fine-tuning of the LVLM is costly, especially when maintaining a reasonably large batch size for contrastive learning, herein, we adopt parameter-efficient training with soft-prompting *combined with* LoRA adapters, both trained under the same loss formulation of Sec. 3.2.

Soft prompting was recently proposed as an efficient task-adaptation approach for both LLM [31] and CLIP [6, 54] models, representing a direct departure from the prompt hand-crafting solution. Specifically, for a given input modality, *i.e.* image and text, we define a set of n modality(m)-specific learnable vectors $[\mathbf{v}_1^m, \mathbf{v}_2^m, \cdots, \mathbf{v}_n^m]$, $\mathbf{v}_i^m \in \mathbb{R}^C$ with C denoting the model's vocabulary embedding size. These vectors can be inserted across the input sequence to adjust the model's behavior. In practice, we opt to replace the tokens belonging to the hard prompts (*i.e.* $\mathbf{x}_p^v$ and $\mathbf{x}_p^t$; see Sec. 3.1) with the learnable vectors, initializing their values with the embeddings of the handcrafted ones.

Adapter fine-tuning: While efficient, the representation power of the soft prompts is somewhat limited. Hence, following best practices, we also attach LoRA [19] adapters to the linear layers located inside Φ_t. Such adapters offer a multifold advantage: lower memory requirements, reduced potential of overfitting during training, and no additional compute requirements during inference.

The model is fine-tuned using these components. Importantly, both have a positive impact on overall accuracy.

3.4. How does the model's behavior change?

Building upon the analysis from Sec. 3.1, we show that our training approach elicits the following *behavioral* changes:

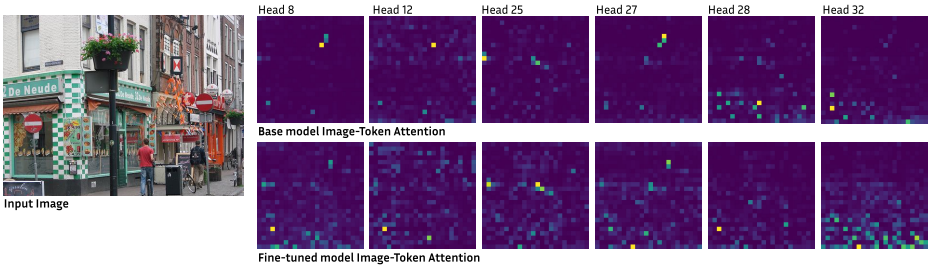

Figure 6. **Attention map between the summary and vision tokens shown for a set of heads**. Notice that post-training, the attention maps densify. This behavioral change can be interpreted as follows: For generative tasks, at every step in the generation process, the model has the chance to look back at the vision tokens, selectively attending to the regions of interest at the current step. In contrast, in a discriminative setting, the model must compress all information present in the image within the summary token.

Table 1. Zero-shot text-image retrieval accuracy on Flickr30K, COCO and nocaps.

Method	**image** retrieval						**text** retrieval					
	Flickr30K		COCO		nocaps		Flickr30K		COCO		nocaps	
	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10
CLIP (`ViT-L`) [40]	67.3	93.3	37.0	71.5	48.6	85.7	87.2	99.4	58.1	87.8	70.0	96.2
BLIP (`ViT-L`) [29]	70.0	95.2	48.4	83.2	62.3	93.4	75.5	97.7	63.5	92.5	72.1	97.7
BLIP2 (`ViT-L`) [30]	74.5	97.2	50.0	86.1	63.0	93.8	86.1	99.4	63.0	93.1	74.4	98.3
OpenCLIP (`ViT-G/14`) [42]	77.8	96.9	48.8	81.5	63.7	93.2	91.5	99.6	66.3	91.8	81.0	98.7
OpenCLIP (`ViT-BigG/14`) [42]	79.5	97.1	51.3	83.0	65.1	93.5	92.9	97.1	67.3	92.6	82.3	98.8
EVA-02-CLIP (`ViT-E/14+`) [44]	78.8	96.8	51.1	82.7	64.5	92.9	93.9	99.8	68.8	92.8	83.0	98.9
EVA-CLIP (`8B`) [45]	80.3	97.2	52.0	82.9	65.3	93.2	94.5	99.7	70.1	93.1	83.5	98.6
EVA-CLIP (`18B`) [45]	83.3	97.9	55.6	85.2	69.3	94.8	**95.3**	99.8	72.8	94.2	85.6	99.2
LLaVA-1.5-7B [35]	59.6	89.3	34.4	69.6	46.9	83.3	65.6	92.3	35.6	70.5	52.1	88.1
VLM2Vec (`Mistral-7B`) [23]	80.1	97.3	52.0	85.6	65.9	94.5	90.3	99.6	68.2	93.2	79.2	98.5
E5-V (`LLaVA-Next-8B`) [22]	79.5	97.6	52.0	84.7	65.9	94.3	88.2	99.4	62.0	89.7	74.9	98.3
E5-V (`LLaVA-1.5-7B`) [22]	76.7	96.9	48.2	82.1	62.0	93.0	86.6	99.0	57.4	88.4	71.9	97.0
VladVA (Ours) (`LLaVA-1.5-7B`)	**85.0**	**98.5**	**59.0**	**88.6**	**72.3**	**96.5**	94.3	**99.9**	**72.9**	**94.4**	**85.7**	**99.5**

(1) The attention map between the summary and vision tokens increases in density. (2) Both the entropy of the output distribution of the summary token and the spread of the cumulative variance of the embeddings increase.

The attention map densification, as exemplified in Fig. 6, shows that, for discriminative tasks, the model gathers evidence from all parts of the image in order to correctly encode the information therein. This is not needed for generation, as at every generation step, the model can "peak back" at the vision tokens and select the required information.

Entropy and cumulative variance: As shown in Fig. 3, our approach results in models where the cumulative variance of the image and text embeddings is significantly more spread out, which translates into richer and better-aligned embeddings, capable of more accurately capturing fine-grained details. Additionally, the model maintains the diversity of output distribution at the summary token, *i.e*. high entropy, as illustrated in Fig. 4.

4. Experiments

We compare our approach with the current state-of-the-art on two tasks of interest in a zero-shot manner: image-text retrieval and compositionality/language understanding.

Models compared: We compare with state-of-the-art models based on the two-towers (independent) approach, which is practical for retrieval purposes and also followed by our method. We cover a wide variety of settings: different models and model sizes, training data, training losses, etc.: CLIP (`ViT-L`) [40] – the original CLIP trained with a contrastive loss on 400M image-text pairs; BLIP (`ViT-L`) [29] – trained on over 120M samples using contrastive, captioning and image-text matching losses; BLIP2 (`T5-XXL`) – improved and scaled-up version of BLIP; OpenCLIP (`ViT-G/14`) [42] – scaled-up version of [40] trained on 2B samples; OpenCLIP (`ViT-BigG/14`) [42], EVA-02-CLIP (`ViT-E/14+`) [44], EVA-CLIP (`8B`) [45] and EVA-CLIP (`18B`) [45] – large contrastively trained models, with up to 18B parameters, fine-tuned from vision encoders trained with Masked Image Modeling (MIM); E5-V (`LLaVA-Next-8B`) and E5-V (`LLaVA-1.5-7B`) – LVLMs finetuned using a text-text contrastive loss. Depending on the task, we also include additional specialized baselines (*e.g*. NegCLIP [52] for compositionality).

Training details: We use a LLaVA-1.5 (7B) [36] model due to its popularity and simplicity (for other models, see supp. material). For LoRA adapters, we set the rank and α to 16.

Table 2. Comparison with state-of-the-art on the SugarCrepe compositionality benchmark.

Method	Params	Replace			Swap		Add	
	(B)	Object	Attribute	Relation	Object	Attribute	Object	Attribute
Human	–	100	99	97	99	100	99	99
CLIP (ViT-L) [40]	0.43	94.1	79.2	65.2	60.2	62.3	78.3	71.5
BLIP (ViT-L) [29]	0.23	96.5	81.7	69.1	66.6	76.8	92.0	85.1
BLIP2 (ViT-L) [30]	1.17	97.6	81.7	77.8	62.1	65.5	92.4	87.4
OpenCLIP (ViT-G/14) [42]	1.37	95.8	85.0	72.4	63.0	71.2	91.5	82.1
OpenCLIP (ViT-BigG/14) [42]	2.54	96.6	87.9	74.9	62.5	75.2	92.2	84.5
EVA-02-CLIP (ViT-E/14+) [44]	5.04	97.1	88.5	74.2	67.3	74.1	91.8	83.9
EVA-CLIP (8B) [45]	8.22	96.4	86.6	74.8	66.1	74.6	91.3	82.0
EVA-CLIP ((18B) [45]	18.3	97.5	88.8	76.1	65.3	76.0	92.4	85.0
NegCLIP [52]	0.15	92.7	85.9	76.5	75.2	75.4	88.8	82.8
LLaVA-1.5-7B [35]	7.06	88.0	81.6	76.1	60.9	58.8	67.0	62.4
VLM2Vec (Mistral-7B) [23]	7.30	97.2	89.0	81.7	62.9	72.5	94.7	88.6
E5-V (LLaVA-Next-8B) [22]	8.36	96.7	89.5	85.3	75.0	70.1	89.2	83.5
E5-V (LLaVA-1.5-7B) [22]	7.06	95.8	86.6	81.6	62.9	64.0	93.5	88.0
VladVA (Ours) (LLaVA-1.5-7B)	7.06	**98.1**	**92.1**	**86.8**	**79.0**	**82.9**	**95.2**	**95.8**

The number of soft prompts is aligned to the length of the tokenized hand-crafted prompt. Unless otherwise stated, we train the models for 7 epochs, using a batch size of 1024, a learning rate of $1e-4$, no weight decay, and AdamW [37] optimizer with default values for β_1 and β_2. During training, the learning rate is decayed according to a cosine scheduler [38]. Depending on the data configuration, we use up to 32 A100 GPUs. All our models and training procedures were implemented using PyTorch [39] and DeepSpeed [41].

We used the following training data: a 4M random subset of OpenImages [24], CC3M (∼2.8M images) [43], and ShareGPT-4V [9]. As no captions are available for OpenImages, we automatically label them with 5 captions using BLIP2 [30]. During training, only one caption is sampled at a time. For longer captions, we directly use the ShareGPT-4V [9] data, which we extend with synthetic short captions produced by BLIP2 in order to enable the training procedure proposed in Sec. 3.2.3. Similarly, CC3M is automatically annotated with long captions using ShareGPT4-V [9].

4.1. Zero-shot image-text retrieval

We test our approach on the standard Flickr30k [51], MS-COCO [34] and nocaps [2] datasets, containing 1,000, 5,000 and 15,100 test samples respectively. For the latter, we simply average the results on the three partitions.

As shown in Tab. 1, across all three datasets, our approach significantly surpasses the current state-of-the-art including models of similar size. It even outperforms the much bigger EVA-CLIP (18B) model (85.0% vs. 83.3%) on Flickr30k, (59.0% vs. 55.6%) on MS-COCO and (72.3% vs. 69.3%) on nocaps in terms of @R1 for image retrieval. Similarly, we outperform the LVLM-based E5-V model by 5.5% on Flickr30k, 7% on MS-COCO, and 6.4% on nocaps.

4.2. Image-text compositionality

Herein, we focus our comparison on the currently most challenging test sets, SugarCrepe [18] and SugarCrepe++ [15] (for Winoground [47] please see supp. material). For SugarCrepe++, we are mostly interested in the Image-to-Text (ITT) setting since the Text-to-Text (TOT) one evaluates the language component of the methods only.

As Tabs. 2 and 3 show, our approach is the best in both SugarCrepe and SugarCrepe++ (ITT). On SugarCrepe, we outperform the 18B EVA-CLIP model on all categories, with particularly large gains on relation replacement (76.1 vs. 86.8), attribution adding (85.0 vs. 95.8), and object swap (65.3 vs. 79.0). The last case is particularly interesting as it directly measures the *bag-of-words* behavior, showcasing significant improvements offered by our method. Additionally, we outperform the E5-V variant based on the same LLaVA-1.5-7B model that we used, and the one based on the heavier LLaVA-Next-8B. A similar trend is observed on SugarCreppe++ where we outperform EVA-CLIP (18B) by up to 10.9% (on object swap) and E5-V (ITT) in all but relation replacement. Thanks to its text-text training, E5-V surpasses our method for the TOT setting, but we note that their loss can be readily incorporated into our framework, leaving this for future work.

5. Ablation studies

5.1. Impact of method's components

We quantify the impact of the proposed method's components by training on a smaller 1M subset, reporting results on SugarCrepe (averaged over each category) and on Flickr30k (R@1 for T2I and I2T).

Impact of adaptation components: We start by measur-

Table 3. Comparison with state-of-the-art on the SugarCrepe++ compositionality benchmark.

Method	Params	Swap Object		Swap Attribute		Replace Object		Replace Attribute		Replace Relation	
	(B)	ITT	TOT	ITT	TOT	ITT	TOT	ITT	TOT	ITT	TOT
Human	–	100.00	96.7	96.7	93.3	100.00	97.00	100.00	98.3	100.00	96.7
CLIP (ViT-L) [40]	0.43	46.0	14.5	44.5	28.7	92.0	81.3	68.8	56.3	53.4	39.1
BLIP (ViT-L) [29]	0.23	46.8	29.8	60.1	52.5	92.6	89.1	71.7	75.0	56.8	57.7
BLIP2 (ViT-L) [30]	1.17	37.9	39.5	51.9	55.4	94.8	96.9	73.2	86.5	65.1	69.6
OpenCLIP (ViT-G/14) [42]	1.37	40.7	27.4	54.2	49.6	93.1	89.4	72.5	73.1	57.6	51.4
OpenCLIP (ViT-BigG/14) [42]	2.54	48.8	28.2	57.7	52.4	94.2	90.5	76.4	72.6	59.4	53.6
EVA-02-CLIP (ViT-E/14+) [44]	5.04	48.4	28.2	56.3	49.4	94.5	88.9	76.3	70.6	59.4	49.4
EVA-CLIP (8B) [45]	8.22	43.6	25.4	55.2	46.9	93.7	85.8	73.4	67.9	59.7	49.2
EVA-CLIP (18B) [45]	18.3	45.2	25.4	55.5	47.6	94.1	85.1	77.0	69.8	60.4	47.8
NegCLIP [52]	0.15	55.3	34.7	58.0	56.5	89.5	94.5	69.4	76.3	52.3	51.6
CLIP-SVLC [13]	0.15	43.0	18.9	48.4	34.6	80.9	91.6	57.0	66.9	47.3	51.3
BLIP-SGVL [17]	0.15	13.2	–	38.8	–	53.8	–	34.4	–	30.7	–
LLaVA-1.5-7B [35]	7.06	23.8	30.7	28.0	29.5	58.1	63.0	46.8	58.1	52.3	63.4
VLM2Vec (Mistral-7B) [23]	7.30	40.7	39.9	48.1	50.0	94.6	96.9	77.0	85.6	67.9	70.7
E5-V (LLaVA-Next-8B) [22]	8.36	50.8	**48.4**	49.7	56.9	93.1	**97.6**	76.1	**87.1**	**74.7**	**84.4**
E5-V (LLaVA-1.5-7B) [22]	7.06	39.5	42.3	40.7	48.5	89.7	94.6	71.7	86.4	72.0	81.5
VladVA (Ours) (LLaVA-1.5-7B)	7.06	**56.1**	36.7	**63.0**	**62.5**	**95.0**	93.0	**78.2**	82.3	71.1	66.3

ing the impact of the efficient adaptation strategy based on soft prompting and adapter-finetuning. For simplicity, we ablate this by training using only the contrastive loss. As the results from Tab. 4 show, both components, individually and jointly, provide notable gains on top of the original LLaVA-1.5-7B model (i.e. the case of no adaptation).

While LoRA fine-tuning performs better than soft-prompting (due to its bigger capacity), the latter alone performs surprisingly well. To understand why, we analyze the changes the soft prompts undergo by finding the closest embedding in the LLM's vocabulary. This results in the following decoded sentences: "</s> '<Summarize the provided image in one word:/ $[" and, " ωaSummarize the provided text in one word:−". The two sentences remain unchanged semantically, with the only characters changed being the ones at the start and the end of the prompt. Intuitively, this allows the model to mark/specialize the token that should gather the visual or textual evidence for discriminative tasks.

Impact of AR loss: We measure the impact of the proposed autoregressive loss on long captions from Sec. 3.2.2. As Tab. 4 shows, the AR loss adds a notable performance boost across all datasets tested. Finally, we note that using the long captions in isolation, without the proposed training strategy and loss, does not result in measurable gains.

5.2. Impact of training dataset size

Although at a relatively small scale (training is expensive due to the LVLM), herein, we aim to examine whether scaling the dataset size benefits the proposed discriminative adaptation of LVLMs. Specifically, we scale our dataset size from 1M to 8.1M samples. As Tab. 5 shows, we obtain steady gains across all metrics, with no signs of immediate saturation. This suggests that some potential is still left untapped, and further scaling could result in extra gains.

Table 4. Impact of adaptation components and AR loss. All models are trained on 1M samples.

Method	AR loss	SugarCrepe			Flickr30k	
		Replace	Swap	Add	T2I	I2T
LLaVA-1.5-7B	✗	81.9	59.8	64.7	59.6	65.6
+ soft-prom.	✗	86.4	66.9	89.3	76.7	91.7
+ adapter-ft.	✗	87.0	69.8	88.8	79.1	91.4
+ adapter-ft. + soft-prom.	✗	87.1	72.0	88.6	79.6	**92.9**
+ adapter-ft. + soft-prom.	✓	**89.5**	**75.5**	**89.5**	**80.6**	91.8

Table 5. Impact of training data size.

Training data	SugarCrepe			Flickr30k	
	Replace	Swap	Add	T2I	I2T
LLava-1.5-7B (0M)	81.9	59.8	64.7	59.6	65.6
OpenImages (1M)	87.1	72.0	88.6	79.6	92.9
OpenImages (4M)	88.2	79.6	89.1	82.3	93.1
+ ShareGPT-4V (1.3M)	91.0	80.3	92.4	83.1	94.0
+ CC3M (2.8M)	**92.3**	**80.9**	**95.5**	**85.0**	**94.3**

6. Conclusions

We introduced a new framework for adapting a *generative* LVLM into a *discriminative* model, unlocking its innate capability for powerful image-text discrimination and enhanced language understanding. Our framework uses both short and long captions for training the LVLM with contrastive and next-token prediction losses respectively. We also presented a parameter-efficient adaptation method using a combination of soft prompting and LoRA adapters. Finally, we showed that our approach results in significant improvements over state-of-the-art models of similar size for image-text retrieval and compositionality benchmarks.

References

[1] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 2

[2] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019. 7

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 2

[4] Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schmidt. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. *arXiv preprint arXiv:2308.06595*, 2023. 2

[5] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 1, 2

[6] Adrian Bulat and Georgios Tzimiropoulos. LASP: Text-to-text optimization for language-aware soft prompting of vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23232–23241, 2023. 5

[7] Adrian Bulat, Yassine Ouali, Ricardo Guerrero, Brais Martinez, and Georgios Tzimiropoulos. Efficient vision-language pre-training via domain-specific learning for human activities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7978–8000, 2024. 2

[8] Adrian Bulat, Yassine Ouali, and Georgios Tzimiropoulos. FFF: Fixing flawed foundations in contrastive pre-training results in very strong vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14172–14182, 2024. 2

[9] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. 4, 7

[10] Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. Measuring and improving chain-of-thought reasoning in vision-language models. *arXiv preprint arXiv:2309.04461*, 2023. 1

[11] Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, et al. Mobilevlm v2: Faster and stronger baseline for vision language model. *arXiv preprint arXiv:2402.03766*, 2024. 2

[12] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024. 1, 2

[13] Sivan Doveh, Assaf Arbelle, Sivan Harary, Eli Schwartz, Roei Herzig, Raja Giryes, Rogerio Feris, Rameswar Panda, Shimon Ullman, and Leonid Karlinsky. Teaching structured vision & language concepts to vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2657–2668, 2023. 8

[14] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 1

[15] Sri Harsha Dumpala, Aman Jaiswal, Chandramouli Sastry, Evangelos Milios, Sageev Oore, and Hassan Sajjad. Sugarcrepe++ dataset: Vision-language model sensitivity to semantic and lexical alterations. *arXiv preprint arXiv:2406.11171*, 2024. 7

[16] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 6(3):e30, 2021. 2

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

[18] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36, 2024. 1, 2, 7

[19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 5

[20] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1

[21] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 2

[22] Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*, 2024. 1, 2, 3, 4, 6, 7, 8

[23] Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*, 2024. 2, 6, 7, 8

[24] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object

detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020. 7

[25] Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*, 2024. 2

[26] Martha Lewis, Nihal V Nayak, Peilin Yu, Qinan Yu, Jack Merullo, Stephen H Bach, and Ellie Pavlick. Does clip bind concepts? probing compositionality in large image models. *arXiv preprint arXiv:2212.10537*, 2022. 1, 2

[27] Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, 2024. 1, 2

[28] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 1, 2

[29] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 1, 6, 7, 8

[30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2, 4, 6, 7, 8

[31] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 2, 5

[32] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021. 1

[33] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024. 2

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1, 7

[35] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 2, 6, 7, 8

[36] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2, 6

[37] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[38] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 7

[39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 7

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

[41] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506, 2020. 7

[42] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2, 6, 7, 8

[43] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 7

[44] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 6, 7, 8

[45] Quan Sun, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. Eva-clip-18b: Scaling clip to 18 billion parameters. *arXiv preprint arXiv:2402.04252*, 2024. 6, 7, 8

[46] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024. 2

[47] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 7

[48] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al.

Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1, 2

[49] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 2

[50] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. 2

[51] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 1, 3, 7

[52] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022. 1, 2, 6, 7, 8

[53] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 1, 2

[54] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 5

[55] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 2