

Fwd2Bot: LVLM Visual Token Compression with Double Forward Bottleneck

Adrian Bulat^{*1,2}

Yassine Ouali^{*1}

Georgios Tzimiropoulos^{1,3}

¹Samsung AI Cambridge ²Technical University of Iasi ³Queen Mary University of London

Abstract

In this work, we aim to compress the vision tokens of a Large Vision Language Model (LVLM) into a representation that is simultaneously suitable for (a) generative and (b) discriminative tasks, (c) is nearly lossless, and (d) is storage-efficient. We propose a novel compression approach, called Fwd2Bot, that uses the LVLM itself to compress the visual information in a task-agnostic manner. At the core of Fwd2bot there exists a “double-forward pass” training strategy, whereby, during the first forward pass, the LLM (of the LVLM) creates a bottleneck by condensing the visual information into a small number of summary tokens. Then, using the same LLM, the second forward pass processes the language instruction(s) alongside the summary tokens, used as a direct replacement for the image ones. The training signal is provided by two losses: an autoregressive one applied after the second pass that provides a direct optimization objective for compression, and a contrastive loss, applied after the first pass, that further boosts the representation strength, especially for discriminative tasks. The training is further enhanced by stage-specific adapters. We accompany the proposed method by an in-depth ablation study. Overall, Fwd2Bot results in highly-informative compressed representations suitable for both generative and discriminative tasks. For generative tasks, we offer a 2× higher compression rate without compromising the generative capabilities, setting a new state-of-the-art result. For discriminative tasks, we set a new state-of-the-art on image retrieval and compositionality.

1. Introduction

LVLMs (*i.e.* LMMs or Visual LLMs) are LLMs [6, 15] that, in addition to text, are capable of integrating and processing visual information as input context [19]. They have been deployed for many tasks and use cases requiring multi-modal (*i.e.* vision-language) understanding and reasoning like image captioning, visual question answering, and multi-modal chatbots. They are typically constructed

by stitching and then fine-tuning together a pre-trained vision encoder (e.g. CLIP [33]) and an LLM on conversational/instructional data, with the LLM jointly processing the visual and language tokens thereafter [19].

A key characteristic that distinguishes LVLMs from LLMs is that, for many tasks, the language instructions are short, and the input sequence length is dominated by the visual tokens. Hence, in this work, our aim is to answer the following question: *“How can we compress LVLM’s visual tokens into a short sequence of summary tokens in order to enable efficient deployment with minimal accuracy degradation?”* We refer to this problem as *Token Compression*.

Several works have been recently proposed aiming to address the relevant problem of (what we call) *Token Reduction* [3, 12, 22, 35]. Although Token Compression and Token Reduction have similar goals (*i.e.* efficient deployment), they operate under different settings and trade-offs. Token Reduction assumes an on-the-fly setting whereby during inference, given an input image and the text instruction, the number of visual tokens must be effectively pruned (*i.e.* reduced) before they are fed to the LLM. Token Compression distinguishes between offline and online processing. During offline processing, an image is compressed in a small number of visual summary tokens in a query/instruction agnostic manner which are stored for further processing. During inference (*i.e.* online processing), the (saved) summary tokens can be directly provided along with the text instruction to the LLM (of the LVLM) for efficient processing of both generative and discriminative tasks. Although Token Reduction methods can also be run in offline-online mode, they do not utilize this distinction to improve compression efficiency, nor can they be used for discrimination.

Going beyond Token Reduction methods, in this work, we set up a training procedure that specifically capitalizes on two-staged processing (*i.e.* offline and online) by trading one-off offline processing with significantly superior accuracy gains (for the same number of summary tokens used) during online inference. Our main methodological contribution is that an excellent way to perform Token Compression is to use the LVLM itself to compress the visual tokens based on a “double-forward pass” strategy. Specifically, our method, called Fwd2Bot, performs a first forward

^{*} Equal contribution. Listing order is random.

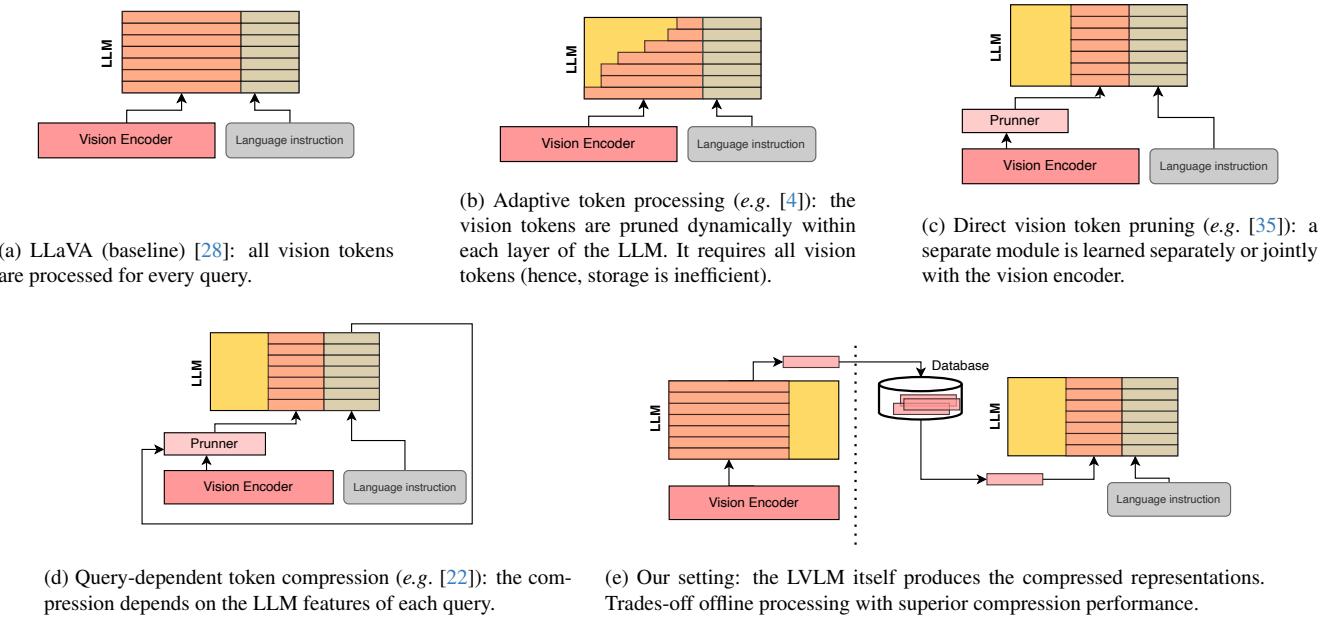


Figure 1. Different methods/paradigms for reducing/compressing the visual tokens in LVLMs.

pass whereby the trainable summary tokens, the image tokens, and the tokens of the prompt “Summarize the image in a few words.” are all fed to the LVLM. This step allows the summary tokens to interact with the image and the prompt tokens by passing through all the layers of the LVLM. Then, we create a bottleneck by performing a second forward pass whereby only the summary tokens and the language instruction are passed through the LLM for optimization with next-token prediction loss. This step allows the summary tokens to be optimized directly for the task at hand.

Going even further beyond previous works, a second methodological contribution of ours is that the summary tokens are optimized not only for autoregressive generation but also for discrimination (*i.e.* image-text retrieval). For this purpose, we introduce a contrastive loss for optimizing the summary tokens. Importantly, we show that the contrastive loss is beneficial for enhancing the accuracy of the generative tasks, too.

Overall, we make the following contributions:

- We introduce Fwd2Bot, a new two-staged offline-online Token Compression method whereby highly condensed visual summary tokens are produced by the LVLM itself. For this purpose, we set up a training procedure whereby two forward passes through the LLM of the LVLM are utilized, creating a visual information bottleneck in-between that allows the training of the summary tokens for the task in hand.
- We show that the learned summary tokens produced by our method can be used not only for auto-regressive generation but also for image-text discrimination (*i.e.* retrieval). For this purpose, we introduce a second loss for

contrastive optimization. Importantly, we show that this loss further enhances the generative abilities of the model.

- For generative tasks, our method achieves a 2x higher compression rate without compromising the generative capabilities, significantly outperforming previous works. For discriminative tasks, we set a new state-of-the-art result for text-image retrieval and compositionality.

2. Related work

2.1. Token Reduction in LVLMs

The current prevalent LVLM architecture consists of a pre-trained LLM, a pretrained vision encoder, and a projector head that maps the vision tokens from the output space of the vision encoder to the input space of the LLM [2, 5, 25, 28, 29, 41, 42, 47]. Despite their remarkable accuracy, a limitation of such approaches is their computational cost. In large part, this is a consequence of the LLM having to process a large number of vision tokens (*e.g.* 576 in [28]). To alleviate this, recently, a series of works focus on reducing their number [3, 12, 22, 35, 44] (see Fig. 1 for a summary of concepts).

PruMerge [35] reduces the number of vision tokens by discarding and/or merging the tokens produced by the vision encoder. The approach itself is training-free and relies on the similarities between the spatial local tokens and the global (CLS) one. In contrast, [23] uses a learnable attentive module that combines and consolidates multi-level feature maps into a smaller set. [3, 12] learn nested representations using either a set of convolutions aimed to capture multi-resolution representation [3, 5] or a transformer

layer [12]. [4, 41] dynamically vary the number of vision tokens within the LLM on a layer-by-layer basis using a rule-based system that takes into consideration the correlation between the tokens themselves or the tokens and the instruction. Building on top of [23], QueCC [22] conditions the token selection process on the user query, first processed by the LLM to produce a feature embedding before being used for conditioning of the compression module.

Our method differs from the above works in several ways. Firstly, it is the only method designed and optimized to benefit from a two-staged process where an offline step is used to produce highly condensed summary tokens in a task-agnostic manner, which can be then efficiently used during inference. Secondly, it is the only method that sets up a training procedure that uses the whole LVLM itself to compress the visual tokens using a “double-forward pass” training strategy. Thirdly, it is the only method that learns summary representations for both generative and discriminative tasks. Finally, in terms of results, our method matches and surpasses the state-of-the-art of [22] without having to recompute the vision tokens for every new query.

2.2. Discriminative LVLMs

Very recently, a series of works [13, 16, 17] have explored the possibility of converting LVLMs into discriminative models, or of pairing LLMs with CLIP vision encoders. For example, [13] directly aligns a pretrained LLM with a pretrained CLIP vision encoder. E5-V [16] through text-only contrastive training converts a generative LVLM into a discriminative one, while [17] expands it to multi-modal retrieval. One major limitation of these approaches is the loss of generative abilities post-adaptation. In this work, we address this very issue, creating a unified model that excels both at generative and discriminative tasks, surpassing recently proposed LVLMs adaptations for image-text retrieval and compositionality.

3. Method

3.1. Preliminaries

We implement our approach on top of the LLaVA-1.5 [28] model, leaving all architectural components unchanged. The LLaVA model consists of a pretrained CLIP vision encoder $g(\cdot)$, a projection matrix \mathbf{W} , and an LLM $f(\cdot)$. The input image \mathbf{X}_v is passed to CLIP to produce vision embeddings $\mathbf{H}_v = g(\mathbf{X}_v)\mathbf{W}$. The language embeddings \mathbf{H}_q are obtained from the input language instruction \mathbf{X}_q . Finally, the concatenated vision and language embeddings are passed to the LLM to compute the answer (output) embeddings $\mathbf{H}_a = f(\mathbf{H}_v; \mathbf{H}_q)$ which is decoded to the corresponding answer (output) sequence \mathbf{X}_a .

Although autoregressive in nature, recently, it was shown that the model can be run in discriminative mode producing

vision-text embeddings for matching à la CLIP [16]. Using the prompt \mathbf{X}_p “summarize the above image in one word” (or similar), the vision embedding is produced as $\mathbf{e}_v = \mathbf{H}_a[-1]$, $\mathbf{H}_a = f(\mathbf{H}_v; \mathbf{H}_p)$ (\mathbf{H}_p is the language embedding of \mathbf{X}_p). Analogously, and given a text query \mathbf{X}_{query} , the text embedding is constructed as $\mathbf{e}_t = \mathbf{H}_a[-1]$, $\mathbf{H}_a = f(\mathbf{H}_{query}, \mathbf{H}_p)$ (\mathbf{H}_{query} is the embedding of a \mathbf{X}_{query}). As in CLIP [33], the image-text similarity is computed as $s = \text{cos_sim}(\mathbf{e}_v, \mathbf{e}_t)$.

3.2. Double forward bottleneck algorithm

The two axes influencing LVLM’s efficiency are the model architecture/size and the input sequence length. In this work, we assume that the architecture is fixed. Hence, the inference cost is defined by the input sequence length, which turns out to be dominated by the length (number) k of the vision embeddings $\mathbf{H}_v \in \mathbb{R}^{k \times d}$. Specifically, for a LLaVA model, $k = 576$, which is significantly higher than the typical text query length and answer [22]. Hence, in this work, our goal is to derive a compressed visual token representation $\mathbf{H}_v^c \in \mathbb{R}^{k' \times d}$ where $k' \ll k$ without compromising the model’s accuracy. Importantly, besides improving subsequent runs, a small sequence length also opens the path to offline pre-processing, whereby one can precompute, store, and re-use the compressed representation \mathbf{H}_v^c without having to process the original image again.

Several approaches for visual token compression have been proposed, including the introduction of new modules between the vision encoder and the LLM [12], finetuning the vision encoder [3], or compressing the vision tokens in an instruction/query dependent manner [22]. Departing from these approaches, we take a totally different path by proposing to leverage the LVLM itself (the LLM of the LVLM in particular) to self-compress the visual tokens. Our motivation for this is multifold. Firstly, LLMs have already excelled at text summarization [30, 46]. Hence, we propose to utilize them for image summarization (*i.e.* compression). However, as summarization with text, due to quantization (*i.e.* tokenization), is inefficient with respect to the sequence length, we instead perform this summarization in a continuous latent space. Secondly, the LLM of the LVLM has already been utilized very recently to compute discriminant image-text embeddings [16]. However, these embeddings cannot be used for generation. Hence, in this paper, based on a “double-forward pass” training strategy, we develop a simple, yet powerful, training bottleneck through which visual summary tokens at the output of the model are trained to highly compress visual information that can be used for both generation and discrimination. Overall, we call our method Fwd2Bot. See Fig. 2 for an overview.

More specifically, given an input image \mathbf{X}_v , we introduce the summary tokens *i.e.* learnable input embeddings $\mathbf{H}_r \in \mathbb{R}^{k' \times d}$ which evolve into the compressed vision em-

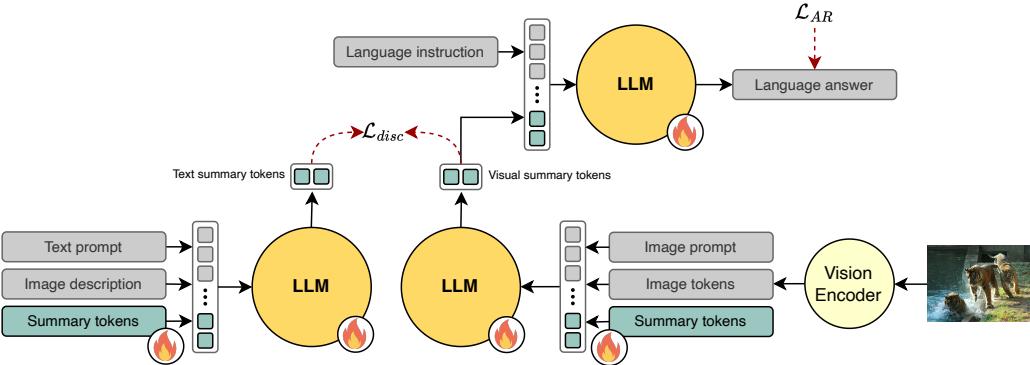


Figure 2. Fwd2Bot training pipeline: A first forward pass from the LLM (of the LVLM) creates a bottleneck by condensing the visual information into a small number of visual summary tokens. Then, using the same LLM (weights of depicted LLMs are shared), a second forward pass processes the language instruction(s) alongside the summary visual tokens for training with a next-token prediction loss \mathcal{L}_{AR} (see Sec. 3.2). Furthermore, a contrastive loss \mathcal{L}_{disc} , applied after the first pass, is utilized to further boost the representation strength, especially for discriminative tasks (see Sec. 3.3). Components marked with are trainable.

embeddings $\mathbf{H}_v^c \in \mathbb{R}^{k' \times d}$ after *a first forward pass from the LLM of the LVLM*:

$$[\cdot, \cdot, \cdot, \mathbf{H}_v^c] = f(\mathbf{H}_v; \mathbf{H}_p; \mathbf{H}_r), \quad (1)$$

where \mathbf{H}_p are the embeddings of the handcrafted prompt \mathbf{X}_p “Summarize the image in a few words.”. Clearly, during this pass, \mathbf{H}_r interact with both \mathbf{H}_v and \mathbf{H}_p . As the transformed embeddings \mathbf{H}_v^c are query-agnostic, for subsequent language instructions/queries, the (same) LLM simply takes as input \mathbf{H}_v^c instead of \mathbf{H}_v .

To learn the compressed representation \mathbf{H}_v^c , during training, we perform *a second forward pass from the LLM of the LVLM* where this time only \mathbf{H}_v^c and the language instructions/embeddings are passed to the LLM. An autoregressive loss is applied at the output of the second forward pass:

$$\mathcal{L}_{AR} = - \sum_{i=1}^L \log(p_\theta(x_i | \mathbf{H}_v^c, \mathbf{X}_{q,<i}, \mathbf{X}_{a,<i})), \quad (2)$$

where θ represents the trainable parameters, $\mathbf{X}_{q,<i}$ and $\mathbf{X}_{a,<i}$ are the query/instruction and, respectively, answer tokens located before the current predicted token x_i . \mathbf{H}_v is obtained from Eq. 1. Note that the weights of LLM are shared between the two forward passes. The flow of gradients through \mathbf{H}_v^c results in a single model that can both compress and reason/generate answers by looking solely at the compressed tokens.

Intuitively, our algorithm can be also interpreted as a form of implicit chain-of-thought in the latent space [9], with the LLM “rephrasing” the content of the vision sequence in a condensed manner for itself. Notably, while the input and output spaces of the LLM are not perfectly aligned, they are sufficiently close to resulting in good

alignment of the compressed representations in just a few hundred iterations, making the whole training process efficient. That is the compressed representations simultaneously lie in the input and output space of the LVLM.

3.3. Discriminative adaptation

As noted above, an important conceptual difference between our approach and prior works in LVLM visual token reduction is that, in our work, the compressed representations lie simultaneously in both the input and output space of the LVLM. Unlike previous approaches, this enables us to directly leverage the compressed representations for CLIP-like discrimination in a zero-shot manner, as detailed in Sec. 3.1. However, in this case, the discriminant performance is suboptimal as there is no explicit loss to encourage the separability of concepts.

To address this, and create a unified compressed representation suitable for both generative and discriminative tasks, we also propose to apply a contrastive loss over \mathbf{H}_v^c , at the output of the first forward pass. Not only does this loss enhance the discriminative properties of the compressed representation, but it can also improve the generative ability of the model thanks to learning a better underlying representation.

Given a dataset consisting of paired image-text samples, the contrastive loss, for a given batch containing B elements, is defined as:

$$\mathcal{L}_{disc} = \frac{1}{B} \sum_{k=1}^b \left(-\log \frac{\exp(s_v^{k,k})}{\sum_j \exp(s_v^{k,j})} - \log \frac{\exp(s_t^{k,k})}{\sum_j \exp(s_t^{j,k})} \right), \quad (3)$$

where $s_v^{k,j} = \text{cos_sim}(\mathbf{e}_v^k, \mathbf{e}_t^j)$ computes the cosine similarity between the k -th image and the j -th caption. $\mathbf{e}_v =$

Table 1. Comparison with various token reduction/compression methods on vision-language understanding tasks.

Method	# Tokens	GQA	MMB	MME	POPE	SQA	TextVQA	VisWiz	VQAv2
LLAVA-1.5 [28]	576	62.0	64.3	1510.7	85.9	66.8	58.2	50.0	78.5
PruMerge [35]	≈32	57.2	60.9	1350.3	76.3	68.5	56.0	45.2	72.0
TokenPacker [23]	36	59.6	62.8	1440.9	83.3	71.0	53.2	50.2	75.0
Matryoshka Multi. [3]	36	60.3	64.8	-	85.5	-	-	52.8	-
Matryoshka Query [12]	36	58.8	63.4	1416.3	81.9	66.8	-	51.0	73.7
QueCC [22]	36	60.5	62.5	1442.0	84.5	70.6	53.3	50.1	75.8
Fwd2Bot (Ours)	32	61.6	64.6	1472.1	85.9	68.5	55.8	53.1	77.1
TokenPacker [23]	16	58.9	62.7	1378.8	83.7	68.1	52.5	50.5	74.4
Matryoshka Query [12]	16	57.6	61.9	1408.5	80.8	67.5	-	49.8	71.1
QueCC [22]	16	59.0	62.2	1408.0	83.4	70.7	51.3	47.7	74.5
Fwd2Bot (Ours)	16	61.0	64.4	1470.0	85.6	67.7	54.2	49.8	76.5
TokenPacker [23]	4	56.2	61.5	1347.6	81.7	68.5	49.2	45.7	70.5
Matryoshka Query [12]	4	53.0	56.5	1176.1	77.6	65.1	-	49.4	64.1
QueCC [22]	4	56.5	62.1	1390.3	81.8	68.6	48.7	45.0	70.6
Fwd2Bot (Ours)	4	58.6	63.3	1403.0	84.3	67.7	52.5	51.6	74.5

$\frac{1}{k'} \sum \mathbf{H}_v^c$ and, $\mathbf{e}_t = \frac{1}{k'} \sum \mathbf{H}_t^c$, respectively. \mathbf{H}_t^c is computed analogously to \mathbf{H}_v^c (Eq. 1), except that it encodes textual data instead of visual, *i.e.* $\mathbf{H}_t^c = f(\mathbf{H}_{query}, \mathbf{H}_p, \mathbf{H}_r)$. \mathbf{H}_t^c is only used as part of the discriminative loss.

3.4. Overall training loss and data

The final model is trained using both losses, autoregressive and discriminative/contrastive:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{AR}} + \mathcal{L}_{\text{disc}}. \quad (4)$$

At a given iteration, depending on the training data, the applicable losses are used. That is, for conversational data sampled from the LLaVA-665k dataset, we apply \mathcal{L}_{AR} . For data sampled from CC3M, we apply $\mathcal{L}_{\text{disc}}$. If a conversational sample also has a caption associated with it, both losses are applied within the same iteration. For efficiency, the sampler will group together such cases. This also ensures that we have sufficiently large batches for contrastive training. The sampler aims for a 1:1 ratio between discriminative and autoregressive.

3.5. Stage-specific adaptation

To enable efficient adaptation, we train our models using LoRA [11] adapters which restrict the weight updates to a low-rank representation, $\Delta W = BA$, $\Delta W \in \mathbb{R}^{d \times m}$, $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times m}$, with $r \ll \min(d, m)$.

Although this works well, we use stage-specific adapters to further enhance the plasticity of the LVLM. We distinguish two stages that correspond to the two forward passes used during training: compression, which summarizes \mathbf{H}_v into \mathbf{H}_v^c , and generation, which produces \mathbf{X}_a given \mathbf{H}_v^c and \mathbf{X}_q . Depending on which operation we perform, compression or answer generation, different LoRA adapter weights A and B are used.

4. Results

4.1. Generative benchmarks

Following [28], we evaluate our approach on a diverse collection of datasets, mainly: GQA [14], MMB [31], MME [26], POPE [24], SQA [32], TextVQA [37], VisWiz [8] and VQAv2 [7]. To ensure fairness, in all cases, we fully align the test-time settings and processing with [28]. In addition to this, we also evaluate our approach for captioning on MS-COCO [27], Flickr30k [43] and NoCaps [1], comparing it to token reduction methods that have models openly available. See supplementary material for results on TextCaps [36].

When comparing our approach with the state-of-the-art token reduction methods for visual-language understanding, as the results from Tab. 1 show, we set a new best result, outperforming prior works using $2.25\times$ fewer tokens (16 vs 36). Our results for 32 and even 16 tokens nearly match the uncompressed LlaVA [28] baseline.

Similarly, when evaluated for zero-shot captioning (Tab. 2), our approach matches LLaVA’s accuracy, significantly outperforming prior methods. This suggests that the proposed approach encodes more information in its compressed tokens. We note that LLaVA saw some MS-COCO images during training; hence, the MS-COCO evaluation is not fully zero-shot for all methods listed.

4.2. Discriminative benchmarks

To measure the discriminative abilities of our model, we evaluate it on a diverse set of retrieval benchmarks: Flickr30k [43], MS-COCO [27], NoCaps [1] and SugarCrepe [10]. The last one measures the compositional capabilities of the model, an area where CLIP and CLIP-like models tend to under-perform.

To offer a full picture, our method is compared against a diverse set of contrastive models, varying in size (from

Table 2. Comparison with various token reduction/compression methods on image captioning.

Method	# Tokens	Flickr30K				COCO				nocaps			
		B@4	CIDEr	MET.	ROUGE	B@4	CIDEr	MET.	ROUGE	B@4	CIDEr	MET.	ROUGE
LLAva-1.5 [28]	576	30.6	81.2	25.0	53.4	32.9	115.4	27.7	56.3	42.9	105.3	28.9	59.8
PruMerge [35]	≈32	18.5	36.3	15.7	40.2	18.5	66.3	18.8	44.9	25.9	58.6	20.0	47.8
Matryoshka Multi. [3]	36	25.4	68.7	24.1	49.9	27.7	102.2	27.2	53.3	36.8	93.6	28.0	56.5
Matryoshka Query [12]	36	26.4	69.5	23.1	50.0	28.0	101.3	26.2	52.7	36.2	90.0	26.8	55.8
Fwd2Bot (Ours)	32	30.0	78.9	25.2	52.9	31.5	113.1	27.9	55.6	42.5	105.9	29.2	59.6
Matryoshka Query [12]	16	24.8	65.2	22.7	49.0	27.6	99.2	26.0	52.5	36.2	90.0	26.8	55.8
Fwd2Bot (Ours)	16	29.0	78.2	25.3	52.7	31.0	112.0	27.9	55.4	42.0	104.7	29.3	59.5
Matryoshka Query [12]	4	20.1	47.5	19.8	44.5	23.2	81.0	23.0	48.6	28.4	63.2	21.1	49.5
Fwd2Bot (Ours)	4	28.4	74.5	24.8	51.8	31.1	111.4	27.9	55.4	41.1	103.4	29.0	59.1

Table 3. 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	R@1	R@10	
Contrastive approaches														
CLIP (ViT-B) [33]	58.8	89.8	30.5	66.8	46.8	85.1	77.8	98.2	51.0	83.5	67.3	95.6		
CLIP (ViT-L) [33]	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) [20]	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) [21]	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) [34]	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) [34]	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+) [38]	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 [39]	80.3	97.2	52.0	82.9	65.3	93.2	94.5	99.7	70.1	93.1	83.5	98.6		
LVLM-based approaches														
LLaVA-1.5-7B [28]	59.6	89.3	34.4	69.6	46.9	83.3	65.6	92.3	35.6	70.5	52.1	88.1		
E5-V (LLaVA-1.5-7B) [16]	76.7	96.9	48.2	82.1	62.0	93.0	86.6	99.0	57.4	88.4	71.9	97.0		
VLM2Vec (Mistral-7B) [17]	80.1	97.3	52.0	85.6	65.9	94.5	90.3	99.6	68.2	93.2	79.2	98.5		
Fwd2Bot (Ours) (LLaVA-1.5-7B)	83.8	98.5	59.0	88.6	72.3	96.5	94.3	99.9	72.9	94.4	85.7	99.5		

0.2B parameters for CLIP (ViT-B) [33] to 8B for EVA-CLIP [39]) and architecture (CLIP [33], BLIP [21] and LVLM-based [16]). Our approach matches and outperforms all of them. In particular, for zero-shot image-to-text and text-to-image retrieval, as the results from Tab. 3 show, we match and outperform larger models, *i.e.* EVA-CLIP (8B vs. 7.06B), despite using 3 orders of magnitude fewer samples for training (2,700M for EVA-CLIP vs. ∼3M for ours). A similar trend can be observed when evaluated for compositionality on SugarCreppe (Tab. 4). An interesting observation about the latter is that models derived from LVLMs (*e.g.*, E5-V and ours) demonstrate significantly stronger compositionality. This suggests that the LVLM run in discriminative mode inherits the strong vision-language understanding of the underlying generative model.

5. Ablation studies & analysis

In this section, we ablate the main components of our approach. See supplementary material for more ablations.

Impact of each loss function: As detailed in Secs. 3.2 and 3.3, our models are trained using two losses: one autoregressive, applied after the second forward pass, and one contrastive, applied over the compressed tokens, after the first pass. In Tab. 5, we report results for a LLaVA model

evaluated using 16 tokens on generative and discriminative tasks. Intuitively, training solely with the discriminative loss (1st row) results in degraded generative performance, as no alignment between the input and output space of the LLM is performed. Moreover, discriminative losses applied over short captions tend to focus on coarse details, missing out on more fine-grained details. Conversely, applying only the generative loss (2nd row) results in degraded retrieval abilities, as no loss explicitly encourages concept separation. We note that the longer the training scheduler is, the more pronounced these degradations are for the two cases.

Finally, combining the two losses (3rd row) results in the best performance across the board. Notice that the two losses are complementary when applied jointly and boost the model’s accuracy on both sets of benchmarks.

Single vs. stage-specific LoRA vs. full fine-tuning: Herein, we compare the effect of training using (a) a single shared LoRA adapter, (b) stage-specific adapters, as proposed in Sec. 3.5, and (c) full fine-tuning. We present the results of these choices in Tab. 6. The best results are obtained using the stage-specific adapters. The fine-tuning run suffers from overfitting to some extent, and its larger training cost makes optimization more difficult.

Fig. 3 further solidifies the need for stage-specific LoRAs, as the optimal representations required during com-

Table 4. Comparison with state-of-the-art on the SugarCrepe compositionality benchmark.

Method	Params (B)	Replace			Swap		Add	
		Object	Attribute	Relation	Object	Attribute	Object	Attribute
Contrastive approaches								
NegCLIP [45]	0.15	92.7	85.9	76.5	75.2	75.4	88.8	82.8
CLIP (ViT-B) [33]	0.15	90.9	80.1	69.2	61.4	64.0	77.2	68.8
CLIP (ViT-L) [33]	0.43	94.1	79.2	65.2	60.2	62.3	78.3	71.5
BLIP (ViT-L) [20]	0.23	96.5	81.7	69.1	66.6	76.8	92.0	85.1
BLIP2 (ViT-L) [21]	1.17	97.6	81.7	77.8	62.1	65.5	92.4	87.4
OpenCLIP (ViT-G/14) [34]	1.37	95.8	85.0	72.4	63.0	71.2	91.5	82.1
OpenCLIP (ViT-BigG/14) [34]	2.54	96.6	87.9	74.9	62.5	75.2	92.2	84.5
EVA-02-CLIP (ViT-E/14+) [38]	5.04	97.1	88.5	74.2	67.3	74.1	91.8	83.9
EVA-CLIP [39]	8.22	96.4	86.6	74.8	66.1	74.6	91.3	82.0
LVLM-based approaches								
LLaVA-1.5-7B [28]	7.06	88.0	81.6	76.1	60.9	58.8	67.0	62.4
E5-V (LLaVA-1.5-7B) [16]	7.06	95.8	86.6	81.6	62.9	64.0	93.5	88.0
VLM2Vec (Mistral-7B) [17]	7.3	97.2	89.0	81.7	62.9	72.5	94.7	88.6
Fwd2Bot (Ours) (LLaVA-1.5-7B)	7.06	98.1	89.5	82.7	77.8	78.1	95.3	93.1

Table 5. Effect of generative and discriminative losses for generation (MMB, MME, TextVQA) and retrieval (Flickr30K, MS-COCO).

Method	MMB	MME	TextVQA	Flickr30K		MS-COCO	
				T2I	I2T	T2I	I2T
Discriminative	46.2	624.3	13.5	84.3	94.8	56.3	73.2
Generative	64.1	1420.1	54.2	61.3	76.0	33.9	47.0
Both	64.4	1470.0	54.2	83.8	94.4	56.8	70.2

Table 6. Single vs. stage-specific LoRA v.s full finetuning for generation (MMB, MME, TextVQA) and retrieval (Flickr30K, MS-COCO).

Method	MMB	MME	TextVQA	Flickr30K		MS-COCO	
				T2I	I2T	T2I	I2T
Fine-tuning	64.3	1413.1	52.9	83.1	94.0	56.2	70.4
Single LoRA	64.3	1410.5	51.8	83.8	94.1	56.5	69.9
Stage LoRA	64.4	1470.0	54.2	83.8	94.4	56.8	70.2

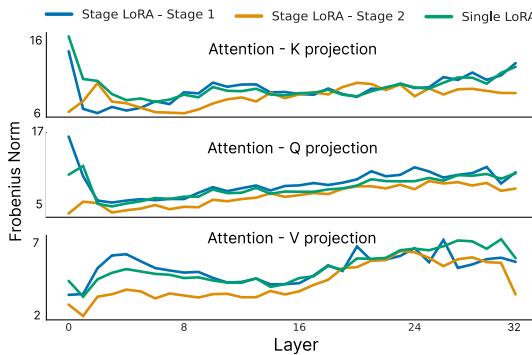


Figure 3. The norm of the learned LoRA weights adjustment $\Delta W = BA$ for a model trained with either a single LoRA or stage-specific LoRAs.

pression (first forward pass) vs downstream inference (second forward pass) are different, especially for earlier layers.

Double forward vs single forward: To showcase the importance of our “double-forward pass” training strategy, we conducted the following experiment: instead of using the LLM itself to compress the vision summary tokens, we use the CLIP vision encoder only, effectively compressing them

in the input space of the LLM. In this case, the loss is directly applied after the LLM, as in LLaVA, using a single forward pass. As shown in Tab. 7, this baseline (1st row) vs ours (2nd row) performs significantly worse.

Table 7. Double vs single forward pass for generation (MMB, MME, TextVQA) and retrieval (Flickr30K, MS-COCO).

Method	MMB	MME	TextVQA	Flickr30K		MS-COCO	
				T2I	I2T	T2I	I2T
Single-forward stage	61.3	1305.0	42.1	80.8	92.1	53.1	66.2
Double-forward (Ours)	64.4	1470.0	54.2	83.8	94.4	56.8	70.2

What do the compressed tokens encode? The compressed representation gradually encodes, from left to right, coarser to finer-grained concepts. This effect can be observed in Fig. 5, where, as the number of tokens increases, the caption generated correctly captures more elements present in the photo, importantly reducing hallucinations. This effect is also corroborated in Fig. 6. There, we mask-out different groups of (4 and 8) tokens, quantitatively measuring the impact of this: earlier tokens induce larger drops in performance (*e.g.* masking the first 8 tokens reduces performance by 10%). However, the performance does not drop to (near) 0, which suggests that there is also some degree of redundancy. The observed behavior can largely be attributed to the causal attention masking used by the LVLM, which encourages a directional information distribution.

How does the model behavior change? To shed light on the changes the model undergoes to act as a self-compressor, we analyze the attention patterns before and after our fine-tuning. The results of this visual analysis are presented in Fig. 4. Looking on the left side, we can observe that LLaVA exhibits a sparse attention pattern across all layers, particularly early on. In contrast, during self-compression, our model attends to all visually important parts of the image, having a significantly denser attention pattern at all layers. Intuitively, in the first case, as the

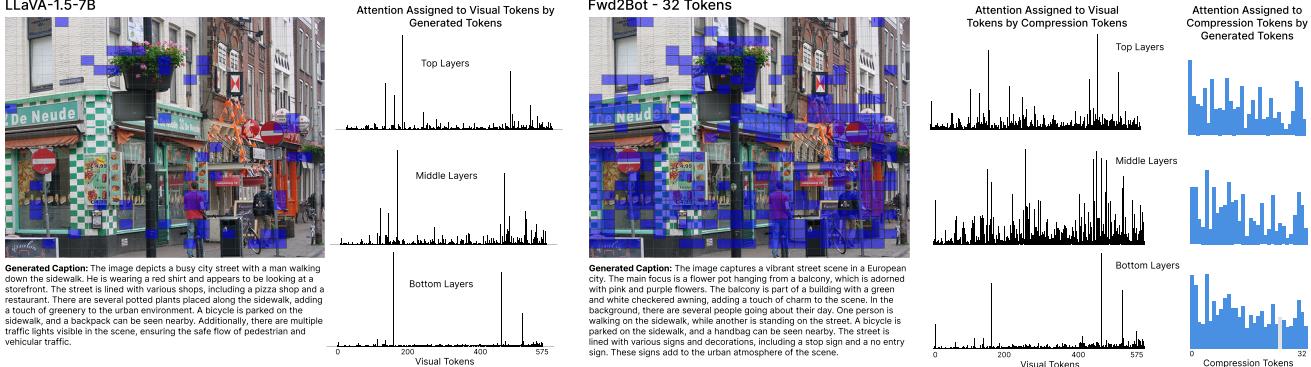


Figure 4. Visualization of attention weights assigned to the 576 visual tokens and the 32 compressed tokens. On the left, we show the cumulative weights assigned to each visual token by the generated tokens for the baseline LLaVa-1.5-7B model. For Fwd2Bot, on the right, we first display the per-visual-token weights assigned by the summary tokens during the first forward pass to produce the summary tokens. We then show the weights assigned to the compressed tokens by the generated tokens during the second forward pass.

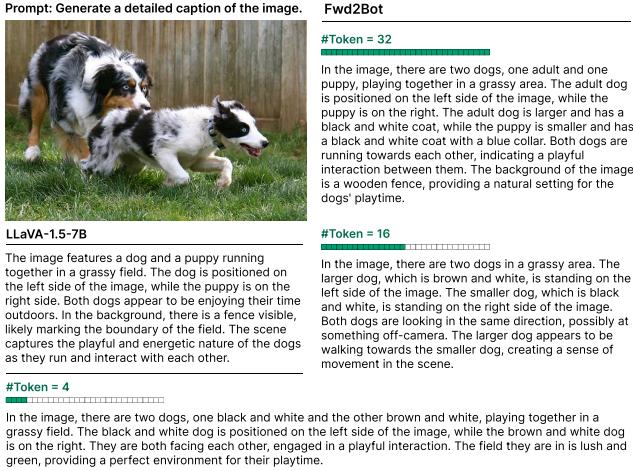


Figure 5. Captioning with variable number of summary tokens.

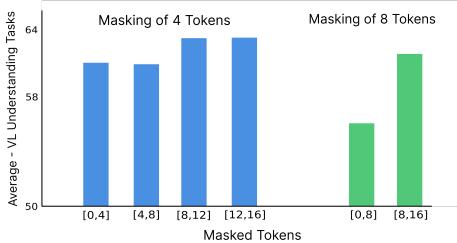


Figure 6. The relative importance of different subsets of visual tokens. We show the mean over the VL understanding tasks when masking specific subsets of the compressed visual tokens.

model has access to all tokens, during generation, the model can peak back at the vision tokens as needed. In contrast, during compression, the LLM must ensure that all visually important details are stored in the compressed representation. Finally, on the right-most part of the figure, we showcase the attention pattern between the generated tokens and the compressed representation obtained during text genera-

tion from compressed representations. Earlier and late-most tokens tend to have a higher attention weight.

6. Conclusions

In this work, we introduced Fwd2Bot, a novel LVLM visual token compression approach that uses the LVLM itself to compress the visual information in a task-agnostic manner, that is trained using a new “double-forward pass” training strategy. This results in a compressed visual representation that is simultaneously suitable for (a) generative and (b) discriminative tasks, (c) is nearly lossless, and (d) is storage-efficient. Performance-wise, for generative tasks, we offer a 2× higher compression rate without compromising the generative capabilities, setting a new state-of-the-art. For discriminative tasks, we also set a new state-of-the-art result on image retrieval and compositionality.

References

- [1] 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. 5
- [2] 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
- [3] Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. Matryoshka multimodal models. In *Workshop on Video-Language Models@ NeurIPS 2024*, 2024. 1, 2, 3, 5, 6, 12
- [4] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pages 19–35. Springer, 2024. 2, 3
- [5] 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
- [6] 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
- [7] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 5
- [8] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 5
- [9] Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhitong Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024. 4
- [10] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcreepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36, 2024. 5
- [11] 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. 5
- [12] Wenbo Hu, Zi-Yi Dou, Liunian Harold Li, Amita Kamath, Nanyun Peng, and Kai-Wei Chang. Matryoshka query transformer for large vision-language models. *arXiv preprint arXiv:2405.19315*, 2024. 1, 2, 3, 5, 6, 12
- [13] Weiquan Huang, Aoqi Wu, Yifan Yang, Xufang Luo, Yuqing Yang, Liang Hu, Qi Dai, Xiyang Dai, Dongdong Chen, Chong Luo, et al. Llm2clip: Powerful language model unlock richer visual representation. *arXiv preprint arXiv:2411.04997*, 2024. 3
- [14] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 5
- [15] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lampe, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 1
- [16] 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. 3, 6, 7, 11
- [17] Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhui Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*, 2024. 3, 6, 7
- [18] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 11
- [19] 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
- [20] 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. 6, 7
- [21] 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. 6, 7
- [22] Kevin Y Li, Sachin Goyal, Joao D Semedo, and J Zico Kolter. Inference optimal vlms need only one visual token but larger models. *arXiv preprint arXiv:2411.03312*, 2024. 1, 2, 3, 5, 11
- [23] Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. Tokenpacker: Efficient visual projector for multimodal llm. *arXiv preprint arXiv:2407.02392*, 2024. 2, 3, 5
- [24] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 5
- [25] 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
- [26] Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pages 405–409, 2024. 5
- [27] 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. 5
- [28] 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, 3, 5, 6, 7, 11, 12
- [29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2

- [30] Yixin Liu, Alexander R Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, et al. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. *arXiv preprint arXiv:2212.07981*, 2022. 3
- [31] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 5
- [32] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 5
- [33] 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, 3, 6, 7
- [34] 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. 6, 7
- [35] Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*, 2024. 1, 2, 5, 6, 12
- [36] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer, 2020. 5, 11
- [37] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 5
- [38] 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
- [39] 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
- [40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 11
- [41] 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, 3
- [42] 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
- [43] 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. 5
- [44] Gaotong Yu, Yi Chen, and Jian Xu. Balancing performance and efficiency: A multimodal large language model pruning method based image text interaction. *arXiv preprint arXiv:2409.01162*, 2024. 2
- [45] 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. 7
- [46] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57, 2024. 3
- [47] 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. 2

Results for larger models: In the main manuscript, we conduct experiments using a LLaVA-1.5 (7B) model. Herein, we validate how our approach behaves when using a larger model, *i.e.* a LLaVA-1.5 (13B). As the results from Table 8 show, the proposed method nearly matches the full LLaVA model’s accuracy using only 16 and even 4 tokens in this case, too.

Additional details regarding test-time inference: In Fig. 7, we depict the test-time inference flow for generative and discriminative tasks. The first step compresses the given image \mathbf{X}_v into its compressed representation \mathbf{H}_v^c . This representation is then stored in a database. Note that while \mathbf{H}_v^c can be computed on the fly, too, the scenario we are mostly interested in is pre-indexing, whereby the image representations are computed offline ahead of time.

Once stored, we can directly operate on this compressed representation for both generative and discriminative tasks. For generative tasks, the same LLM used for compression takes as input a user-provided instruction and the compressed image representation, producing an answer autoregressively (Fig. 7, top-right corner). For discriminative tasks, in order to measure the text-to-image similarity, à la CLIP, and again using the same LLM, we pass the user query (image description) to the LLM, producing a set of

embeddings. We can measure the similarity between the given image description by taking the cosine similarity between the sum of the precomputed compressed vision tokens and the text embeddings newly produced by the LLM (Fig. 7, left side).

Efficiency analysis: Generally, for a transformer model [40], the total inference-time compute cost [18], excluding nonlinearities, biases, normalizations, residuals and embeddings, which are negligible is equal to: $\text{FLOPs} \approx \mathcal{O}(N \cdot T)$. In other words, the number of FLOPs depends on the number of parameters, N , and the sequence length, T . For a more detailed breakdown, see [18].

Hence, for a LLaVA model, the total inference cost is approximatively: $\text{FLOPs}_{LLaVA} \approx \mathcal{O}(N_v \cdot V) + \mathcal{O}(N_{LLM} \cdot (V + Q + G))$. N_v and N_{LLM} represents the number of parameters of the vision, and LLM, respectively. V denotes the number of vision tokens (576 in LLaVA), Q of query/instruction tokens, and G of generated tokens.

For our approach, the initial cost to obtain the compressed representation \mathbf{H}_i^c is equal to $\mathcal{O}(N_v \cdot V) + \mathcal{O}(N_{LLM} \cdot (V + K))$, with $K \ll V$ representing the number of compressed tokens. This process is aimed to be run offline, as a pre-indexing step, resulting in the storage of $2Kd$ bytes per image. We note that this representation is directly suitable for both generation and retrieval. Once computed, the cost of subsequent queries is equal to: $\mathcal{O}(N_{LLM} \cdot (K + Q + G))$. In contrast, the current state-of-the-art approach, QueCC [22], requires $\mathcal{O}(N_{LLM} \cdot (K + 2Q + G))$ due to the dependency on the user

query/instruction for token compression. Moreover, from a storage point of view, in [22], all V (576) tokens must be stored or, alternatively, recomputed if an image from the database is queried again.

Additional discriminative comparisons with other token-summarization approaches. We note that our approach is the only one that compresses the vision tokens into a representation suitable both for generative and discriminative tasks, requiring no additional forward passes. However, herein, for completeness, we evaluate on our suite of discriminative tasks the current state-of-the-art token compression models that offered pretrained models. This is achieved by following the zero-shot setup described in Paper Section 3.1 and [16]. Unsurprisingly, as the results from Table 9 and 10 show, our approach significantly surpasses the other methods we compare with.

Additional zero-shot image captioning evaluations: In addition to the evaluation from the main manuscript, herein, we evaluate our approach for zero-shot captioning on TextCaps [36], a dataset for image captioning with reading comprehension. As the results from Table 11 show, we generally match the full-tokens LLaVA’s model performance. Importantly, our results remain stable as the number of compressed tokens decreases.

Full-attention vs causal: Vicuna, and hence LLaVA, much like the rest of the generative LLMs, employs causal attention masking in order to restrict the past states from

Table 8. Token compression performance on vision-language understanding tasks using a LLaVA-1.5 13B model.

Method	# Tokens	GQA	MMB	MME	POPE	SQA	TextVQA	VisWiz	VQAv2
LLAva-1.5 [28]	576	63.3	67.7	1531.0	86.2	71.6	61.3	53.6	80.0
Fwd2Bot (Ours)	32	62.2	67.6	1465.1	85.3	72.5	59.6	54.0	78.7
Fwd2Bot (Ours)	16	61.8	67.3	1473.5	85.0	72.4	57.5	54.2	78.4
Fwd2Bot (Ours)	4	59.9	66.4	1390.1	84.4	71.1	53.6	52.7	75.9

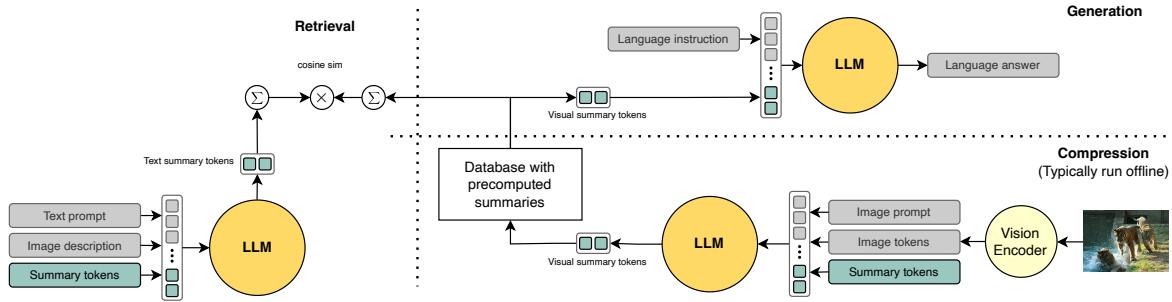


Figure 7. Test-time inference, depicting: compression (lower-right), generation (upper-right) and discrimination (left). Notice that in all cases we use the same LLM. The compressed embeddings are suitable for both set of tasks and are generally expected to be pre-computed offline.

Table 9. Zero-shot text-image retrieval accuracy on Flickr30K, COCO and nocaps.

Method	Tokens	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
LLaVA-1.5-7B [28]	576	59.6	89.3	34.4	69.6	46.9	83.3	65.6	92.3	35.6	70.5	52.1	88.1
PruMerge [35]	18	34.7	67.9	18.4	47.9	25.8	62.7	38.3	74.3	19.8	49.9	28.2	65.2
Matryoshka Multi. [3]	16	57.9	88.5	34.1	69.7	45.5	83.2	63.8	91.7	36.4	72.5	48.0	86.2
Matryoshka Query [12]	16	53.6	85.9	29.8	65.4	40.5	80.0	59.4	90.3	34.1	69.6	45.4	84.7
Fwd2Bot (Ours)	16	83.8	98.5	59.0	88.6	72.3	96.5	94.3	99.9	72.9	94.4	85.7	99.5

Table 10. Comparison on the SugarCrepe compositionality benchmark.

Method	Tokens	Replace			Swap		Add	
		Object	Attribute	Relation	Object	Attribute	Object	Attribute
LLaVA-1.5-7B [28]	576	88.0	81.6	76.1	60.9	58.8	67.0	62.4
PruMerge [35]	18	88.0	74.4	69.7	62.5	57.3	81.4	66.0
Matryoshka Multi. [3]	16	90.3	81.4	80.1	70.2	67.9	75.7	75.8
Matryoshka Query [12]	16	89.3	81.4	79.2	70.6	64.7	73.8	73.6
Fwd2Bot (Ours) (LLaVA-1.5-7B)	7.06	98.1	89.5	82.7	77.8	78.1	95.3	93.1

Table 11. Comparison with various token compression methods on TextCaps dataset for image captioning in terms of BLEU-4 (B@4), CIDEr score, METEOR (MET.) and ROUGE-L.

Method	Tokens	B@4	CIDEr	MET.	ROUGE
LLaVA-1.5 [28]	576	27.1	90.4	21.9	46.2
PruMerge [35]	≈32	17.6	62.8	17.0	39.7
Matryoshka Multi. [3]	36	25.1	94.8	23.0	46.3
Matryoshka Query [12]	36	21.0	70.0	19.9	42.6
Fwd2Bot (Ours)	32	26.5	90.6	22.4	46.1
Matryoshka Query [12]	16	20.1	62.5	19.3	41.7
Fwd2Bot (Ours)	16	26.4	90.5	22.5	46.3
Matryoshka Query [12]	4	15.2	42.0	16.5	37.4
Fwd2Bot (Ours)	4	25.4	86.1	22.0	45.7

attending the future ones. While necessary for autoregressive modeling, it’s unclear why it would be for vision token compression too, as there is no preferential direction for image processing. Hence, herein, we explore the effect of changing the attention pattern from causal to bidirectional (*i.e.* full) attention for the compression forward pass, while keeping it causal for the subsequent answer generation ones. In this instance, the stage (*i.e.* compression vs generative) specific LoRAs also take the role of adjusting the attention pattern and information flow. Analyzing the results from Table 12 we can observe gains for discriminative tasks and degradation for generative ones. This suggests that a direct finetuning under a different attention pattern is suboptimal, likely requiring a pre-alignment step. Moreso, the LoRA adapters may limit the ability of the model in shifting its attention pattern.

Finetuning checkpoint choice: The natural starting point for our approach is the LLaVA model itself. However, for completeness, we also try to directly finetune from the Vicuna LLM itself. As the results from Table 13 show, starting

Table 12. Compression with Bidirectional vs Causal attention for generation (MMB, MME, TextVQA) and retrieval (Flickr30K, MS-COCO).

Method	MMB	MME	TextVQA	Flickr30K		MS-COCO	
				T2I	I2T	T2I	I2T
Bidirectional	60.2	1310.1	48.4	83.6	94.8	57.9	72.2
Causal	64.4	1470.0	54.2	83.8	94.4	56.8	70.2

from a model already optimized for vision-language understanding results in a notable performance boost. To compensate for this, likely, a longer training scheduler is needed and potentially a full model finetuning, as in LLaVA.

Table 13. Impact of the pre-trained checkpoint for generation (MMB, MME, TextVQA) and retrieval (Flickr30K, MS-COCO).

Method	MMB	MME	TextVQA	Flickr30K		MS-COCO	
				T2I	I2T	T2I	I2T
Vicuna	60.3	1296.3	48.2	81.2	92.5	54.3	67.4
LLaVA	64.4	1470.0	54.2	83.8	94.4	56.8	70.2