ABC: Achieving Better Control of Visual Embeddings using VLLMs

Anonymous authors
Paper under double-blind review

Abstract

Visual embedding models excel at zero-shot tasks like visual retrieval and classification. However, these models cannot be used for tasks that contain ambiguity or require user instruction. These tasks necessitate an embedding model which outputs can use a natural language instruction to control the representation of a visual embedding. Existing CLIP-based approaches embed images and text independently, and fuse the result. We find that this results in weak interactions between modalities, and poor user control over the representation. We introduce **ABC**, an open-source multimodal embedding model that uses a vision-language model backbone to deeply integrate image features with natural language instructions. **ABC** achieves best-for-size performance on MSCOCO image-to-text retrieval and is the top performing model on classification and VQA tasks in the Massive Multimodal Embedding Benchmark. With a strongly unified vision-language representation, **ABC** can use natural language to solve subtle and potentially ambiguous visual retrieval problems. To evaluate this capability, we design `CtrlBench`, a benchmark that requires interleaving textual instructions with image content for correct retrieval. **ABC** advances the state of visual embeddings, outputting high-quality visual representations with natural language control.

1 Introduction

Visual embeddings have become a foundational representation in computer vision. Image embedding models have become the state of the art for many zero-shot tasks, including visual retrieval (Chen et al., 2024) and image classification (Yu et al., 2022). Since the release of CLIP (Radford et al., 2021), its dual encoder architecture has remained the state of the art for producing high-quality visual embeddings (Yu et al., 2022; Sun et al., 2023; Chen et al., 2024). However, CLIP only supports *separately* embedding images or text (Radford et al., 2021). Therefore, complex visual embedding tasks which require additional specification are impossible. For example, a CLIP model cannot distinguish which is the correct answer in Figure 1. Both candidates are plausible unless the user provides additional instruction. Therefore, an embedding that interleaves vision and natural language is essential.

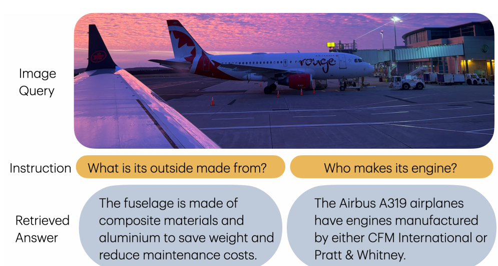


Figure 1: A visual embedding where natural language is required to solve ambiguity

We find that existing approaches suffer from two problems: **(1)** Reliance on generic or human-annotated instructions (Jiang et al., 2024b; Zhang et al., 2025a). Unlike the abundant noisy image-text data used to train CLIP models Schuhmann et al. (2022); Changpinyo et al. (2021); Sharma et al. (2018), datasets containing multiple modalities and diverse instruction phrases are significantly more scarce and challenging to curate at scale. This is problematic for embedding models, which require massive datasets for the best representations (Cherti et al., 2023a). Often vague instructions are used repetitively during training, resulting in overfitting of the instruction (Wei et al., 2023; Gudibande et al., 2023). **(2)** Weak interaction between modalities. Previous works fuse embeddings outputted by CLIP models Zhang et al. (2024); Wei et al.

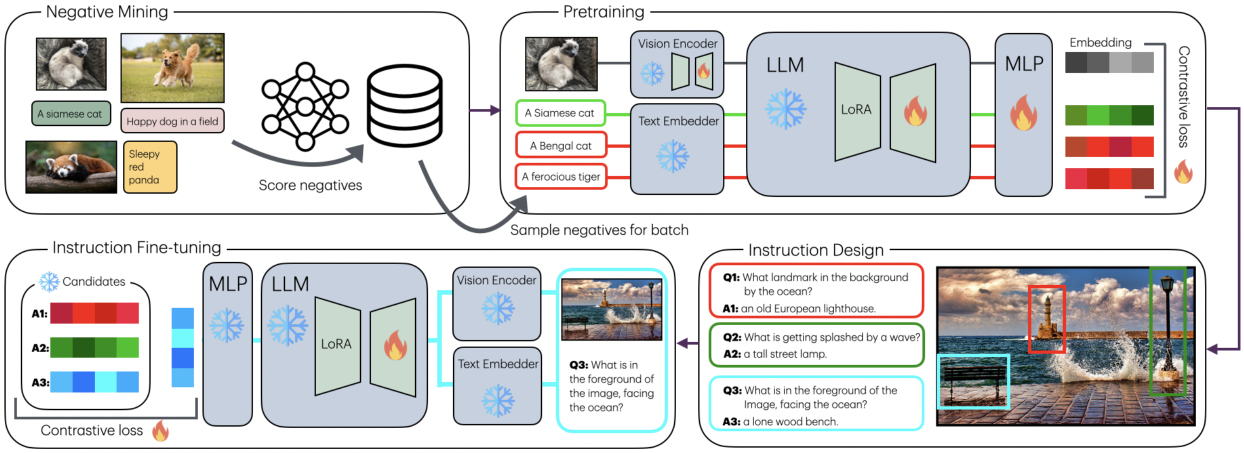


Figure 2: An overview of **ABC**'s training regime. First, we mine negatives to augment each batch in our pretraining dataset with *almost plausible* text candidates. Then we pretrain our model using contrastive loss on independently embedded text-image pairs. Once our model has learned to output high-quality uncontrolled visual embeddings, we introduce instructions. We create multiple instructions and candidate captions for a single image, with each one focusing on different aspects of the image. Lastly, we fine-tune using multiple captions for same image, the positive caption for each query corresponds to the natural language instruction. Therefore, **ABC** uses the instruction to distinguish the correct candidate for the query.

(2023). This approach prevents deeper interaction between modalities, resulting in superficial use of the instructions Jiang et al. (2024b).

To this end, we introduce **ABC**, a model that uses natural language instructions to control visual embeddings. **ABC**'s Vision Large language Model (VLLM) backbone allows it to integrate natural language instructions when crafting visual embeddings. We find that training our model has two fundamental challenges: **(1)** Extracting useful contrastive embeddings from a pretrained generative VLLM. **(2)** Designing an instruction fine-tuning method that lets users modify multimodal embeddings using natural language instructions. To train **ABC**, we adopt a multi-stage training process. In the initial pretraining stage, we use contrastive training with carefully selected negatives to develop a model that generates embeddings, similar to CLIP. In the instruction fine-tuning stage, we train a lightweight adapter to project queries consisting of images and natural language instructions into the embedding space of the pretrained model. Our synthetic instructions to correspond to different aspects *of the same image*, therefore, **ABC** learns to use the instruction for fine-grained control. Our training results in a model that produces powerful and controllable visual embeddings.

ABC achieves impressive zero-shot performance in retrieval, classification, and visual question answering (VQA) tasks. In MSCOCO (Lin et al., 2015) image-to-text retrieval, our model outperforms all CLIP models containing at most 8 billion parameters. Furthermore, our model outperforms all other models on the zero-shot classification and VQA splits of MMEB (Jiang et al., 2024b), a multimodal embedding benchmark spanning 19 tasks. Lastly, we design `CtrlBench` to measure our model's ability to use natural language instructions to control retrieval. Using `CtrlBench`, we show that **ABC** can accomplish visual retrieval tasks that are fundamentally ambiguous without utilizing natural language instructions.

Our contribution is threefold. **(1) ABC**: an open-source multimodal embedding model that uses natural language instructions to control visual embeddings. We demonstrate that **ABC** produces powerful embeddings by benchmarking on zero-shot tasks *and* robustly utilizes natural language instructions to control embeddings. We study how key design choices—including vision encoder resolution, contrastive loss temperature, and VLLM backbone—impact embedding quality. **(2)** We provide a *decoupled* methodology for adapting VLLMs into dense embedding models. Previous work integrates instructions during the large-scale contrastive training run. We show that these stages can be decoupled, resulting in a lightweight and adaptable fine-tuning stage that requires only 100 training steps using synthetic instructions. **(3)** Lastly, we introduce `CtrlBench`, a novel benchmark for measuring instruction-controlled retrieval. `CtrlBench` *requires* the model to interleave modalities to retrieve the correct response.

2 Related Works

Visual embeddings. Since Radford et al. (2021) introduced CLIP, various aspects of the original model have been improved. DataComp (Gadre et al., 2023) demonstrated the importance of data filtering during contrastive pretraining to create CLIP models. Training CLIP models requires large-scale pretraining and device parallelization to achieve the throughput and batch sizes needed to produce the best performing models (Cherti et al., 2023b; Sun et al., 2023; Yu et al., 2022). InternVL-G Chen et al. (2024) is the largest and best performing open-weight CLIP model to date, a 14 billion parameter model trained to create a vision encoder for the InternVL family of VLLMs.

Multimodal embeddings. MagicLens (Zhang et al., 2024) and UniIR (Wei et al., 2023) are multimodal embedding models that fuse CLIP embeddings. By combining multiple CLIP models, they take image-text pairs and align them with image-text pairs using contrastive loss. VLM2Vec Jiang et al. (2024b) and E5-V Jiang et al. (2024a) also use VLLM backbones to produce multimodal embeddings. VLM2Vec uses contrastive training to align text-image queries to text-image candidates; alternatively, E5-V is trained only by aligning text pairs. Jiang et al. (2024b) introduced *Massive Multimodal Embedding Benchmark* (MMEB), a benchmark for a variety of multimodal embedding tasks.

Extracting Embeddings from LLMs. Several architectural modifications have been proposed for adapting LLMs to produce dense embeddings Wang et al. (2024a); BehnamGhader et al. (2024); Lee et al. (2024). LLM2VEC BehnamGhader et al. (2024) changed the attention mask to be bidirectional, which allows information to flow between all tokens, improving embedding quality. Several methods for extracting dense embedding from LLMs have been proposed, these include picking the last token Jiang et al. (2024b), mean token pooling BehnamGhader et al. (2024) or an additional attention layer Lee et al. (2024). In the text-only domain, the Mistral-7b Jiang et al. (2023) backboned NV-Embed Lee et al. (2024) has achieved the top score on MTEB Muennighoff et al. (2023), a popular text embedding benchmark.

Mined negatives and synthetic instructions. Negative mining is a technique in contrastive training where *somewhat relevant candidates* are included in each batch de Souza P. Moreira et al. (2024). This improves the quality of embeddings by ensuring they can differentiate between subtly different candidates. This is often used for training text embedding models, and features prominently in NV-Embed's training Lee et al. (2024). Synthetic instruction data has recently emerged as a way to scale instruction data in multimodal embedding models Xiao et al. (2025); Zheng et al. (2024). MegaPairs Zhou et al. (2024) uses a VLLM to create open-ended instructions to train multimodal retrievers. They use this data synthesis pipeline to train two versions of their MMRet model, one uses CLIP Radford et al. (2021) as its backbone and the other uses LLava-NeXT Liu et al. (2024a).

3 ABC

Figure 2 is an overview of our training regime and model architecture. Our training regime consists of 2 distinct stages; pretraining and instruction fine-tuning. The pretraining stage uses self-supervised training on image-caption pairs to adapt features used for generative modeling into features for dense embeddings. As our pretraining does not require instructions, it can be easily scaled using any large image-captioning data source (Changpinyo et al., 2021; Schuhmann et al., 2022). In the second stage, we train using a small set of queries consisting of images and synthetic text instructions. We create synthetic instructions that correspond to different aspects of the same image and use these as negatives. This approach improves **ABC**'s ability to use instructions to control visual representations.

3.1 Model Design

Pretraining Architecture. During pretraining images and text are embedded independently, and aligned using contrastive loss, much like CLIP (Radford et al., 2021). Text and images are converted to hiddens $\boldsymbol{h}^{(0)} \in \mathbb{R}^{s \times d}$ using the embedding layer and vision encoder, respectively. The LLM module is shared between both modalities. To adapt the LLM to output dense embeddings, we make several architectural changes. Following BehnamGhader et al. (2024), we enable bidirectional attention, allowing all tokens to attend to all other tokens. To create our dense embeddings, we average over the sequence dimension in the last hidden layer $\boldsymbol{h}^{(l)}$ (equation 1), and project the result $\boldsymbol{x} \in \mathbb{R}^d$ using a residually connected MLP layer (equation 2).

$$\boldsymbol{x} = \frac{1}{s}\sum_{i}^{s} \boldsymbol{h}_i^{(l)} \tag{1}$$

$$MLP(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{A}g(\boldsymbol{B}\boldsymbol{x}) \tag{2}$$

Where $\boldsymbol{A}$ and $\boldsymbol{B}$ are parameter matrices and g is the element-wise SELU function (Klambauer et al., 2017). To train our backbone, we apply LoRA (Hu et al., 2021) adapters on both the vision encoder and LLM modules. We optimize the contrastive loss temperature hyperparameter τ during pretraining.

Instruction Fine-tuning Architecture. After pretraining, **ABC** can project queries consisting of images *or* text into a shared embedding space. However, to control the representation of visual embeddings, we must project queries consisting of images *and* text instructions into the embedding space. To accomplish this, we train a low-rank LoRA adapter on the LLM component. The adapter contrastively aligns image and instruction queries with text candidates embedded using the pretrained model. We fuse the pretrained LoRA weights into the base model and freeze the MLP adapter layer. Following the standard VLLM instruction format, we insert the text instruction tokens after the image tokens in our prompt template:

`<Image> Instruction : <Instruction>` (3)

We do not back-propagate through text candidate embeddings, only the queries containing an image and instruction. This ensures that our image-text queries share the same features as text and images embedded using the pretrained model. We can easily alternate between embedding with or without instructions by disabling the instruction fine-tuned LoRA adapter. We freeze the temperature τ during fine-tuning.

3.2 Pretraining Data

To create our pretraining dataset we employ negative mining on Conceptual Captions (Sharma et al., 2018). We derive the mined negatives for our dataset as follows: **(1)** We do a small pretraining run using only in-batch negatives. **(2)** We use the resulting model to calculate similarity scores between all images and captions in our pretraining dataset. This approach avoids a circular dependence on a third-party embedding model to train our embedding model. Therefore, it is easily extensible to modalities where an existing embedding model is not publicly available. **(3)** To prevent our negatives from being too similar to our positive samples, we set a similarity threshold $\epsilon \in [0, 1]$. We only sample negatives that have a similarity score of at most ϵ times the similarity score of the correct candidate. We randomly choose our mined hard negatives from the 100 candidate captions below the threshold. This results in the text candidates shown in Figure 3. The mined text negatives are clearly relevant, but the correct caption is still the best answer.

Figure 3: A sample from our pretraining dataset. The positive caption (green) is the best caption for the image. The mined negatives (red) are relevant but not the best choice.

In our pretraining run, each batch consists of N image queries (without instructions) and M text candidates. We include $M - N$ mined text negatives in each batch. Therefore, each image query has $\frac{M}{N} - 1$ corresponding mined negatives. Our pretraining loss function is given by Equation (4).

$$\mathcal{L}_{pretrain} = -\sum_{i}^{N} \log \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{y}_i/\tau)}{\sum_{j=1}^{N}(\exp(\frac{\boldsymbol{x}_i^\top \boldsymbol{y}_j}{\tau}) + \sum_{k=1}^{\frac{M}{N}-1}\exp(\frac{\boldsymbol{x}_i^\top \boldsymbol{n}_j^k}{\tau}))} \tag{4}$$

For a given query $\boldsymbol{x}_i$, $\boldsymbol{n}_i^k$ is its k_{th} corresponding mined hard negative and and $\boldsymbol{y}_i$ is its positive caption. τ is the temperature hyperparameter used to scale the loss. $\boldsymbol{n}_j^k$ is a mined negative for $\boldsymbol{x}_i$ when $i = j$, otherwise it acts as a regular in-batch negative. The embeddings for $\boldsymbol{x}_i$, $\boldsymbol{y}_i$ and $\boldsymbol{n}_j^k$ are unit normalized before loss is computed. We scale the loss by $\frac{1}{N}$, the number of image queries in the batch.

3.3 Instruction Fine-tuning Data

To create our instruction fine-tuning dataset we use Visual Genome (Krishna et al., 2016), a dataset comprised of images with captioned bounding boxes. When choosing which bounding boxes to use for each image, we filter by the size of the bounding box (width × height). We exclude the 5 largest bounding boxes, as we find they often do not represent specific objects or aspects of the scene. For each image, we randomly choose 4 bounding boxes and their respective captions. We then prompt GPT-4o (OpenAI, 2024) to create instructions corresponding to each bounding box caption. We sample multiple captions from each image so that they can be used as negatives for each other during instruction fine-tuning. This ensures that each image query has multiple realistic captions in its batch. Therefore, *utilizing the instruction is required to choose the correct text candidate unambiguously.* We provide the prompt and generation settings used to create our instruction fine-tuning dataset in Appendix A. We instruction fine-tune using exclusively in-batch negatives. However, we group all queries that contain the same image into the same batch. This ensures that queries always have multiple relevant text candidates.

4 Experiments

We evaluate two aspects of the model. As controlling a poor-quality embedding is useless, we first assess the quality of embeddings output by **ABC**. We evaluate our model on a variety of zero-shot retrieval and classification tasks (Section 4.2). We study how variable resolution capabilities (Section 4.3), the settings for τ when contrastively training a VLLM (Section 4.4) and VLLM backbone choice (Section 4.5) effect embedding quality. Secondly, we measure **ABC**'s ability to use natural language to control visual embeddings on several visual question answering tasks (Section 4.6). Motivated by shortcomings in VQA benchmarks, we construct `CtrlBench`. A benchmark designed to require the model to jointly use the image and instruction to retrieve the most appropriate caption.

4.1 Training Settings

We initialize **ABC** using Qwen2-VL-7B (Wang et al., 2024b). We train our negative mining model using only in-batch negatives with a batch size of 256 for 1000 steps. We set $\epsilon = 0.95$ and sample 7 mined negatives for each image. We pretrain using batches of 512 image queries and 4096 text candidates sharded across 8 NVIDIA A100-SXM4-80GB GPUs (Qu et al., 2021) for 4000 steps. We use a LoRA adapter with a rank of 64 and an alpha of 128. We limit the number of tokens output by the vision encoder to 512 during

	MSCOCO (5K test set)						**Flickr30K** (1K test set)					
	Image → Text			Text → Image			Image → Text			Text → Image		
Model	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP (Radford et al., 2021)	58.4	81.5	88.1	37.8	62.4	72.2	88.0	98.7	99.4	68.7	90.6	95.2
ALIGN Jia et al. (2021)	58.6	83.0	89.7	45.6	69.8	78.6	88.6	98.7	99.7	75.7	93.8	96.8
FLAVA (Singh et al., 2022)	42.7	76.8	-	38.4	67.5	-	67.7	94.0	-	65.2	89.4	-
FILIP * (Yao et al., 2021)	61.3	84.3	90.4	45.9	70.6	79.3	89.8	99.2	99.8	75.0	93.4	96.3
CoCa * (Yu et al., 2022)	66.3	86.2	91.8	51.2	74.2	82.0	92.5	**99.5**	**99.9**	**80.4**	**95.7**	**97.7**
OpenCLIP-G (Cherti et al., 2023a)	67.3	86.9	92.6	**51.4**	74.9	**83.0**	92.9	99.3	99.8	79.5	95.0	97.1
EVA-02-CLIP-E+ (Sun et al., 2023)	68.8	87.8	92.8	51.1	**75.0**	82.7	**93.9**	99.4	99.8	78.8	94.2	96.8
ABC (ours)	**69.2**	**87.9**	**93.2**	47.6	72.1	80.6	90.7	99.0	99.5	74.6	92.6	95.45

Table 1: Comparison of retrieval performance on MSCOCO Lin et al. (2015) and Flickr30K Plummer et al. (2016) datasets (Karpathy split). Best performance is **bold**, second best is underlined. * indicates a closed-weight model.

	CLIP	OpenCLIP	SigLIP	UniIR	MagicLens	BLIP2	MMRet	ABC (ours)
Classification (9 tasks)								
ImageNet-1K Russakovsky et al. (2015)	55.8	63.5	45.4	58.3	48.0	10.3	49.1	**71.2**
HatefulMemes Kiela et al. (2021)	51.1	51.7	47.2	**56.4**	49.0	49.6	51.0	52.1
VOC2007 Everingham et al.	50.7	52.4	64.3	66.2	51.6	52.1	74.6	**81.4**
SUN397 Xiao et al. (2010)	43.4	68.8	39.6	63.2	57.0	34.5	60.1	**71.8**
Place365 López-Cifuentes et al. (2020)	28.5	37.8	20.0	36.5	31.5	21.5	35.3	**40.7**
ImageNet-A Djolonga et al. (2020)	25.5	14.2	42.6	9.8	8.0	3.2	31.6	**49.4**
ImageNet-R Hendrycks et al. (2021)	75.6	83.0	75.0	66.2	70.9	39.7	66.2	**86.8**
ObjectNet Barbu et al. (2019)	43.4	51.4	40.3	32.2	31.6	20.6	49.2	**67.7**
Country-211 Radford et al. (2021)	**19.2**	16.8	14.2	11.3	6.2	2.5	9.3	18.5
All Classification	43.7	48.8	43.2	44.5	39.3	26.0	47.4	**60.0**

Table 2: Zero-shot classification results on MMEB (Jiang et al., 2024b). We compare with pretrained CLIP models (Radford et al., 2021; Cherti et al., 2023a), instruction finetuned models derived from CLIP (Wei et al., 2023; Zhang et al., 2024) and other LLM backbone approaches (Li et al., 2023; Zhou et al., 2024).

training. For our optimizer, we use AdamW (Loshchilov & Hutter, 2019) with a learning rate of 4×10^{-5}, betas of 0.9 and 0.999 and a weight decay of 10^{-3}. We warm-up for 3% of training steps and initialize the temperature τ as 7×10^{-2}. In our instruction fine-tuning stage, we use a lower rank LoRA adapter with rank and alpha of 16 and 32, respectively. Our instruction fine-tuning stage can be short, as our VLLM backbone is already instruction fine-tuned. Therefore, we only instruction fine-tune for 100 steps. Each batch contains 128 unique images, with each image appearing four times, paired with a different instruction and a corresponding positive text candidate.

Modifications to save VRAM. Due to our use of LoRA Hu et al. (2021), the VRAM requirements for gradients and optimizer state is relatively low. Consequentially, the model activations used by autograd for the backward pass account for most of VRAM used during training. To address this, we make aggressive use of activation checkpointing, recomputing the activations of each decoder block during the backward pass. Furthermore, we modify the VLLM backbone to skip the logits and cross-entropy loss computation. With the large vocabulary size of modern LLMs, the output layer has become the most memory-intensive layer (Wijmans et al., 2024). As we only use the hidden state for calculating embeddings, the logits tensor is unnecessary. We find that this simple change saves up to 11 GB of memory per device, allowing us to use to a significantly larger batch size.

4.2 Zero-shot Retrieval and Classification

We first evaluate the quality of **ABC** embeddings, as controlling low-quality visual embeddings is useless. To test embedding quality, we use two zero-shot tasks: retrieval from a pool of candidates and image classification. For retrieval, we use MSCOCO (Lin et al., 2015) and Flickr30K (Plummer et al., 2016). For image classification, we use MMEB Jiang et al. (2024b), which provides splits of many standard image classification tasks, such as ImageNet-1K (Russakovsky et al., 2015), VOC-2007 (Everingham et al.), and ObjectNet (Barbu et al., 2019). We note that the classification labels in MMEB are short, often only one or two words. This is problematic for image-captioning models that have been trained on full sentences (Yu et al., 2022). To alleviate this issue, we use Radford et al. (2021)'s technique of embedding classification labels in a sentence template. We use "`A photo of a {label}.`" as our template for all classification evaluations.

In Table 1, our model demonstrates strong image-to-text retrieval capabilities, achieving competitive performance to models that have been contrastively trained with hundreds of GPUs and massive batch sizes (Sun et al., 2023; Cherti et al., 2023a). We achieve the best performance in MSCOCO image-to-text retrieval. Comparatively, our text-to-image retrieval is weaker. This follows from our training, where we use image-to-text negatives, not text-to-image negatives. Averaged across all classification tasks (Table 2), we achieve 11.2% better accuracy than the next-best model . On ImageNet-1K (Russakovsky et al., 2015), **ABC** has 7.7% better zero-shot accuracy than OpenCLIP. Overall, **ABC**'s performance on established benchmarks such as MSCOCO, Flickr30K and ImageNet-1K indicates that the pretrained model's embeddings are useful.

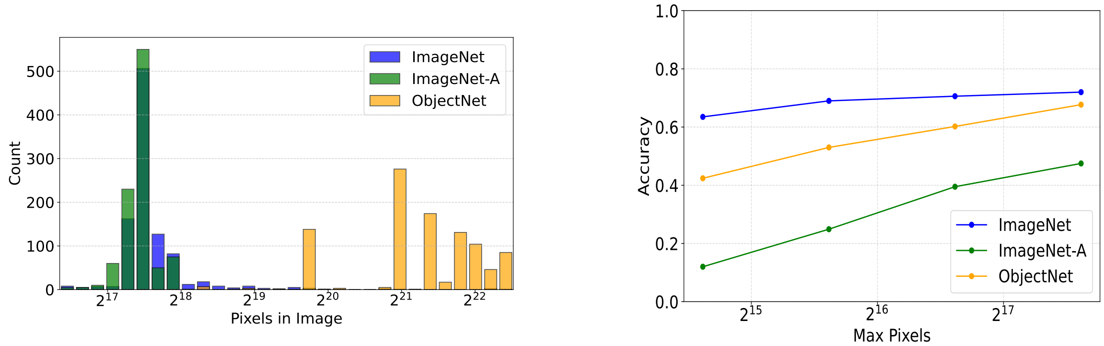


Figure 4: **(Left)** Pixel distributions of benchmarks. ObjectNet (Barbu et al., 2019) contains significantly higer resolution than ImageNet (Russakovsky et al., 2015), which is downscaled. **(Right)** Accuracy on benchmarks when increasing the pixel count (tokens) used by vision encoder when creating embeddings.

In Section 4.6, we evaluate the fine-tuned adapter's ability to embed images with text instructions into the pretrained model's embedding space.

4.3 Image Resolution and Embedding Quality

Our VLLM backbone, Qwen2-VL-7B (Wang et al., 2024b), supports image inputs with variable resolution. This allows the user to effectively trade-off image resolution with inference speed by adjusting the number of tokens output by the vision encoder. We examine how this trade-off influences embedding quality for image classification. We find that performance on certain tasks, like ObjectNet (Barbu et al., 2019) and ImageNet-A (Djolonga et al., 2020), strongly correlate with the resolution used in the VLLM vision encoder (Figure 4). On average, ObjectNet images have 13 times more pixels than those from ImageNet-1K. When the number of tokens produced by the vision encoder is scaled up, we see a large improvement on ObjectNet, with accuracy increasing by 23.4%. However, lower resolution benchmarks like ImageNet-1K do not benefit nearly as much from scaling resolution. Notably, a smartphone camera takes photos with significantly higher resolution than images in ImageNet-1K or ObjectNet, by default (Apple Inc., 2024). This motivates the need for benchmarks containing larger resolution images. Existing low-resolution benchmarks may underestimate the capabilities of models that can natively utilize high-resolution images. Furthermore, we find that correctly classifying the "natural adversarial examples" of ImageNet-A (Djolonga et al., 2020) is largely a function of resolution. Our accuracy on ImageNet-A increases from only 12% to 47.5% simply by scaling the resolution used by the vision encoder during evaluation.

4.4 Temperature and Loss Dynamics

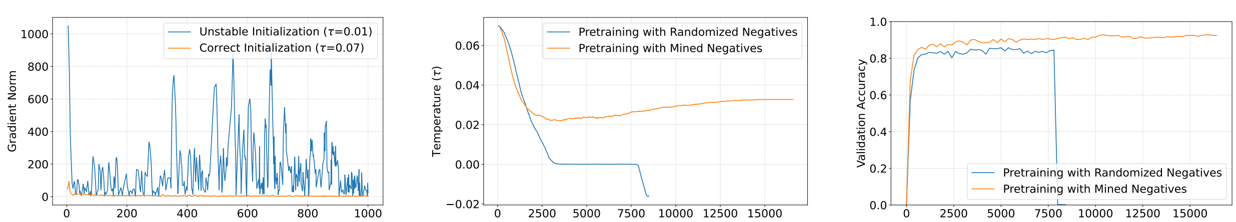


Figure 5: Visualization of temperature (τ) and its effects on training dynamics. **(Left)** The effect of temperature initializations on gradient norm while training our negative mining model. **(Middle)** Temperature behavior during pretraining, without and without mined negatives. **(Right)** Validation accuracy during pretraining, with and without mined negatives.

Model	$\text{Accuracy}_{\text{val}}$	MMLU_{val}	$\text{MMBench}_{\text{EN}}$
LLaVA-NeXT	62.6	35.3	68.7
InternVL-2-8B	65.9	51.8	81.7
Qwen2-VL-7B	**70.2**	**54.1**	**83.0**

Table 3: Validation accuracy vs. generative task performance.

The setting of the temperature hyperparameter (τ) is known to be crucial for contrastive training (Jia et al., 2021). However, τ is often treated as a fixed hyperparameter (Jiang et al., 2024b; Wei et al., 2023). A large fixed setting of τ results in a significantly worse model (Wang & Liu, 2021), whereas a too small setting can result in training instability (Figure 5). Therefore, we find that optimizing τ throughout the pretraining process is crucial. In shorter runs like training our negative mining model, we find that shows that *both* the initialization and optimization of τ is crucial. For our negative mining model, we find $\tau = 0.07$ to be a temperature initialization that is both stable and performant.

Temperature in pretraining. During pretraining, we find that *the mined negatives are essential for training stability.* Figure 5 shows our pretraining run with a batch size of 128 images queries and 1024 text candidates. If randomized negatives are used instead of our chosen mined negatives, the optimizer tends to push τ very close to 0. This results in numerical instability, loss spikes, and eventually catastrophic failure. When our mined negatives are used in our pretraining, the temperature stabilizes at a very reasonable value of around 0.03.

4.5 VLLM Backbone

We explore how the choice of VLLM backbone has an effect on the quality of our multimodal embedding. We ablate over 3 popular choices of VLLM backbone, all at approximately the 8 billion parameter scale: Qwen2-VL-7B (Wang et al., 2024b), our chosen backbone. InternVL-2-8B from the internVL family of VLMs (Chen et al., 2024). Lastly, LLaVA-NeXT (Liu et al., 2024a) with Mistral-7B (Jiang et al., 2023) as its LLM component. We pretrain each model for 1000 steps with a query batch size of 128 and candidate batch size of 1024. Table 3 shows the validation accuracy of **ABC** with different backbones. We find that our backbone choice, Qwen2-VL-7B, produces the best results. We also note each backbone's performance on two standard generative VLLM benchmarks: MMMU Yue et al. (2024) and MMBench Liu et al. (2024b). We find that performance after contrastive training strongly correlates with the performance of the backbone on generative tasks. This indicates that training better VLLMs naturally results in better backbones for our embedding model.

4.6 Controlling Embeddings using Natural Language

Visual Question Answering is a widely adopted task used to measure control over visual embeddings (Zhang et al., 2025b; Wei et al., 2023; Jiang et al., 2024b). For a VQA task, the model creates a

	CLIP	**OpenCLIP**	**SigLIP**	**UniIR**	**MagicLens**	**BLIP2**	**MMRet**	**ABC (ours)**
VQA (10 tasks)								
OK-VQA Marino et al. (2019)	7.5	11.5	2.4	25.4	12.7	8.7	<u>28.0</u>	**48.1**
A-OKVQA Schwenk et al. (2022)	3.8	3.3	1.5	8.8	2.9	3.2	<u>11.6</u>	**37.3**
DocVQA Mathew et al. (2021b)	4.0	5.3	4.2	6.2	3.0	2.6	<u>12.6</u>	**28.5**
InfographicsVQA Mathew et al. (2021a)	4.6	4.6	2.7	4.6	5.9	2.0	**10.6**	<u>7.9</u>
ChartQA Masry et al. (2022)	1.4	1.5	<u>3.0</u>	1.6	0.9	0.5	2.4	**11.7**
Visual7W Zhu et al. (2016)	4.0	2.6	1.2	<u>14.5</u>	2.5	1.3	9.0	**25.6**
ScienceQA Lu et al. (2022)	9.4	10.2	7.9	12.8	5.2	6.8	<u>23.3</u>	**26.3**
VizWiz Gurari et al. (2018)	8.2	6.6	2.3	24.3	1.7	4.0	<u>25.9</u>	**29.4**
GQA Hudson & Manning (2019)	41.3	52.5	<u>57.5</u>	48.8	43.5	9.7	41.3	**60.1**
TextVQASingh et al. (2019)	7.0	10.9	1.0	15.1	4.6	3.3	<u>18.9</u>	**35.4**
All VQA	9.1	10.9	8.4	16.2	8.3	4.2	<u>18.4</u>	**31.0**

Table 4: Zero-shot VQA results on MMEB (Jiang et al., 2024b). Best is **bold**, second best is <u>underlined</u>.

joint embedding of an image and a pertaining question. The model is expected to relate the joint embedding to a text embedding of the question's answer, measured by cosine similarity. The model succeeds if the distance to the correct answer is the closest among a pool of text candidates. We measure **ABC**'s VQA abilites using the VQA split of MMEB (Jiang et al., 2024b), which contains 10 unique VQA tasks.

Table 4 shows **ABC**'s performance on the VQA split of MMEB. In just 100 training steps using instructions, our model surpasses MMRet (Zhou et al., 2024) on VQA, which was trained with millions of instructions. Interestingly, CLIP derivatives fine-tuned with instructions such as MagicLens (Zhang et al., 2024) and UniIR (Wei et al., 2023) do not perform significantly better than their respective baselines, despite MMEB providing instructions for each task. This evidences that encoder-only embedding models struggle to effectively utilize natural language instructions (Jiang et al., 2024b).

Problems with VQA. A VQA benchmark should require the use of both modalities *together*, and using only the image or the text should be insufficient to accomplish the task. We note that ensuring this property requires inspecting not only the queries but the candidate pool as well. Consider Figure 6, an example from the Visual7W task in MMEB. When the candidate pool is examined,[1] it is clear that there is only one answer that is plausible, given the image. Therefore, the instruction is not required to accomplish the VQA task. This a common pitfall when adapting open-ended generative VQA benchmarks into fixed candidate pool embedding tasks.

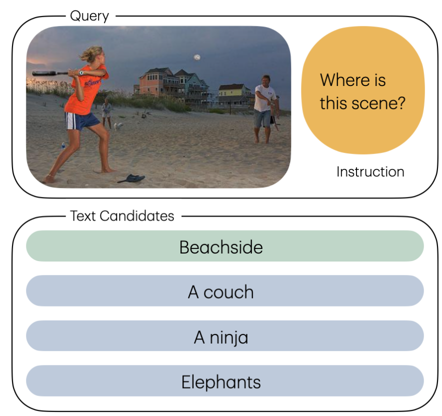


Figure 6: An example of a VQA task from the Visual7W (Zhu et al., 2016) split of MMEB (Jiang et al., 2024b).

`CtrlBench`. Motivated by this shortcoming in VQA, we construct `CtrlBench`, an instruction-controlled retrieval benchmark. `CtrlBench` has a similar format to the Flicker30K (Plummer et al., 2016) test split. However, instead of each image having multiple valid captions, we instead provide an instruction from which the model can infer the most relevant text candidate. Therefore, `CtrlBench` tests both retrieval and instruction following capabilities. To construct `CtrlBench`, we sample a 1000 test images from VisualGenome (Krishna et al., 2016). To create 5000 instructions and text candidate pairs, we generate 5 instructions for each image. As each image has 5 associated captions in the candidate pool, a model that cannot utilize instructions can (in expectation) achieve at most 20% R@1 on the benchmark. To differentiate `CtrlBench` from our fine-tuning distribution, we use Gemini 2.5 Flash (Team et al., 2025) to rephrase all instructions and candidates. We deduplicate captions from the dataset to prevent ambiguity in our text candidates, resulting in `CtrlBench` being slightly smaller than 1000 images.

Model	R@1	R@5	R@10
UniIR Wei et al. (2023)	0.0	0.0	0.1
MagicLens Zhang et al. (2024)	11.7	28.4	38.5
VLM2Vec Jiang et al. (2024b)	26.44	56.5	68.9
GME-Qwen2VL-2B Zhang et al. (2025a)	9.4	24.6	33.9
GME-Qwen2VL-7B Zhang et al. (2025a)	10.7	27.6	36.9
ABC (ours)	**41.0**	**56.5**	**80.0**

Table 5: Recall@K on `CtrlBench`.

Table 5 shows the performance of multimodal embedding models on `CtrlBench`. We compare with 2 instruction fine-tuned CLIP derivatives: UniIR Wei et al. (2023) and MagicLens Zhang et al. (2024) as well as VLM2Vec Jiang et al. (2024b) and GME (Zhang et al., 2025a), two concurrent works that adapt VLLMs into multimodal embedding models. We find that neither of the CLIP-based architectures have above 20% R@1 on `CtrlBench`, the performance that indicates that the model is non-trivially utilizing instructions. We

[1]For brevity, the candidate pool in Figure 6 consists of a subset of the whole candidate pool. Nevertheless, the entire candidate pool for the query suffers from being irrelevant to the instruction.

find that VLLM architectures are better at instruction-controlled retrieval. **ABC** and VLM2Vec have R@1 above 20%, proving that they can non-trivially utilize instructions.

5 Discussion and Limitations

5.1 Distribution Overlap in Training and Evaluation

Throughout our experiments, we found that several training data sources overlap with commonly used benchmarks. Several benchmark are constructed from the same underlying data source as training data or directly derived from training data Barbu et al. (2019); Russakovsky et al. (2015); Plummer et al. (2016); Lin et al. (2015). This makes evaluating "zero-shot" capabilities difficult. Our fine-tuning dataset and `CtrlBench` sample 12,800 and 1,000 images from VisualGenome (Krishna et al., 2016), respectively. We mitigate distribution overlap between `CtrlBench` and our fine-tuning data by using an LLM to rephrase all text in `CtrlBench`. However, this is not a perfect solution, and a key limitation in our work. To the author's knowledge, our pretraining data (Sharma et al., 2018) does not overlap with any benchmarks we evaluate against, and all other evaluations (Tables 1, 2 and 4) are out-of-distribution.

5.2 Decoupling Pretraining and Instruction Fine-tuning

We find that adapting an instruction fine-tuned VLLM to use instructions is relatively cheap. Our instruction fine-tuning completes in less than an hour on a couple A100 GPUs. This allowed us to quickly iterate on our instruction fine-tuning stage, without pretraining from scratch. Comparatively, training a VLLM to create high-quality embeddings is much more resource intensive. Our pretraining run barely fits into the 640 GB of VRAM provided by a single A100 node, and takes several days to complete. Furthermore, our work indicates that these models could benefit substantially from further scaling (Appendix C).

5.3 Important Factors during Pretraining

Prior work has largely focused on what architectural adaptations to the make the VLLM to convert it into an embedding model. These include the attention mask, how the embedding is pooled and adapter architecture (BehnamGhader et al., 2024; Lee et al., 2024; Jiang et al., 2024b). Throughout our experiments, we find that most of these choices are often interchangeable or only produce marginal improvements (Appendix B). However, We find that many of the crucial factors when training CLIP models are also important when adapting VLMs. In particular, well-chosen data, batch size and number of samples seen during training are all important factors (Appendix C), just like with CLIP models (Gadre et al., 2023; Cherti et al., 2023a).

6 Conclusion

We introduce **ABC**, a multimodal embedding model that leverages a VLLM backbone to control image representations via natural language instructions. It achieves strong zero-shot results on a variety of multimodal tasks, spanning retrieval, classification, and VQA. **ABC** consists of a multi-stage training process, which isolates the computationally expensive contrastive pretraining from a lightweight instruction finetuning phase. We explore what factors are the most crucial when adapting VLMs to output multimodal embeddings. In particular, we find that vision encoder resolution, contrastive loss temperature and VLLM backbone choice are all important factors. Lastly, we design `CtrlBench` to measure our model's ability to use instructions to accomplish subtle natural language guided retrieval tasks.

References

Apple Inc. iphone 16 – tech specs. https://support.apple.com/en-ca/121029, 2024. Accessed: 2025-01-22.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc,

E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/97af07a14cacba681feacf3012730892-Paper.pdf.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. Llm2vec: Large language models are secretly powerful text encoders, 2024. URL https://arxiv.org/abs/2404.05961.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. *CoRR*, abs/2102.08981, 2021. URL https://arxiv.org/abs/2102.08981.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, 2024. URL https://arxiv.org/abs/2312.14238.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2023a. doi: 10.1109/cvpr52729.2023.00276. URL http://dx.doi.org/10.1109/CVPR52729.2023.00276.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829. IEEE, June 2023b. doi: 10.1109/cvpr52729.2023.00276. URL http://dx.doi.org/10.1109/CVPR52729.2023.00276.

Gabriel de Souza P. Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. Nv-retriever: Improving text embedding models with effective hard-negative mining, 2024. URL https://arxiv.org/abs/2407.15831.

Josip Djolonga, Jessica Yung, Michael Tschannen, Rob Romijnders, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Matthias Minderer, Alexander D'Amour, Dan Moldovan, Sylvain Gelly, Neil Houlsby, Xiaohua Zhai, and Mario Lucic. On robustness and transferability of convolutional neural networks. *CoRR*, abs/2007.08558, 2020. URL https://arxiv.org/abs/2007.08558.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. URL "http://host.robots.ox.ac.uk/pascal/VOC/voc2007".

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets, 2023. URL https://arxiv.org/abs/2304.14108.

Luyu Gao, Yunyi Zhang, Jiawei Han, and Jamie Callan. Scaling deep contrastive learning batch size under memory limited setup, 2021. URL https://arxiv.org/abs/2101.06983.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. The false promise of imitating proprietary llms, 2023. URL https://arxiv.org/abs/2305.15717.

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, 2018. URL https://arxiv.org/abs/1802.08218.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization, 2021. URL https://arxiv.org/abs/2006.16241.

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, 2021. URL https://arxiv.org/abs/2106.09685.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering, 2019. URL https://arxiv.org/abs/1902.09506.

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 Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/jia21b.html.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

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, 2024a. URL https://arxiv.org/abs/2407.12580.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks, 2024b. URL https://arxiv.org/abs/2410.05160.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Casey A. Fitzpatrick, Peter Bull, Greg Lipstein, Tony Nelli, Ron Zhu, Niklas Muennighoff, Riza Velioglu, Jewgeni Rose, Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, Helen Yannakoudakis, Vlad Sandulescu, Umut Ozertem, Patrick Pantel, Lucia Specia, and Devi Parikh. The hateful memes challenge: Competition report. In Hugo Jair Escalante and Katja Hofmann (eds.), *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pp. 344–360. PMLR, 06–12 Dec 2021. URL https://proceedings.mlr.press/v133/kiela21a.html.

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/5d44ee6f2c3f71b73125876103c8f6c4-Paper.pdf.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016. URL https://arxiv.org/abs/1602.07332.

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, 2024. URL https://arxiv.org/abs/2405.17428.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. URL https://arxiv.org/abs/2301.12597.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024a. URL https://arxiv.org/abs/2310.03744.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024b. URL https://arxiv.org/abs/2307.06281.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Pan Lu, Swaroop Mishra, Tony 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. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Alejandro López-Cifuentes, Marcos Escudero-Viñolo, Jesús Bescós, and Álvaro García-Martín. Semantic-aware scene recognition. *Pattern Recognition*, 102:107256, 2020. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2020.107256. URL https://www.sciencedirect.com/science/article/pii/S0031320320300613.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge, 2019. URL https://arxiv.org/abs/1906.00067.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning, 2022. URL https://arxiv.org/abs/2203.10244.

Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V Jawahar. Infographicvqa, 2021a. URL https://arxiv.org/abs/2104.12756.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images, 2021b. URL https://arxiv.org/abs/2007.00398.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark, 2023. URL https://arxiv.org/abs/2210.07316.

OpenAI. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, 2016. URL https://arxiv.org/abs/1505.04870.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5835–5847, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.466. URL https://aclanthology.org/2021.naacl-main.466/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2015. URL https://arxiv.org/abs/1409.0575.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. URL https://arxiv.org/abs/2210.08402.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge, 2022. URL https://arxiv.org/abs/2206.01718.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics*, 2018. URL https://api.semanticscholar.org/CorpusID:51876975.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read, 2019. URL https://arxiv.org/abs/1904.08920.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model, 2022. URL https://arxiv.org/abs/2112.04482.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale, 2023. URL https://arxiv.org/abs/2303.15389.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor,

Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Roopali Vij, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek

Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka

Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models, 2025. URL `https://arxiv.org/abs/2312.11805`.

Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss, 2021. URL `https://arxiv.org/abs/2012.09740`.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models, 2024a. URL `https://arxiv.org/abs/2401.00368`.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024b. URL `https://arxiv.org/abs/2409.12191`.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. Uniir: Training and benchmarking universal multimodal information retrievers, 2023. URL `https://arxiv.org/abs/2311.17136`.

Erik Wijmans, Brody Huval, Alexander Hertzberg, Vladlen Koltun, and Philipp Krähenbühl. Cut your losses in large-vocabulary language models, 2024. URL `https://arxiv.org/abs/2411.09009`.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010. doi: 10.1109/CVPR.2010.5539970.

Rui Xiao, Sanghwan Kim, Mariana-Iuliana Georgescu, Zeynep Akata, and Stephan Alaniz. Flair: Vlm with fine-grained language-informed image representations. *CVPR*, 2025.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training, 2021. URL `https://arxiv.org/abs/2111.07783`.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models, 2022. URL `https://arxiv.org/abs/2205.01917`.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2024. URL `https://arxiv.org/abs/2311.16502`.

Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. Magiclens: Self-supervised image retrieval with open-ended instructions. In *The Forty-first International Conference on Machine Learning (ICML)*, pp. to appear, 2024.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Gme: Improving universal multimodal retrieval by multimodal llms, 2025a. URL https://arxiv.org/abs/2412.16855.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Gme: Improving universal multimodal retrieval by multimodal llms, 2025b. URL https://arxiv.org/abs/2412.16855.

Kecheng Zheng, Yifei Zhang, Wei Wu, Fan Lu, Shuailei Ma, Xin Jin, Wei Chen, and Yujun Shen. Dreamlip: Language-image pre-training with long captions. In *ECCV*, 2024.

Junjie Zhou, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, Defu Lian, and Yongping Xiong. Megapairs: Massive data synthesis for universal multimodal retrieval, 2024. URL https://arxiv.org/abs/2412.14475.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images, 2016. URL https://arxiv.org/abs/1511.03416.

A Instruction Prompt Design

We use the following prompt template when using GPT-4o to generate instructions:

"<Image> Given this image, provide a user prompt where the following caption would be a reasonable answer: {Caption}. Only return the prompt."

Where {Caption} is the caption for the bounding box from Visual Genome. We use a temperature of 0 (deterministic) when generating instructions. We choose this template to collect a diverse set of instructions that a user would plausibly ask.

B Architecture Ablation

LLM2Vec (BehnamGhader et al., 2024) introduced several architecture modifications for adapting LLMs into text embedding models. In the multimodal setting, we find that using casual vs. bidirectional attention is marginal for improving model accuracy (Figure 7).

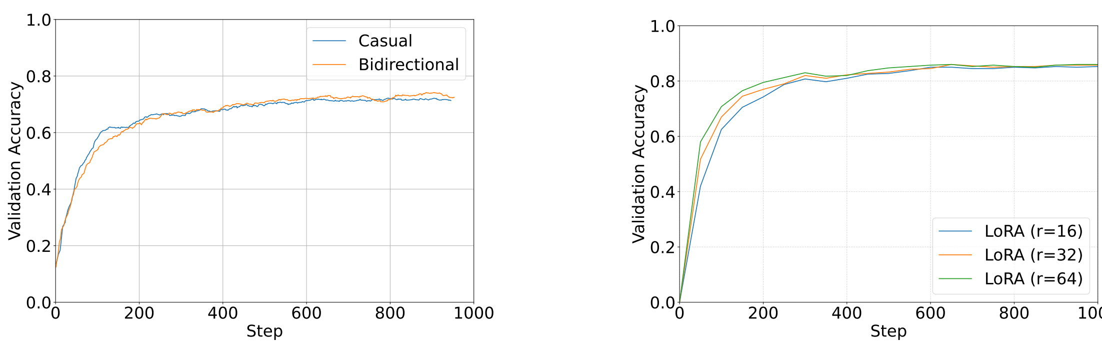
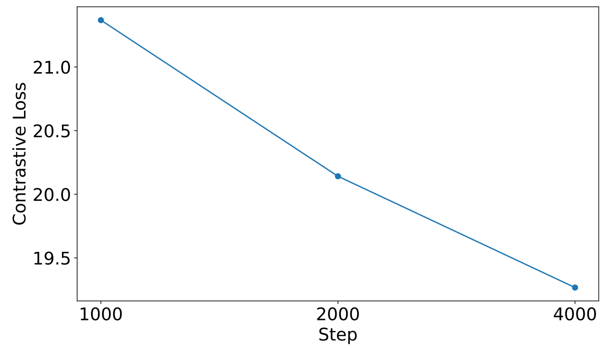


Figure 7: **(Left)** Causal vs Bidirectional attention. **(Right)** LoRA adapter rank ablation.

On average, bidirectional attention *slightly* outperforms casual attention. Ablating over adapter rank, we find that different rank adapters tend to converge to the same accuracy, with higher rank adapters performing a fraction of a percentage better. Overall, better data and scaling (Appendix C) present a much more promising direction for improving embedding quality.

C Scaling Training



We find that batch size and number of samples seen during training are both important factors. By increasing our batch size by 4x, from 128 queries and 1024 candidates to our full pretraining run size (512 queries and 4096 candidates), our validation accuracy on an 800 sample validation batch using in-batch negatives increases from 92.5% to 95.0%. We control for total samples seen by training the smaller batch size model for 4x more steps. Interestingly, this effect is even more pronounced on OOD validation data. For example, on a similar batch of MSCOCO data, accuracy increases from 73.8% to 84.0%. Therefore, techniques for scaling batch size under limited VRAM, such as GradCache Gao et al. (2021), are a promising direction to improve our pretrained model. Furthermore, we find that more steps (more samples seen) is also a straightforward way to increase the pretrained model's performance. As shown by Appendix C, doubling step count steadily decreases loss during our pretraining run.

D Comparison to VLM2Vec

With the MMEB benchmark, Jiang et al. (2024b) also include a mutltimodal embedding model: VLM2Vec. We exclude VLM2Vec from Table 1 and **??** as it has been trained on MSCOCO and several tasks from the MMEB benchmark. We provide the comparison of the two models here while specifying which tasks are out of distribution (OOD) for VLM2Vec. All tasks are OOD for our model.

	VLM2Vec	**ABC (ours)**
ImageNet-1K (Russakovsky et al., 2015)	65.6	**71.2**
HatefulMemes Kiela et al. (2021)	**67.1**	52.1
VOC2007 Everingham et al.	**88.6**	81.4
SUN397 Xiao et al. (2010)	**72.7**	71.8
Place365 López-Cifuentes et al. (2020)	**42.6**	40.7
ImageNet-A (Djolonga et al., 2020)	19.3	**49.4**
ImageNet-R (Hendrycks et al., 2021)	70.2	**86.8**
ObjectNet Barbu et al. (2019)	29.5	**67.7**
Country-211 Radford et al. (2021)	13.0	**18.5**
Average	52.1	**60.0**
Average OOD	34.9	**52.6**

Table 6: Classification results for **VLM2Vec** and **ABC (ours)**. Gray background indicates that **VLM2Vec** has been trained on this task.

Table 6 compares the performance of our model to VLM2Vec Jiang et al. (2024b). Gray rows indicate tasks that **VLM2Vec** is trained on. On average, VLM2Vec performs better on tasks that are in its distribution, while our model outperforms on tasks where both models are OOD. Interestingly, our model performs better on ImageNet-1K even though it is not trained on it.

E `CtrlBench`

Q: What is depicted in the image?
A: This is a signage

Q: What can be seen growing near the base of the signpost?
A: long limp green blades of plant

Q: What does the bottom sign say?
A: a sign that says plums

Q: What is written on the top sign in this image?
A: The word Apples written in white on a black background

Q: What does the middle sign in the image say?
A: A dark grey sign that says Pears.

Q: What is featured prominently in the foreground of the image, facing the ocean?
A: lone wood bench

Q: What do you notice about the water in this coastal scene?
A: pretty blue ocean water

Q: What is the structure visible on the horizon in this coastal landscape?
A: a distant, stone lighthouse

Q: What landmark is visible in the background by the ocean?
A: old European lighthouse

Q: What do you notice about the weather in this seaside photo?
A: blue sky in the distance

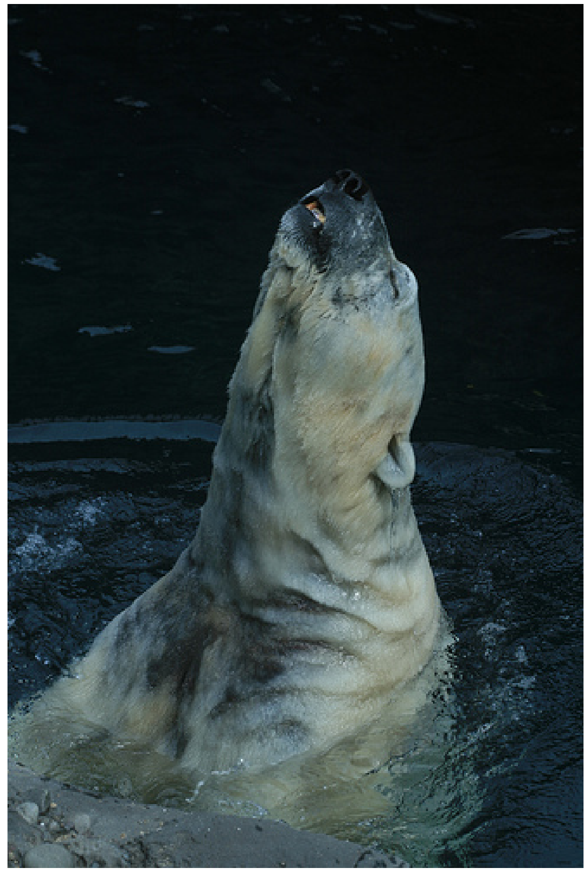
Q: What is happening in the top left corner of the image?
A: Tennis ball in the air

Q: What is the player about to hit with his racket?
A: A round tennis ball

Q: What color are the shorts the tennis player is wearing?
A: black and yellow shorts

Q: What action is taking place on the tennis court in the image?
A: A man serving a tennis ball.

Q: What's prominently dividing the ground area in the image?
A: White line on a court



Q: What part of the animal is prominently displayed in the image?
A: the neck of a polar bear

Q: What colors can you see on the bear in the image?
A: black, brown and white bear neck

Q: What stands out to you about the bear's appearance in this image?
A: the bear has such tiny ears for such a huge animal

Q: What is the bear doing with its eyes in this picture?
A: the bear has his eyes closed

Q: What's the polar bear swimming in?
A: a body of water