Measuring Progress in Fine-grained Vision-and-Language Understanding

Emanuele Bugliarello[*,ð,c] **Laurent Sartran**[ð] **Aishwarya Agrawal**[ð]
Lisa Anne Hendricks[‡,ð] **Aida Nematzadeh**[‡,ð]

[ð]DeepMind [c]University of Copenhagen

Abstract

While pretraining on large-scale image–text data from the Web has facilitated rapid progress on many vision-and-language (V&L) tasks, recent work has demonstrated that pretrained models lack "fine-grained" understanding, such as the ability to recognise relationships, verbs, and numbers in images. This has resulted in an increased interest in the community to either develop new benchmarks or models for such capabilities. To better understand and quantify progress in this direction, we investigate four competitive V&L models on four fine-grained benchmarks. Through our analysis, we find that X-VLM (Zeng et al., 2022) consistently outperforms other baselines, and that modelling innovations can impact performance more than scaling Web data, which even degrades performance sometimes. Through a deeper investigation of X-VLM, we highlight the importance of both novel losses and rich data sources for learning fine-grained skills. Finally, we inspect training dynamics, and discover that for some tasks, performance peaks early in training or significantly fluctuates, never converging.

1 Introduction

Fine-grained multimodal skills (*e.g.*, understanding relationships and recognising verbs) require identifying and relating various entities across both image and text modalities. Vision-and-language models (VLMs) need such skills to robustly perform well on real-world vision-and-language (V&L) applications; *e.g.*, a *coarse-grained* model tested on image retrieval to "find an image where something is *on* a sofa" might incorrectly return an image of a cat sitting *below* the sofa. As another example, in captioning, a model might incorrectly describe an image where "someone is *selling* a sweater" as "someone is *buying* a sweater," if it does not have a precise understanding of the two verbs.

However, common V&L benchmarks (*e.g.*, Lin et al., 2014; Goyal et al., 2017; Suhr et al., 2019) do not explicitly shed light on such fine-grained understanding. Indeed, in the last few years, there has been an increase in the number of benchmarks which demonstrate that current, coarse-grained models struggle with fine-grained understanding (Hendricks and Nematzadeh, 2021; Parcalabescu et al., 2022; Salin et al., 2022; Thrush et al., 2022). Meanwhile, more models have been designed specifically to learn a better mapping between visual and textual modalities (*e.g.*, Yao et al., 2022a,b; Zeng et al., 2022; Gao et al., 2022). While such models perform well on coarse-grained retrieval and other downstream tasks, they have not been directly evaluated on fine-grained understanding. Consequently, it is unclear if the performance gains are due to tighter, more fine-grained representations introduced by model innovations at the pretraining stage. To fill this gap, we analyse several recent models with innovations designed for a better image–text alignment and their corresponding baselines on a suite of fine-grained benchmarks. We centre our study on three key questions.

First we consider: *Which models perform well on fine-grained tasks?* To answer this, we evaluate models from four different model families trained with different amounts of pretraining data, as well as recent architectures that leverage frozen large language models (LLMs). We observe that **modelling innovations have more impact than simply scaling image captions** from the Web. Furthermore, explicitly modelling localisation can improve performance, but it is crucial *how* it is done, and simply using localisation data is not enough.

Our observations motivate our next question: *How do data and losses impact fine-grained understanding?* We focus our study on the best performing model, X-VLM (Zeng et al., 2022), which learns to map specific objects and regions (not a full image) to a label (word or phrase describing the

*Work completed during an internship at DeepMind. ‡denotes equal senior contribution. Correspondence to: Emanuele Bugliarello <emanuele@di.ku.dk>.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers, pages 1559–1582
July 9-14, 2023

Benchmark	Task	Examples	Subtasks	Example Subtasks
		Fine-grained Tasks		
SVO-Probes	Verb understanding	48K	3	subject, verb, object
VALSE	V&L grounding	14K	6	existence, counting, spatial relations
VSR	Spatial reasoning	2K	7	adjacency, directional, proximity relationships
Winoground	Compositional reasoning	800	8	pragmatics, object swap, relation swap
		Coarse-grained Tasks		
COCO	Retrieval	25K	0	N/A
Flickr30K	Retrieval	5K	0	N/A

Table 1: Overview of our benchmarks. For consistency, we report the number of examples as the number of positive image–text pairs in each evaluation dataset.

region). We reformulate the X-VLM loss to better disentangle the contribution of data and losses, observing that more data does not improve performance unless paired with **losses designed to learn a mapping between regions and labels**. Furthermore, the diversity of class labels is important for performance on coarse-grained retrieval, and region descriptions (as opposed to single word labels) are crucial for performance on fine-grained tasks.

Finally, it is unclear if all fine-grained skills are learned at the same time during training so we consider: *How does fine-grained understanding evolve during training?* Surprisingly, we find that while performance steadily improves on coarse-grained retrieval tasks through training, **performance fluctuates substantially on many fine-grained tasks**, with some skills, like counting, becoming increasingly *worse*. Additionally, performance across different fine-grained tasks that should test for similar skills are not always well correlated.

Contributions. In this work, we **1)** provide in-depth analyses of how data and modelling decisions impact performance on fine-grained tasks, and **2)** further disentangle the gains given by data and pretraining losses on our best performing model (X-VLM). Our results suggest that to make progress in fine-grained understanding, modelling innovations (*e.g.*, through object-centric losses) as well as data quality and richness are more effective than scaling up Web data alone. Finally, we **3)** shed light on VLMs' pretraining dynamics and suggest that future work should revisit pretraining strategies in order to consistently improve across several tasks.

2 Benchmarks

We describe the recent (English) benchmarks proposed to measure fine-grained V&L understanding in zero-shot setups.[1] See Table 1 for an overview.

SVO-Probes (Hendricks and Nematzadeh, 2021) focuses on verb understanding: it tests whether a model can identify if an image matches a sentence, and includes negative images which differ on a specific part of speech (Subject, Verb, and Object). The dataset consists of 421 verbs and over 48K image–sentence pairs.[2] The authors show that their baselines fail more in situations requiring verb understanding than other parts of speech.

VALSE (Parcalabescu et al., 2022) consists of six tasks that cover basic linguistic phenomena, such as plurality, actions and coreference. For each task, given a visual input, a model is asked to distinguish real captions from foils (Shekhar et al., 2017), where a foil is constructed from a caption by altering a word or phrase that realises a specific linguistic phenomenon (*e.g.*, semantic number of nouns). The authors show that VLMs can identify objects in images, but struggle to ground their interdependence with specific linguistic indicators.

VSR (Liu et al., 2023) tests for 65 types of visual spatial relationships (*e.g.*, `under`, `in front of`) grouped into seven categories (*e.g.*, adjacency, orientation). Each sample consists of an image–sentence pair; a model needs to predict whether the sentence correctly describes the spatial relation between two objects in the image. We evaluate models in a zero-shot setup on the 'random' split.[3]

Winoground (Thrush et al., 2022) is an expert-curated benchmark aiming to test models' compositional reasoning. Given two images and two captions, the goal is to match them correctly; wherein both captions contain the same set of words, but in a different order. The authors define three scores: Text (whether a model can match the correct caption for a given image), Image (vice versa), and Group (whether a model can match each pair). Several competitive VLMs have been shown to often perform close to or below random chance.

We also report zero-shot performance on coarse-grained retrieval in **Flickr30K** (Young et al., 2014) and **COCO** (Lin et al., 2014) in our analysis.

[1] We note that two more datasets require fine-grained skills to be solved and that they are not part of our analysis. ImageCoDe (Krojer et al., 2022) requires comparing a caption within a multi-image context, a setup not suitable for zero-shot evaluation of current single-image VLMs. Yuksekgonul et al. (2023) propose the ARO benchmark to evaluate VLMs' attribution, relation, and order understanding. However, the data had not been released as of the ACL deadline.

[2] Only 30,578 pairs were available as of Nov 2022.

[3] Note that VSR has recently been updated, but we expect the findings from our experiments to hold on the revised splits.

Model	Loss			Data		Downstream		
	CL	Text	Obj Det	Unsupervised	Supervised	VQAv2	NLVR2	RefCOCO+
ALBEF$_{4M}$	✔	MLM	-	4M: COCO+SBU+VG+CC$_{3M}$	-	74.7	80.5	-
ALBEF$_{14M}$	✔	MLM	-	14M: 4M + CC$_{12M}$	-	76.0	83.1	-
BLIP$_{14M}$	✔	LM	-	CapFilt/B(14M)	-	77.6	82.3	-
BLIP$_{129M}$	✔	LM	-	CapFilt/B(14M + LAION)	-	78.2	83.1	-
BLIP$_{129M}$-CapFilt/L	✔	LM	-	CapFilt/L(14M + LAION)	-	78.3	82.2	-
BLIP-ViT/L$_{129M}$	✔	LM	-	CapFilt/L(14M + LAION)	-	-	-	-
PEVL$_{14M}$	✔	MLM	MLM	14M	RefCOCO{,+,g}+F30KE+GQA+VCR+VG	-	-	74.5
X-VLM$_{4M}$	✔	MLM	Regress	4M	COCO + VG	78.1	84.2	71.0
X-VLM$_{16M}$	✔	MLM	Regress	14M	COCO + VG + Objects365 + OpenImages	78.4	84.4	76.9

Table 2: Overview of core evaluated models. All the models use contrastive learning (CL), cross-attention and a (masked) language modelling objective. Fine-grained models also predict object locations from supervised data.

3 Evaluated Models

Recent work has shown that two components are crucial ingredients of strong coarse-grained VLMs (*e.g.*, Li et al., 2021; Alayrac et al., 2022; Chen et al., 2023): **1)** a contrastive objective that aligns vision and language modalities, and **2)** a cross-attention mechanism that fuses the two modalities. As we are interested in high performance on both fine- and coarse-grained tasks, to select models for our study, we surveyed recent work that uses these building blocks,[4] but also incorporates new losses or data that can potentially improve fine-grained V&L understanding. We find that many recent models build on ALBEF (Singh et al., 2022; Yang et al., 2022; Hao et al., 2023) (which we also study as a coarse-grained baseline).

Other than strong performance on coarse-grained and downstream tasks, we also considered: **1)** the possibility to study the role of new modelling innovations and data for fine-grained skills, and **2)** the availability of open-source code and pretrained weights. This resulted in four models briefly described next (more details in App. A.1). Table 2 codifies the main differences in pretraining objectives and data used by these models. Recall that previous work does not evaluate these models on fine-grained benchmarks.

ALBEF (Li et al., 2021), with strong downstream performance, matches all our criteria and serves as a coarse-grained baseline. ALBEF is a dual-stream encoder (Bugliarello et al., 2021) that first encodes images and captions independently, and then fuses them with cross-modal attention.

BLIP (Li et al., 2022b) uses an autoregressive language model (LM), and employs a dataset bootstrapping technique (CapFilt) to generate synthetic captions and to remove noisy pairs from large-scale Web data. BLIP outperforms ALBEF on most coarse-grained downstream tasks; thus, we study BLIP as another coarse-grained baseline to test if its generative LM and data contributions also lead to better fine-grained understanding.

PEVL (Yao et al., 2022b) is a fine-grained model building on ALBEF, but leverages more supervised datasets such as referring expressions, captions with visual coreferences, object detection and region descriptions data, etc. (see Table 2). Unlike ALBEF, PEVL is explicitly trained to learn fine-grained representations of entities by predicting their coordinates in a unified masked language modelling framework (similar to Pix2Seq, Chen et al., 2022): bounding box coordinates corresponding to a given entity are added in the caption as "`A cat < 10 73 206 175 > is napping.`"

X-VLM (Zeng et al., 2022) is our second fine-grained model that enhances ALBEF by adding both new losses and additional supervised data. In contrast to PEVL, X-VLM models visual position through an additional bounding box prediction head that regresses the object's bounding box (bbox) coordinates. The authors use both object detection labels and region descriptions to learn coarse- and fine-grained alignments (we provide an in-depth analysis of this model in Section 5).

We remark that PEVL and X-VLM were the only open-source fine-grained VLMs at the time of our evaluation, and both of them build on top of ALBEF. In addition to these core models, we also evaluate a dual-encoder network (CLIP; Radford et al. 2021) as well as recent architectures that rely on frozen, autoregressive (L)LMs: ClipCap (Mokady et al., 2021), Flamingo (Alayrac et al., 2022) and BLIP-2 (Li et al., 2023). As these models perform generally worse than our best fine-grained model, X-VLM, and differ significantly from it, we do not discuss their performance further. For more details, we refer the reader to Tables 6 to 11 in App. B.1.

[4]By studying models with well-established modules, we expect our findings to be more informative for future work.

	Model	SVO Avg.	VALSE Avg.	VSR Test Avg.	Winoground Text	Winoground Image	Winoground Group
	Random	50.0	50.0	50.0	25.0	25.0	12.5
	CLIP_{400M}	81.6	64.0	N/A	30.7	10.5	8.0
	BLIP-2_{129M}	86.5	74.0	61.5	43.0	22.0	18.2
1	ALBEF_{4M}	87.6	69.1	57.3	29.2	15.5	11.0
2	$\text{X-VLM}_{4M}{}^{\sharp}$	<u>88.9</u>	<u>72.4</u>	<u>63.0</u>	<u>44.0</u>	**26.7**	**21.5**
3	ALBEF_{14M}	88.6	69.4	58.3	32.5	16.2	12.7
4	BLIP_{14M}	48.7	67.8	49.7	36.5	18.5	14.5
5	$\text{PEVL}_{14M}{}^{\sharp}$	86.2	68.9	57.5	33.2	15.7	12.2
8	$\text{X-VLM}_{16M}{}^{\sharp}$	**90.0**	**74.5**	**64.3**	**46.7**	<u>24.5</u>	<u>21.2</u>
9	BLIP_{129M}	<u>51.4</u>	68.8	46.9	<u>35.5</u>	15.0	11.7
10	BLIP_{129M}-CapFilt/L	51.2	68.2	48.7	34.7	<u>15.2</u>	<u>12.2</u>
11	BLIP-ViT/L_{129M}	50.8	<u>70.3</u>	<u>50.3</u>	34.7	14.5	<u>12.2</u>

Table 3: Overall performance of core evaluated models on fine-grained benchmarks; the highest values for a given data size and the overall best values are marked with underline and bold, respectively. $\sharp$ marks fine-grained models. For a detailed breakdown of task performance and full comparison with prior arts, see App. B.1.

4 Which Fine-grained Models Perform Well on Fine-grained Tasks?

We compare two strong VLMs (ALBEF and BLIP) with two models with explicit object modelling (*i.e.*, fine-grained; X-VLM and PEVL). We evaluate on fine-grained tasks (see Table 3) to determine if recent object-centric models improve on tasks designed to measure fine-grained skills—an evaluation missing from previous work. We also include results on CLIP and BLIP-2 in Table 3 to highlight how well fine-grained models perform, even though pretrained with less data and having fewer parameters (as shown in Table 6 in App. B.1).

Experimental setup. All our fine-grained benchmarks only require models to predict a matching score for a given image–text pair, a common task that current V&L models—including all of our evaluated models—are pretrained to solve. On VSR, a model's prediction is correct if the matching score is greater/lower than 50% for a true/false label. On the other benchmarks, a model's prediction is correct if the score for the positive image–text pair is higher than the score of the negative pair(s).[5] We evaluate the public models released by the authors on GCP.[6] Code to reproduce our analysis is online.[7]

ALBEF vs. BLIP. We first compare our two coarse-grained baselines. A key difference between ALBEF and BLIP is that the former is trained with masked language modelling (MLM), while the latter uses autoregressive language modelling (LM) for text; with BLIP outperforming ALBEF on downstream tasks when pretrained on the same 14M images. Performing the same comparison on fine-grained benchmarks, we find that ALBEF_{14M} outperforms BLIP_{14M} on all tasks (largely on SVO-Probes and VSR) except on Winoground. Likewise, Table 6 (App. B.1) shows that other visual-conditional LMs, such as ClipCap models, also struggle with fine-grained understanding. This might be due to the fact that our evaluation relies on image–text alignments and does not test for generation, where the LM objective is often preferred. Given these results and the fact that ALBEF is more similar to our fine-grained models, we compare against ALBEF in most of our discussion.

Effectively modelling object positions improves fine-grained understanding. Overall, we find that X-VLM consistently outperforms all other evaluated approaches (see Table 3). This trend holds in both the 4M and 16M pretraining setups. When trained on the same 4M images as the ALBEF baseline, X-VLM with explicit object modelling, notably improves over all benchmarks (gaining 1.3pp on SVO-Probes, 3.3pp on VALSE, 5.7pp on VSR, and 14.8/11.2/11.5pp on Winoground). Importantly, X-VLM_{4M} also outperforms ALBEF_{14M} (trained on 10M more data points). This result shows the importance of explicit object modelling for a range of fine-grained tasks, including ones that are dissimilar to the supervised localisation task (*e.g.*, verb understanding).

X-VLM_{16M}, which adds CC_{12M} as well as object detection data from OpenImages and Objects365 to X-VLM_{4M}'s data, achieves even higher overall gains in most fine-grained benchmarks. On VALSE, it closes the gap with a larger model trained on supervised data from many downstream tasks (12-in-1; Lu et al. 2020), and on VSR it achieves similar accuracy to LXMERT (Tan and Bansal, 2019) fine-tuned on 50% of VSR training data (67.9pp). Moreover, on Winoground, X-VLM_{4M} significantly outperforms previous coarse-grained models, including a large-scale dual-encoder (CLIP, Group score of 8.0; Radford et al., 2021) and a strong, larger cross-modal Transformer ($\text{UNITER}_{\text{Large}}$, Group score of 10.5; Chen et al., 2020), as shown in Table 6 in App. B.1.

Not all object modelling improves fine-grained understanding. Like X-VLM, PEVL also mod-

[5] We evaluate SVO-Probes using *pairwise ranking accuracy* to benchmark models without a binary classification head (we note that Hendricks and Nematzadeh 2021 used accuracy).

[6] https://cloud.google.com/.

[7] https://github.com/e-bug/fine-grained-evals.

els visual locations of objects. However, it does so by expecting (masked) bbox locations as part of its input caption. Surprisingly, PEVL_{14M} performs much worse than X-VLM_{16M} on all tasks; in fact, it performs on par with the ALBEF_{14M} baseline, despite being originally initialised with its checkpoint and further tuned to model visual object locations.[8] We conjecture that modelling objects as input prompts is less beneficial than directly predicting object locations with a classification head (X-VLM), as the former does not directly influence the object's representations in the text modality.

Modelling objects has more impact than increasing data. In Table 3, we observe that, not surprisingly, increasing data for a given family (*e.g.*, ALBEF_{4M} to ALBEF_{14M}) results in improved performance on most benchmarks. However, interestingly, the *fine-grained* X-VLM_{4M}, trained on 4M data points, outperforms all BLIP_{129M} variants—a coarse-grained model trained on 129M data points (compare row 2 with rows 9–11). Similarly, while increasing the data from 4M to 14M results in improvements across most tasks for the coarse-grained ALBEF_{14M}, these performance gaps are smaller than what we gain from modelling objects on top of ALBEF_{4M}. That is, the average performance gap between ALBEF_{4M} and X-VLM_{4M} is bigger (+5.2pp) than that observed when increasing data from ALBEF_{4M} to ALBEF_{14M} (+1.0pp). This result highlights that simply scaling data, without modelling innovations, might not be enough for notable improvements on fine-grained tasks.

We also find that scaling data can *hurt* performance on some benchmarks. For example, on Winoground Image and Group scores, X-VLM_{16M} and BLIP-ViT/L_{129M} perform worse than their corresponding models trained on less data, X-VLM_{4M} and BLIP_{14M}, respectively.[9] Looking at performance by subtasks, we find that scaling Web data leads to worse performance on several of them, such as Image scores in most Winoground tasks, and VALSE's `existence`, `counting adversarial` and `coreference` for BLIP-ViT/L_{129M} (more details in App. B.1). We conjecture that pretraining on noisy Web data—where the language in an image–text pair does not always faithfully describe the image—might diminish the fine-grained alignments learned from smaller, cleaner datasets (Hendricks et al. 2021 report similar trends on coarse-grained tasks).

Takeaways. We observe that modelling object positions in images provides a strong signal for fine-grained understanding; but, *how* we model this information is crucial: simply pretraining a model with bbox positions in input does not lead to better off-the-shelf representations. We also see bigger gains on fine-grained tasks when modelling objects compared to scaling the pretraining data.

5 Data & Losses for Fine-grained Tasks

Recent fine-grained models build on coarse-grained ones by introducing additional training data (*e.g.*, object detection data in X-VLM and PEVL) and new losses (*e.g.*, bounding box regression loss in X-VLM). We study how data and losses influence fine-grained understanding, focusing on X-VLM as it outperforms other models on fine-grained benchmarks. While Zeng et al. (2022) perform ablations to show the importance of their new objective function, they do not study the impact of data and losses independently; moreover, they do not evaluate on find-grained benchmarks. We start with a description of X-VLM, emphasising details in its pretraining procedure, that we reveal to have significant impact on the final performance.

5.1 What are X-VLM Data and Losses?

The X-VLM architecture consists of the same modules as ALBEF: a vision, a text, and a cross-modal Transformer (Vaswani et al., 2017) encoder (see App. A.1 for details). Given an image–text pair, ALBEF performs two forward passes (as shown in Figure 1): first, the model computes a contrastive learning loss ($\mathcal{L}_{\text{CL}}$) and an image–text matching loss ($\mathcal{L}_{\text{ITM}}$). In a second pass, it masks text inputs to compute a visually-grounded masked language modelling loss, $\mathcal{L}_{\text{MLM}}$. After the two forward passes, ALBEF is trained with $\mathcal{L}_{\text{A}} = \mathcal{L}_{\text{CL}} + \mathcal{L}_{\text{ITM}} + \mathcal{L}_{\text{MLM}}$.

Data. While ALBEF is only pretrained on image–caption data, X-VLM additionally pretrains on object and region detection data. Object detection data consists of an object or attribute–object label (*e.g.*, "dog" or "brown dog"), an image, and

[8] We evaluate three different models released by the authors, which differ in their pretraining and fine-tuning data. All the variants perform similarly, and as a result, we only report PEVL_{14M}, which underwent a second-stage pretraining on multiple supervised tasks (App. B.1 lists all the models).

[9] While BLIP_{129M} performs worse than BLIP_{14M} on a few benchmarks, this might be because the data size is significantly increased without scaling the model size. Thus, we compare against BLIP-ViT/L_{129M}, which uses a larger image encoder.

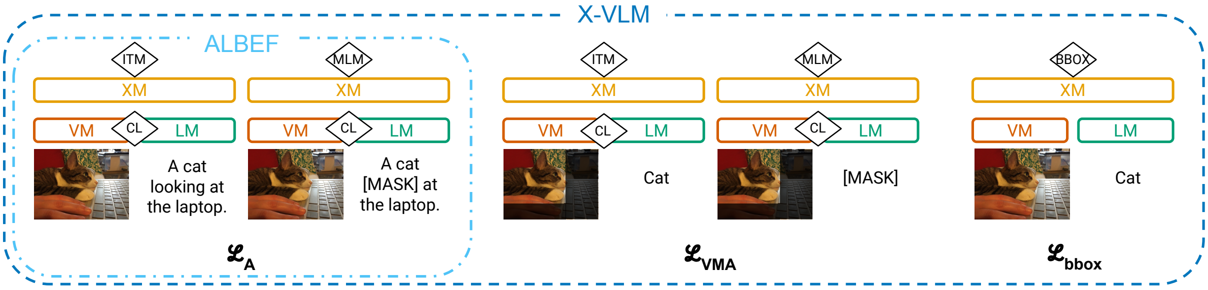

Figure 1: Overview of the pretraining objectives effectively used by ALBEF and X-VLM.

a bounding box; region detection data consists of a short phrase (*e.g.*, "a cute brown dog"), an image, and a bounding box. Other multimodal Transformer models have used detection data (Hendricks et al., 2021; Li et al., 2020; Bugliarello et al., 2021; Zhang et al., 2021), but usually the bounding boxes are discarded, and objects or region descriptions are paired with the *entire* image. In contrast, a close examination of the X-VLM codebase[10] reveals that X-VLM effectively makes use of bounding boxes.

BBOX loss. To take advantage of additional bounding box (bbox) data, X-VLM introduces an objective, $\mathcal{L}_{\text{bbox}}$, which learns to regress to object locations from object detection and region description data (see Figure 1 for an overview).

VMA loss. The X-VLM paper presents two losses, $\mathcal{L}_{\text{A}}$ and $\mathcal{L}_{\text{bbox}}$. However, $\mathcal{L}_{\text{A}}$ operates over two input types: image–text pairs from captioning data and image–text–bbox triplets from object detection data. Thus, it is hard disentangle the impact of the data and the losses on performance. We reformulate $\mathcal{L}_{\text{A}}$ into two losses,[11] operating over: (a) image–text pairs, $\mathcal{L}_{\text{A}}$, as in ALBEF; or (b) image–text–bbox pairs, that we denote *visually masked* ALBEF loss, $\mathcal{L}_{\text{VMA}}$. For $\mathcal{L}_{\text{VMA}}$, the visual and cross-modal encoders only attend to the image patches that correspond to the object bbox coordinates via an attention mask (see Figure 1). This results in an object-centric visual view for grounding the text label through the pretraining objectives. To compute this loss, in addition to the three forward passes described so far (CL and ITM, MLM, and BBOX losses), X-VLM performs two more passes: one where image patches outside a bounding box region are masked out to compute the *visually masked* CL and ITM loss, and another where text is additionally masked for the *visually masked* MLM loss. Section 5.3 shows both the data and pretraining techniques are key to the final model performance.

5.2 Experimental Setup

We re-implement ALBEF and X-VLM in JAX to ensure full control of modelling, data, and initialisation decisions.[12] We initialise both models with a 224×224 ViT-B/16 visual encoder (Steiner et al., 2022), and $\text{BERT}_{\text{BASE}}$ (Devlin et al., 2019) weights in the text and cross-modal layers. Similar to Bugliarello et al. (2021), we pretrain our models on the *exact same* 4M and 14M datasets used by the authors (Table 2), but note that only 1.8M and 11.2M data points were available for $\text{CC}_{3\text{M}}$ and $\text{CC}_{12\text{M}}$, respectively. For object detection data, we use the COCO and VG annotations released by the X-VLM authors. Following Zeng et al. (2022), we pretrain our models for 200K steps using the official hyperparameters (see App. A for more details).

5.3 Results

Table 4 shows the overall zero-shot performance of our ablations on three fine-grained benchmarks and two coarse-grained retrieval tasks. Row 0 is our ALBEF re-implementation, while row 10 corresponds to our X-VLM pretrained following the implementation of Zeng et al. (2022). Our controlled study allows us to quantify how each technique (losses, data, implementation details) in X-VLM contributes towards fine-grained understanding.

Data ablation. We first investigate the role of supervised detection data used to learn fine-grained relationships in X-VLM by pretraining the model, using its standard training objectives, and adding different data sources (rows 1–6).

Looking at rows 1–3, we find that region descriptions from VG (VG_{RD}) are the most useful,

[10]https://github.com/zengyan-97/X-VLM.

[11]Our reformulation is equivalent to X-VLM, but it allows us to disentangle the impact of data and losses on performance.

[12]To verify our implementation, we compare an ALBEF model trained in our codebase with one trained in the original codebase, obtaining an absolute difference below 1pp in Recall@1 on zero-shot Flickr30K and COCO retrieval tasks.

	Data				Loss			SVO-Probes	VALSE	VSR Random	Flickr30K		COCO	
	$\mathcal{D}_A$	$COCO_{OD}$	VG_{OD}	VG_{RD}	$\mathcal{L}_A$	$\mathcal{L}_{VMA}$	$\mathcal{L}_{bbox}$	Avg.	Avg.	Test Avg.	TR@1	IR@1	TR@1	IR@1
0	✓				✓			85.9	68.7	59.3	76.3	59.8	60.9	45.7
1	✓	✓			(✓)	(✓)	(✓)	85.9	69.1	58.6	72.8	59.5	60.8	46.1
2	✓		✓		(✓)	(✓)	(✓)	86.0	68.6	59.7	77.1	62.7	63.3	47.5
3	✓			✓	(✓)	(✓)	(✓)	86.6	70.3	61.1	79.4	62.3	**64.8**	**49.1**
4	✓	✓	✓		(✓)	(✓)	(✓)	85.6	67.5	60.7	77.2	60.7	63.3	47.3
5	✓	✓		✓	(✓)	(✓)	(✓)	86.5	67.6	60.1	77.2	61.4	62.9	47.6
6	✓		✓	✓	(✓)	(✓)	(✓)	**86.9**	**71.1**	**62.5**	**79.7**	**63.4**	64.4	**49.1**
7	(✓)	(✓)	(✓)	(✓)	✓			85.9	69.3	58.2	75.5	58.9	61.9	45.8
8	(✓)	(✓)	(✓)	(✓)	✓	✓		86.5	69.1	59.0	77.5	62.3	63.0	47.6
9	(✓)	(✓)	(✓)	(✓)	✓		✓	86.0	67.9	60.5	78.0	60.5	62.1	47.6
10	✓	✓	✓	✓	✓	✓	✓	**86.9**	69.8	61.9	78.3	63.0	64.6	48.6

Table 4: Overall performance of X-VLM ablations pretrained on the *exact same* data. Rows 0 and 10 are our re-implementation of ALBEF and X-VLM, respectively. Rows 3, 4, and 8 correspond to "w/o object," "w/o region," and "w/o bbox" ablations in Zeng et al. (2022). We find that $\mathcal{L}_{VMA}$ is crucial towards X-VLM's performance, and that VG_{RD} yields richer signal for both coarse- and fine-grained tasks.

single-source signal for the model, resulting in improvements in both fine- and coarse-grained tasks. This variant is either close to or surpasses the final X-VLM variant (row 10) in all the tasks. We attribute this success to both its size (3.7M data points) and language format, wherein noun phrases, rather than simple labels, describe a given entity. In addition, object detection data from VG (VG_{OD}) leads to similar fine-grained results as $COCO_{OD}$, but significantly better zero-shot retrieval performance. VG_{OD} is not only larger than $COCO_{OD}$, but also includes a more diverse set of classes.[13]

We hypothesise that a *large number of classes* (as in VG_{OD}) is important for coarse-grained retrieval tasks, and *more descriptive phrases* of VG_{RD} (rather than single labels) significantly impact fine-grained tasks. To verify this, we disentangle the effect of data size and type: specifically, we re-train rows 2–3 on a subset of VG with the same number of images and annotations as in $COCO_{OD}$. Figure 2 confirms our hypothesis: even when controlled for size, VG_{RD} leads to notably better performance than $COCO_{OD}$. On coarse-grained datasets, VG_{OD} largely outperforms $COCO_{OD}$.

Looking at multi-source supervised data (rows 4–6), our best performing model combines VG_{OD} and VG_{RD} data (row 6) and, surprisingly, adding $COCO_{OD}$ does not boost performance.

Loss ablation. We investigate the role of the two objectives used during supervised pretraining of X-VLM (rows 7–9). We see that training an ALBEF model on object detection data as-is (row 7) results in similar performance as pretraining it on standard

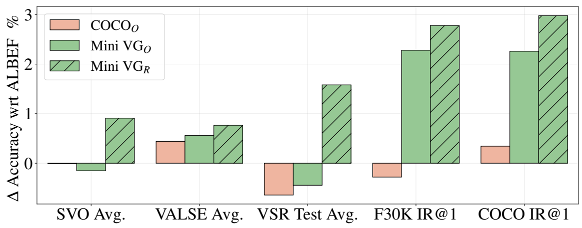

Figure 2: Performance on our benchmarks of X-VLM wrt ALBEF in a controlled setup with a single supervised dataset having the same number of images and annotations. Region descriptions give the highest gains.

image–caption data. That is, *just adding more data is not enough*; additional supervision in the form of the X-VLM pretraining objectives is crucial.

Compared to $\mathcal{L}_{bbox}$ (row 9), our reformulation makes it clear that $\mathcal{L}_{VMA}$ (row 8) leads, on average, to both higher fine-grained accuracy and higher recall on retrieval tasks. One potential explanation is that the visually masked forward pass directly influences the representation learned by the contrastive loss, as well as the cross-modal representations. In contrast, the regression loss only occurs after cross-modal interaction, suggesting that better alignment is important in both contrastive and cross-modal features. Finally, X-VLM achieves its best performance when combining $\mathcal{L}_{VMA}$ and $\mathcal{L}_{bbox}$.

Takeaways. Our reformulation of X-VLM allows us to conduct a careful analysis in a controlled setup on how data and losses influence X-VLM performance. We show that more data does not improve performance unless paired with additional supervisory signal, in the form of either the visually masked ALBEF loss or bbox regression. Given our observations and the fact that, as seen in Section 4 and App. B.1, X-VLM largely outperforms

[13] $COCO_{OD}$ and VG_{OD} have 80 and 50k labels respectively.

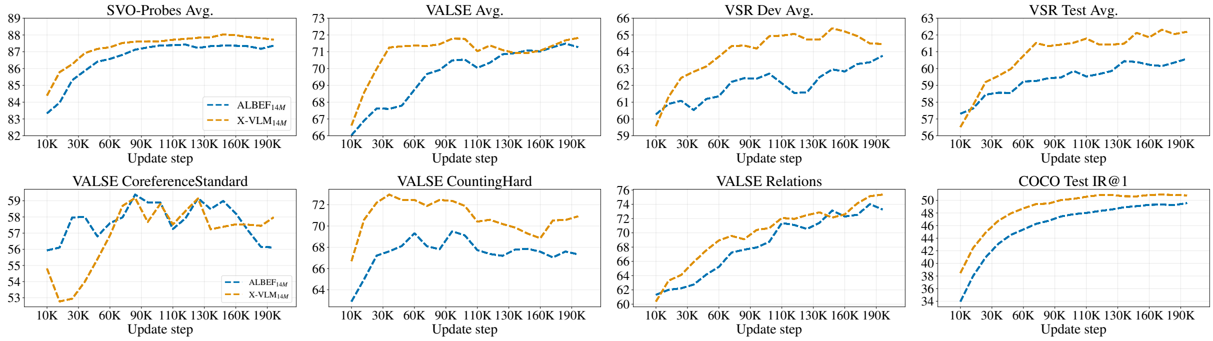

Figure 3: Training dynamics of our ALBEF$_{14M}$ and X-VLM$_{14M}$ models. Models' overall performance converges at different rates on different fine-grained benchmarks (top row). Performance on specific skills varies drastically, with some skills that deteriorate after an initial peak (bottom row). Retrieval performance (bottom right) does not capture this diversity in dynamics. Values smoothed with 0.6 factor for better visualisation. Full results in App. B.2.

the large-scale CLIP and BLIP-2 models on fine-grained tasks such as VALSE and Winoground, we believe that a promising direction in fine-grained understanding will require careful model and loss design with rich data sources like VG, not just scaling up with (potentially) noisy data.

6 Dynamics of Fine-grained Tasks

We now analyse the dynamics of fine-grained skills for our models to investigate (i) when and whether they are acquired, and (ii) how they relate to one another, especially when they aim at measuring similar capabilities. For example, does action understanding in VALSE correlate with verb understanding in SVO-Probes? Are there skills that vastly differ from each other that they would require different modelling contributions (*e.g.*, counting)?

Experimental setup. We evaluate checkpoints (every 10K steps) from pretraining our ALBEF and X-VLM re-implementations with 4M and 14M data points. We focus on 14M results as we see similar patterns with 4M (see App. B.2). When evaluating correlation patterns, we report both Pearson and Spearman correlation coefficients.

Different skills, different patterns. Figure 3 (top) shows how the average model performance evolves during pretraining for the four benchmarks. Interestingly, the performance on these benchmarks converges at different rates: both ALBEF and X-VLM models easily improve on SVO-Probes. Moreover, we observe that modelling objects (à la X-VLM) leads not only to better fine-grained understanding after 200K steps (Tables 3 and 4), but also to remarkably quicker learning rates.

Figure 3 (bottom) shows performance on indicative VALSE tasks, as well as on coarse-grained image retrieval on COCO. While some skills, such as `spatial relations` understanding, are learned progressively during pretraining, others, such as `counting`, *degrade* after a first, short learning phase. Finally, other skills, such as `coreference resolution`, *oscillate* significantly throughout pretraining, showing how models can not properly acquire them. This is in contrast to the coarse-grained COCO retrieval task for which the performance steadily increases over time. We conclude that it is particularly important to examine the training dynamics of fine-grained tasks, and that a single checkpoint might be inadequate for a number of skills. Results on all tasks are provided in App. B.2, including on Winoground for an ALBEF$_{4M}$ that we pretrained on GCP using the original codebase.

Same skills, same patterns? We next investigate whether closely-related tasks in different benchmarks have high correlation throughout pretraining. While we find that VALSE `action replacement` and SVO-Verb have a +55/67% Pearson/Spearman correlation, there is a -13/11% correlation between VALSE `actant swap` and SVO-Subject.

Looking at VALSE `spatial relations`, we find high correlation (+75/65%) with average VSR performance, and especially with relations such as `on top of`, `on`, `inside`, `by`, and `in`; mostly belonging to the 'Topological' category in VSR. On the other hand, we find almost no correlation with several 'Directional' (*e.g.*, `across from`) and 'Orientation' (*e.g.*, `parallel to`) relations, as well as with some 'Topological' ones (*e.g.*, `touching`); and even negative correlation (-40% or less) with `alongside`, `below`, `toward`, `part of` and `near`.

Finally, surprisingly, VSR dev and test splits are *not* positively correlated for all relations. While average performance is highly correlated (+77/78%), only a few relations have Pearson/Spearman coefficients larger than 30% (`in`, `on`, `above`, `within`, and `consists of`). On the other hand, `near`, `ahead of` and `adjacent to` are negatively correlated between dev and test sets, and most relations show very low correlations between the two sets. As a result, improvement in a given relation type on the dev set, will likely not transfer at test time.

Takeaways. When tested on fine-grained benchmarks, we observe that, compared to ALBEF, X-VLM is more sample efficient as it achieves higher performance with fewer training steps. Also, while some tasks steadily improve during pretraining, for others, the performance *degrades* or *fluctuates*. Moreover, surprisingly, the performance of tasks measuring similar skills but from different benchmarks do *not* always positively correlate.

7 Discussion

While recent pretrained VLMs achieve impressive performance on various downstream benchmarks (such as visual question answering and image retrieval), recent benchmarks have highlighted that they still struggle with tasks that require *fine-grained* understanding—where a model needs to correctly align various aspects of an image to their corresponding language entities. Yet, it is still not known to which extent recent fine-grained VLMs (*e.g.*, Zeng et al., 2022; Yao et al., 2022b; Li et al., 2022a; Dou et al., 2022) fare on such benchmarks. We address this gap by evaluating strong and fine-grained models on four benchmarks (Hendricks and Nematzadeh, 2021; Parcalabescu et al., 2022; Liu et al., 2023; Thrush et al., 2022), and encourage future work to report zero-shot fine-grained performance on our selection of benchmarks, especially if models are not open-source.

Our work contributes to a growing thread of research devoted to understand what is learned by pretrained VLMs, such as studying cross-attention patterns (Cao et al., 2020), cross-modal input ablations (Frank et al., 2021), probing linguistic and visual structure (Milewski et al., 2022; Salin et al., 2022; Nikolaus et al., 2022), robustness to words order (Akula et al., 2020; Thrush et al., 2022), and incorrectly fusing image and language modalities (Diwan et al., 2022). Here, we show that object modelling through a prediction loss (as done in X-VLM) results in notable improvements across all benchmarks, outperforming models trained on much larger amounts of Web data. Our analysis highlights that teaching VLMs concepts of objects (*e.g.*, by masking irrelevant parts of the image) is crucial for effectively learning fine-grained skills. Though our models rely on supervised data to learn better localisation, we hope our findings can encourage researchers to design better loss functions for image–text mapping from unsupervised, Web-scale data as well.

Finally, our results also highlight the challenges of evaluating fine-grained understanding: the recent benchmarks capture a variety of subtasks (from counting to relation understanding); to perform well on these subtasks, a model requires different skills. Indeed, we observe that, during training, model performance does not always increase for all subtasks, and in particular, fluctuates a lot for counting, coreference resolution, and various spatial relations. An important future direction is designing models that perform well on a larger range of these subtasks, where improving on one subtask does not degrade performance on the rest. It is unclear why benchmarks do not always correlate; possible reasons include the data itself (images selected for analysis, annotator instructions), or that different competencies are required for different fine-grained tasks. We hope future work can explore this further, possibly by closely examining data in fine-grained benchmarks or expanding the models used in analysis beyond what we used here.

Limitations

Our work focuses on assessing recent English VLMs on tasks which require fine-grained understanding. Here, we outline limitations that we believe are important considerations for future work.

First, we only examined a limited number of models. These include (i) strong coarse-grained models, such as ALBEF, CLIP, FLAMINGO and BLIP-2, and (ii) two strong fine-grained models, PEVL and X-VLM, that build on ALBEF. While we believe our selection of models is representative of strong components in pretrained VLMs (such as dual-encoder and cross-modal interactions), we could not easily evaluate different approaches towards fine-grained understanding (*e.g.*, Yao et al., 2022a; Li et al., 2022a) as the corresponding models and code are not open-source. We hence hope our study will motivate future work to report zero-shot performance on fine-grained benchmarks.

Second, we evaluate our models in a zero–shot setting using image–text matching. Future work could consider how fine-grained understanding improves when fine-tuning for specific tasks. As opposed to relying on image–text matching scores, alternative methods like input ablations, visualising attention or activations could also be used to gain an understanding of potential failure modes.

Third, though we note specific areas where model performance fluctuates a lot during pretraining, we look forward to future research that improves performance for various such areas, like existence and counting.

Finally, some datasets we use are quite small. For example, Winoground only has 1,600 data points. We hope that our analysis sheds light on the kinds of skills models struggle with and encourages more and larger datasets that test for these skills.

Ethics Statement

All datasets used in this work have been previously published. Multimodal datasets frequently include social biases (Meister et al., 2022), and we expect the models trained on them to reflect the biases in these datasets. Datasets also include images of people, and there is no mechanism for people to remove themselves from these datasets.

Multimodal models have many downstream uses. Some examples of beneficial applications include: more advanced image and video retrieval, visual description systems to aid the visually impaired, and interfaces which allow users to more seamlessly interact with smart home devices. Harmful applications might include surveillance, especially when imagery of people is being used without their consent, or fine-tuning a model to retrieve harmful content, such as pornographic material.

In this work, we aim to understand how models perform on fine-grained tasks which highlights current failure modes of our models. We hope insights from our work can inspire (i) novel models which perform well on a broad set of fine-grained tasks, as well as (ii) more high quality data to stress test our models. We hope our work also helps those who might use multimodal models in downstream applications better anticipate how well these models might perform on their tasks.

Acknowledgements

The authors would like to thank the anonymous reviewers, Antoine Miech, Ravichandra Addanki, Wojciech Stokowiec, Chris Dyer and the DeepMind Language Team for feedback on this project.

References

Arjun Akula, Spandana Gella, Yaser Al-Onaizan, Song-Chun Zhu, and Siva Reddy. 2020. Words aren't enough, their order matters: On the robustness of grounding visual referring expressions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6555–6565, Online. Association for Computational Linguistics.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*.

Igor Babuschkin, Kate Baumli, Alison Bell, Surya Bhupatiraju, Jake Bruce, Peter Buchlovsky, David Budden, Trevor Cai, Aidan Clark, Ivo Danihelka, Antoine Dedieu, Claudio Fantacci, Jonathan Godwin, Chris Jones, Ross Hemsley, Tom Hennigan, Matteo Hessel, Shaobo Hou, Steven Kapturowski, Thomas Keck, Iurii Kemaev, Michael King, Markus Kunesch, Lena Martens, Hamza Merzic, Vladimir Mikulik, Tamara Norman, George Papamakarios, John Quan, Roman Ring, Francisco Ruiz, Alvaro Sanchez, Rosalia Schneider, Eren Sezener, Stephen Spencer, Srivatsan Srinivasan, Wojciech Stokowiec, Luyu Wang, Guangyao Zhou, and Fabio Viola. 2020. The DeepMind JAX Ecosystem.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2021. Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs. *Transactions of the Association for Computational Linguistics*, 9:978–994.

Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. 2020. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VI*, volume 12351 of *Lecture Notes in Computer Science*, pages 565–580. Springer.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3558–3568.

Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J. Fleet, and Geoffrey Hinton. 2022. A unified sequence interface for vision tasks. In *Advances in Neural Information Processing Systems*.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023. PaLI: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Dasha Valter Kevin Robinson, Sharan Narang, Gaurav Mishra, Adams Yu, Yanping Huang Vincent Zhao, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. 2022. Why is Winoground hard? Investigating failures in visuolinguistic compositionality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2250, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, Jianfeng Gao, and Lijuan Wang. 2022. Coarse-to-fine vision-language pre-training with fusion in the backbone. In *Advances in Neural Information Processing Systems*.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. 2021. Vision-and-language or vision-for-language? On cross-modal influence in multimodal transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9847–9857, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, Rongrong Ji, and Chunhua Shen. 2022. Pyramid-CLIP: Hierarchical feature alignment for vision-language model pretraining. In *Advances in Neural Information Processing Systems*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. 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 (CVPR)*.

Xiaoshuai Hao, Yi Zhu, Srikar Appalaraju, Aston Zhang, Wanqian Zhang, Bo Li, and Mu Li. 2023. MixGen: A new multi-modal data augmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 379–389.

Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. 2021. Decoupling the role of data, attention, and losses in multimodal transformers. *Transactions of the Association for Computational Linguistics*, 9:570–585.

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

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack William Rae, and Laurent Sifre. 2022. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*.

Drew A. Hudson and Christopher D. Manning. 2019. 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 (CVPR)*.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. 2021. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings* of Machine Learning Research, pages 4651–4664. PMLR.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar. Association for Computational Linguistics.

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 Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision*, 123(1):32–73.

Benno Krojer, Vaibhav Adlakha, Vibhav Vineet, Yash Goyal, Edoardo Ponti, and Siva Reddy. 2022. Image retrieval from contextual descriptions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3426–3440, Dublin, Ireland. Association for Computational Linguistics.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The Open Images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *Int. J. Comput. Vision*, 128(7):1956–1981.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. 2019. Set Transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3744–3753. PMLR.

Juncheng Li, Xin He, Longhui Wei, Long Qian, Linchao Zhu, Lingxi Xie, Yueting Zhuang, Qi Tian, and Siliang Tang. 2022a. Fine-grained semantically aligned vision-language pre-training. In *Advances in Neural Information Processing Systems*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR.

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

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*, volume 34, pages 9694–9705. Curran Associates, Inc.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision – ECCV 2020*, pages 121–137, Cham. Springer International Publishing.

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

Fangyu Liu, Guy Edward Toh Emerson, and Nigel Collier. 2023. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11–20.

Nicole Meister, Dora Zhao, Angelina Wang, Vikram V Ramaswamy, Ruth Fong, and Olga Russakovsky. 2022. Gender artifacts in visual datasets. *arXiv preprint arXiv:2206.09191*.

Victor Milewski, Miryam de Lhoneux, and Marie-Francine Moens. 2022. Finding structural knowledge in multimodal-BERT. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5658–5671, Dublin, Ireland. Association for Computational Linguistics.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. ClipCap: CLIP prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Mitja Nikolaus, Emmanuelle Salin, Stephane Ayache, Abdellah Fourtassi, and Benoit Favre. 2022. Do vision-and-language transformers learn grounded predicate-noun dependencies? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1538–1555, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, Dublin, Ireland. Association for Computational Linguistics.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2641–2649.

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. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Emmanuelle Salin, Badreddine Farah, Stéphane Ayache, and Benoit Favre. 2022. Are vision-language transformers learning multimodal representations? a probing perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11248–11257.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8429–8438.

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

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

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. FLAVA: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15638–15650.

Andreas Peter Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. 2022. How to train your ViT? Data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

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

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. 2021. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10347–10357. PMLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. 2022. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15671–15680.

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

Yuan Yao, Qianyu Chen, Ao Zhang, Wei Ji, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2022b. PEVL: Position-enhanced pre-training and prompt tuning for vision-language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 11104–11117, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yan Zeng, Xinsong Zhang, and Hang Li. 2022. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 25994–26009. PMLR.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. VinVL: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5579–5588.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

A Experimental Setup

In this section, we provide further details on the experimental setups that we used for our studies.

A.1 Evaluated Models: Details

We provide more details on the models we use to evaluate progress in fine-grained V&L understanding. See Table 5 for an overview.[14]

ALBEF (Li et al., 2021) is a recent VLM that has gained popularity due to its design choices, effectively combining core components in V&L learning, such as a contrastive objective and cross-attention, that result in strong downstream performance. ALBEF is a dual-stream encoder (Bugliarello et al., 2021) that first encodes images and captions independently with a vision (ViT; Dosovitskiy et al. 2021; Touvron et al. 2021) and text (BERT; Devlin et al. 2019) Transformer, respectively; and then fuses them in a cross-modal Transformer. The model is pretrained with three objectives: masked language modelling (MLM), unimodal image–text contrastive learning and cross-modal image–text matching. We refer to the original work for more details. While ALBEF does not explicitly train for fine-grained understanding, it serves as an important baseline since our three other models build on top of it.

BLIP (Li et al., 2022b) is a unified V&L understanding and generation model, that can be applied to a wide range of downstream tasks. A key component to BLIP's success is CAPFILT: a dataset boostrapping method which the authors use to generate synthetic captions and removing noisy pairs from large-scale Web data. Moreover, unlike any other model we evaluate, BLIP uses an autoregressive language modelling (LM) objective to convert visual information into coherent captions, allowing us to evaluate the potential benefits of this objective to learn fine-grained relationships. BLIP is not explicitly trained for fine-grained understanding, however, we believe it is important to assess whether generative language modelling and its data contributions that enhance downstream performance also lead to better fine-grained skills.

PEVL (Yao et al., 2022b) explicitly connects image regions and text tokens through cross-modal position modelling. Similar to Pix2Seq (Chen et al., 2022), PEVL expresses visual positions in text by appending the bounding box coordinates corresponding to a given (annotated) entity in the caption, surrounded by two special tokens ‘<’ and ‘>’: “`A cat < 10 73 206 175 > is napping.`” The bounding box coordinates are discretised and added to the text vocabulary. Starting from an ALBEF$_{14M}$ checkpoint, PEVL is pretrained by recovering masked text and position tokens through a generalised MLM objective. The model was trained on a diverse corpus of referring expressions, captions with visual coreferences, question answering, commonsense reasoning, object detection and region descriptions data (Table 2). Unlike ALBEF, PEVL is explicitly trained to learn fine-grained, grounded representations of entities by predicting their coordinates in a unified MLM framework. We evaluate three different models released by the authors, which differ in their pretraining and fine-tuning data: PEVL$_{14M}$, underwent a second-stage pretraining on multiple supervised tasks (Table 5); PEVL$_{GRD}$, which was further fine-tuned for position-output tasks such as phrase grounding (Plummer et al., 2015); and PEVL$_{VRD}$, which was fine-tuned for the position-input task of visual relation detection (Krishna et al., 2017).

X-VLM (Zeng et al., 2022) also aims at learning to locate visual concepts in the image given the associated texts. Similar to the ALBEF architecture, the model consists of an image encoder, a text encoder, and a cross-modal encoder. However, unlike PEVL, X-VLM models visual position through an additional bounding box prediction head: given the visually grounded representation of an object label, the model is trained to regress the object's bounding box (bbox) coordinates. The authors use both object detection labels and region descriptions to learn multi-grained alignments. The pretraining objective is a linear combination of this bbox loss and the losses defined in ALBEF to align texts and visual concepts (for more details, see Section 5).

In addition to the above models, which we extensively discuss, we also evaluate the following models, based on dual-encoder and frozen LLMs.

CLIP (Radford et al., 2021) is a widely used dual-encoder network. The model consists of two encoders, one for images and one for text, trained to represent both modalities in a joint space via an unsupervised contrastive objectives over more than 400M image–text pairs from the Web. Due

[14] Each model's text and multimodal layers were originally initialised with the weights of BERT$_{BASE}$ (Devlin et al., 2019).

Model			Data			
Name	ViT	Img Res	Datasets	# Img	# Cap	# Ann
ALBEF$_{4M}$	DeiT-B/16	256×256	4M: COCO+SBU+VG+CC$_{3M}$	4.0M	5.1M	-
ALBEF$_{14M}$	DeiT-B/16	256×256	14M: 4M + CC$_{12M}$	14.1M	15.2M	-
BLIP$_{14M}$	ViT-B/16	224×224	CapFilt/B(14M)	14.1M	15.2M	-
BLIP$_{129M}$	ViT-B/16	224×224	CapFilt/B(14M + LAION)	129.1M	130.2M	-
BLIP$_{129M}$-CapFilt/L	ViT-B/16	224×224	CapFilt/L(14M + LAION)	129.1M	130.2M	-
BLIP-ViT/L$_{129M}$	ViT-L/16	224×224	CapFilt/L(14M + LAION)	129.1M	130.2M	-
PEVL$_{14M}$	ALBEF$_{14M}$	256×256	14M→RefCOCO{,+,g}+F30KE+GQA+VCR+VG	14.4M	15.2M	4.7M
PEVL$_{GRD}$	PEVL$_{14M}$	512×512	PEVL$_{14M}$ →RefCOCO{,+,g}+F30KE	14.4M	15.2M	4.7M
PEVL$_{VRD}$	PEVL$_{14M}$	512×512	PEVL$_{14M}$ →VG	14.4M	15.2M	6.2M
X-VLM$_{4M}$	Swin-B/32	224×224	4M	4.0M	5.1M	6.2M
X-VLM$_{16M}$	Swin-B/32	224×224	14M + Objects365 + OpenImages	17.4M	16.2M	12.4M

Table 5: Overview of core evaluated models. All the models cross-attend to visual features, and use contrastive learning (CL) and a (masked) language modelling objective. Fine-grained models also predict object locations. Unsupervised pretraining data includes COCO (Lin et al., 2014), SBU (Ordonez et al., 2011), VG (Krishna et al., 2017), CC$_{3M}$ (Sharma et al., 2018), CC$_{12M}$ (Changpinyo et al., 2021) and LAION (Schuhmann et al., 2021). Supervised data additionally includes RefCOCO and RefCOCO+ (Kazemzadeh et al., 2014), RefCOCOg (Mao et al., 2016), F30KE (Plummer et al., 2015), GQA (Hudson and Manning, 2019), VCR (Zellers et al., 2019), Objects365 (Shao et al., 2019) and OpenImages (Kuznetsova et al., 2020). In Table 2, we also list downstream performance on VQAv2 (Goyal et al., 2017), NLVR2 (Suhr et al., 2019) and RefCOCO+.

to its simplicity and wide adoption, we report its performance as a strong, representative baseline.

ClipCap (Mokady et al., 2021) is an autoregressive encoder–decoder network. The image encoder is a pretrained CLIP model, while the text decoder is a pretrained GPT-2 (Radford et al., 2019) language model. The authors propose to learn a lightweight Transformer-based network to map CLIP embeddings into a fixed length prefix. The mapping network and the text decoder are fine-tuned to learn how to generate captions, while the CLIP image encoder is frozen. At inference time, the model generates the caption word after word, starting from the CLIP-based prefix. We report performance for the two released versions—one fine-tuned on COCO, the other on CC$_{3M}$—by ranking positive and negative samples on their likelihood.

Flamingo (Alayrac et al., 2022) is a state-of-the-art VLM capable of tackling a wide range of vision and language tasks from a few input/output examples. To achieve this, the model relies on a pretrained CLIP-like image encoder and a strong pretrained LLM (Hoffmann et al., 2022), both kept frozen. To ingest images and videos, the model learns a small fixed number of visual tokens (Lee et al., 2019; Jaegle et al., 2021). The model is pretrained to generate text from a sequence of text tokens interleaved with images and/or videos.

BLIP-2 (Li et al., 2023) is the most recent, state-of-the-art VLM based on frozen large image encoders and frozen LLMs (Zhang et al., 2022; Chung et al., 2022). Like ClipCap, BLIP-2 learns a mapping network, which in this case is a Transformer model initialised from BERT$_{BASE}$. The mapping network learns visual query tokens to map the visual representations to the frozen LLM in two stages: a V&L representation stage, and a generative learning stage. The model was pretrained with the same objectives and on the same 129M image–caption data as BLIP. Following the authors' setup for image–text retrieval and matching, we use the BLIP-2 model after the first-state pretraining.

A.2 Re-implementation Setup

We re-implement ALBEF and X-VLM in JAX (Babuschkin et al., 2020) to ensure full control of modelling, data, and initialisation decisions.[15] We note ALBEF's vision encoder is initialised with a pretrained ViT-B/16 encoder (Touvron et al., 2021) with an input resolution of 256×256 pixels, but X-VLM adopts a more efficient Swin-B/32 (Liu et al., 2021) encoder with input resolution of 224×224 pixels. In our re-implementation we initialise both models with a ViT-B/16 with a 224×224 input resolution pretrained on ImageNet-

[15]To verify our implementation, we compare an ALBEF model trained in our codebase with one trained in the original codebase. Specifically, we pretrain both models on COCO by initialising their visual encoder with a CLIP ViT-B/16 model, and their text encoder with a BERT$_{BASE}$ model. The two models perform similarly on both zero-shot Flickr30K and COCO retrieval tasks with a gap below 1pp Recall@1.

	Model Name	Size	SVO-Probes Avg.	VALSE Avg.	VSR Random Test Avg.	Winoground Text	Image	Group	Flickr30K TR@1	IR@1	COCO TR@1	IR@1
	Random		50.0	50.0	50.0	25.0	25.0	12.5	0.1	0.1	0.02	0.02
	LXMERT	263M	-	59.6	72.5†	19.2	7.0	4.0	-	-	-	-
	UNITER$_{\text{Large}}$	303M	-	-	-	38.0	14.0	10.5	80.7	66.2	64.1	48.8
	12-in-1	270M	-	75.1	-	-	-	-	-	67.8†	-	68.0†
	CLIP (ViT-B/32)	151M	81.6	64.0	N/A	30.7	10.5	8.0	88.0	68.7	58.4	37.8
	CLIPCAP$_{\text{CC3M}}$	295M	83.1	65.7	N/A	12.2	14.7	5.5	26.4	44.1	6.7	24.3
	CLIPCAP$_{\text{COCO}}$	295M	84.1	68.5	N/A	12.2	14.7	5.5	27.8	52.2	8.1	38.4
	FLAMINGO	80B	88.4	**75.3**	N/A	-	-	-	-	-	-	-
	BLIP-2	1.2B	86.5	74.0	61.5	43.0	22.0	18.2	**95.5**	**86.7**	**80.7**	**64.2**
1	ALBEF$_{\text{4M}}$	500M	87.6	69.1	57.3	29.2	15.5	11.0	85.2	69.4	69.7	51.1
2	X-VLM$_{\text{4M}}$♯	239M	88.9	72.4	63.0	44.0	**26.7**	**21.5**	85.3	71.9	70.8	55.6
3	ALBEF$_{\text{14M}}$	500M	88.6	69.4	58.3	32.5	16.2	12.7	90.9	75.9	73.2	54.8
4	BLIP$_{\text{14M}}$	638M	48.7	67.8	49.7	36.5	18.5	14.5	82.6	78.4	70.4	57.3
5	PEVL$_{\text{14M}}$♯	500M	86.2	68.9	57.5	33.2	15.7	12.2	74.9	60.0	45.9	33.2
6	PEVL$_{\text{GRD}}$♯	502M	88.5	69.5	57.7	36.2	15.0	12.0	71.8	77.6	42.8	37.7
7	PEVL$_{\text{VRD}}$♯	502M	84.8	64.5	59.5	31.2	12.0	7.5	68.0	55.7	38.3	30.6
8	X-VLM$_{\text{16M}}$♯	239M	**90.0**	74.5	**64.3**	**46.7**	24.5	21.2	87.7	74.9	71.6	56.1
9	BLIP$_{\text{129M}}$	638M	51.4	68.8	46.9	35.5	15.0	11.7	90.2	79.5	71.9	58.6
10	BLIP$_{\text{129M}}$-CAPFILT/L	638M	51.2	68.2	48.7	34.7	15.2	12.2	89.1	79.7	72.2	57.8
11	BLIP-VIT/L$_{\text{129M}}$	1.1B	50.8	70.3	50.3	34.7	14.5	12.2	90.4	80.6	74.2	59.3

Table 6: Overall performance of our evaluated models on fine-grained benchmarks and zero-shot retrieval tasks. The overall best values for each task are marked in **bold**. ♯ marks the fine-grained models. † denotes performance after task fine-tuning. X-VLM significantly outperforms the other models that we evaluate on fine-grained tasks.

Model	Existence quantifiers	Plurality number	Counting balanced	sns.†	adv.†	Sp.rel.‡ relations	Action repl.†	actant swap	Coreference standard	clean	Foil-it!	Avg.
Random	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
GPT-2	58.0	51.9	51.6	49.8	45.3	75.0	66.8	76.9	54.5	50.0	80.7	60.1
CLIP	66.9	56.2	62.1	62.5	57.5	64.3	75.6	68.6	52.1	49.7	88.8	64.0
LXMERT	78.6	64.4	62.2	69.2	42.6	60.2	54.8	45.8	46.8	44.2	87.1	59.6
12-in-1	**95.6**	72.4	**76.7**	**80.2**	77.3	67.7	65.9	58.9	**75.7**	**69.2**	86.9	75.1
CLIPCAP$_{\text{CC3M}}$	66.3	54.8	49.4	50.1	51.5	83.2	75.5	87.9	45.1	45.2	94.7	65.7
CLIPCAP$_{\text{COCO}}$	74.9	60.6	55.0	53.0	53.0	**89.7**	71.0	86.5	47.5	49.0	**97.1**	68.5
FLAMINGO	63.6	59.8	58.2	55.2	**80.2**	**89.7**	**86.7**	**92.8**	72.2	65.4	97.0	**75.3**
BLIP-2	83.6	**79.6**	70.2	68.7	68.0	65.6	84.4	63.2	62.6	58.7	96.0	74.0
ALBEF$_{\text{4M}}$	71.3	78.8	62.2	65.1	59.8	73.1	73.6	58.4	52.4	55.8	95.5	69.1
X-VLM$_{\text{4M}}$	80.0	77.8	69.0	68.4	72.5	74.8	77.3	65.0	50.1	48.1	92.5	72.4
ALBEF$_{\text{14M}}$	69.5	76.0	61.5	61.0	64.5	70.7	77.6	60.5	55.9	61.5	96.1	69.4
BLIP$_{\text{14M}}$	82.4	73.8	61.8	62.6	63.7	65.2	74.7	55.2	52.3	42.3	92.3	67.8
PEVL$_{\text{14M}}$	89.7	65.5	66.0	66.2	57.3	67.9	73.5	59.4	58.2	56.7	90.9	68.9
PEVL$_{\text{GRD}}$	91.1	63.9	70.0	70.9	63.2	62.4	74.4	57.1	53.8	49.0	92.6	69.5
PEVL$_{\text{VRD}}$	83.8	61.8	62.8	70.3	40.4	64.5	68.1	53.2	47.7	42.3	94.1	64.5
X-VLM$_{\text{16M}}$	83.6	78.7	71.5	72.0	74.8	73.1	79.2	64.6	60.0	49.0	91.9	74.5
BLIP$_{\text{129M}}$	78.2	75.9	63.4	63.4	58.5	66.2	75.2	59.0	56.4	52.9	93.2	68.8
BLIP$_{\text{129M}}$-CAPFILT/L	75.4	75.0	64.7	68.8	53.0	66.7	73.0	60.6	48.2	51.0	93.8	68.2
BLIP-VIT/L$_{\text{129M}}$	73.3	77.7	68.2	67.6	61.2	71.8	75.3	60.8	51.1	45.2	96.1	70.3

Table 7: Performance on the VALSE benchmark according to pairwise ranking accuracy. Best results are in **bold**. †**sns.** Counting small numbers. **adv.** Counting adversarial. **repl.** Action replacement. ‡ **Sp.rel.** Spatial relations.

21k (Steiner et al., 2022), to ensure that different initialisation is not responsible for the results.

We pretrain our models on the same 4M and 14M datasets that were originally used by the authors (Table 2), but note that only 1.8M and 11.2M data points were available for CC_{3M} and CC_{12M}, respectively. For object detection data, we use the data points released by the X-VLM authors, and interleave captioning and detection data with a 2:1 ratio following their official implementation. Following (Zeng et al., 2022), we pretrain our models for 200K steps using a batch size of 512 and 1024 samples for ALBEF and X-VLM, respectively. We pretrain once, using the same hyperparameters

Model	Object			Relation			Both			1 Main Pred			2 Main Preds		
	Text	Image	Group	Text	Image	Group	Text	Image	Group	Text	Image	Group	Text	Image	Group
Random	25.00	25.00	12.50	25.00	25.00	12.50	25.00	25.00	12.50	25.00	25.00	12.50	25.00	25.00	12.50
MTurk Human	92.20	90.78	88.65	89.27	90.56	86.70	76.92	57.69	57.69	87.33	85.62	82.53	95.37	96.30	93.52
LXMERT	22.70	9.22	6.38	17.60	5.58	2.58	15.38	7.69	3.85	19.18	8.56	5.14	19.44	2.78	0.93
$\text{UNITER}_{\text{Large}}$	39.01	12.77	9.93	36.05	14.16	9.87	50.00	19.23	19.23	40.07	16.44	13.36	32.41	7.41	2.78
CLIP (ViT-B/32)	34.75	7.80	6.38	22.75	8.58	5.58	**80.77**	**42.31**	**38.46**	35.27	13.01	10.27	18.52	3.70	1.85
$\text{ClipCap}_{\text{CC3M}}$	14.18	17.02	7.80	11.16	12.02	3.43	11.54	26.92	11.54	13.70	16.10	6.51	8.33	11.11	2.78
$\text{ClipCap}_{\text{COCO}}$	12.77	17.02	5.67	12.88	9.87	3.86	23.08	34.62	19.23	14.73	16.44	6.85	10.19	7.41	1.85
BLIP-2	47.52	**27.66**	**21.99**	38.20	17.60	14.59	61.54	30.77	30.77	48.63	26.37	22.26	27.78	10.19	7.41
ALBEF_{4M}	29.79	12.77	8.51	26.61	15.02	10.73	50.00	34.62	26.92	33.22	19.18	14.04	18.52	5.56	2.78
X-VLM_{4M}	46.10	**27.66**	**21.99**	41.63	**24.46**	19.31	53.85	**42.31**	**38.46**	47.60	**30.48**	25.68	34.26	**16.67**	**10.19**
$\text{ALBEF}_{\text{14M}}$	29.79	15.60	9.22	30.90	14.16	12.02	61.54	38.46	**38.46**	35.27	18.49	14.38	25.00	10.19	8.33
BLIP_{14M}	41.13	24.11	17.73	32.19	14.16	11.16	50.00	26.92	26.92	42.12	21.92	18.15	21.30	9.26	4.63
PEVL_{14M}	31.21	14.89	10.64	33.48	14.59	11.59	42.31	30.77	26.92	36.30	19.52	15.75	25.00	5.56	2.78
PEVL_{GRD}	39.01	14.89	12.77	33.91	13.73	10.30	42.31	26.92	23.08	37.67	17.47	15.07	32.41	8.33	3.70
PEVL_{VRD}	26.95	10.64	7.09	32.19	12.45	6.87	46.15	15.38	15.38	31.85	11.64	8.22	29.63	12.96	5.56
$\text{X-VLM}_{\text{16M}}$	**48.23**	23.40	19.86	**44.21**	23.18	**20.17**	61.54	**42.31**	**38.46**	**51.03**	29.11	**26.03**	**35.19**	12.04	8.33
$\text{BLIP}_{\text{129M}}$	37.59	17.02	10.64	34.76	12.02	10.73	30.77	30.77	26.92	40.07	18.84	14.73	23.15	4.63	3.70
$\text{BLIP}_{\text{129M}}$-CapFilt/L	34.04	16.31	11.35	33.48	13.30	11.16	50.00	26.92	26.92	38.70	19.18	15.41	24.07	4.63	3.70
$\text{BLIP-ViT/L}_{\text{129M}}$	35.46	16.31	13.48	32.62	12.88	11.59	50.00	19.23	11.54	39.04	17.81	15.07	23.15	5.56	4.63

Table 8: Results on Winoground by linguistic tag. Best results are in **bold**.

Model	Symbolic			Pragmatics			Same Image Series		
	Text	Image	Group	Text	Image	Group	Text	Image	Group
Random	25.00	25.00	12.50	25.00	25.00	12.50	25.00	25.00	12.50
MTurk Human	96.43	92.86	92.86	58.82	41.18	41.18	95.65	91.30	91.30
LXMERT	28.57	3.57	3.57	17.65	5.88	0.00	8.70	4.35	0.00
$\text{UNITER}_{\text{Large}}$	39.29	28.57	17.86	35.29	0.00	0.00	4.35	8.70	0.00
CLIP (ViT-B/32)	39.29	3.57	3.57	35.29	5.88	5.88	8.70	0.00	0.00
$\text{ClipCap}_{\text{CC3M}}$	21.43	21.43	10.71	5.88	5.88	0.00	0.00	8.70	0.00
$\text{ClipCap}_{\text{COCO}}$	25.00	25.00	14.29	23.53	17.65	**17.65**	13.04	13.04	0.00
BLIP-2	42.86	28.57	25.00	41.18	**23.53**	**17.65**	21.74	13.04	4.35
ALBEF_{4M}	42.86	25.00	17.86	17.65	17.65	5.88	8.70	0.00	0.00
X-VLM_{4M}	50.00	**32.14**	**32.14**	41.18	**23.53**	**17.65**	30.43	**26.09**	**13.04**
$\text{ALBEF}_{\text{14M}}$	39.29	14.29	14.29	17.65	0.00	0.00	26.09	4.35	4.35
BLIP_{14M}	39.29	25.00	17.86	23.53	17.65	**17.65**	8.70	4.35	0.00
PEVL_{14M}	35.71	14.29	14.29	29.41	11.76	5.88	13.04	8.70	4.35
PEVL_{GRD}	35.71	7.14	7.14	29.41	11.76	11.76	26.09	8.70	4.35
PEVL_{VRD}	42.86	10.71	7.14	23.53	5.88	0.00	**34.78**	17.39	8.70
$\text{X-VLM}_{\text{16M}}$	42.86	21.43	17.86	**47.06**	11.76	5.88	26.09	4.35	4.35
$\text{BLIP}_{\text{129M}}$	**57.14**	14.29	14.29	35.29	11.76	11.76	26.09	0.00	0.00
$\text{BLIP}_{\text{129M}}$-CapFilt/L	50.00	14.29	14.29	35.29	5.88	5.88	21.74	0.00	0.00
$\text{BLIP-ViT/L}_{\text{129M}}$	39.29	14.29	14.29	29.41	0.00	0.00	13.04	0.00	0.00

Table 9: Results on Winoground by visual tag. Best results are in **bold**.

Model	Subj.	Verb	Obj.	Avg.
Random	50.0	50.0	50.0	50.0
CLIP (ViT-B/32)	83.6	79.0	88.1	81.6
$\text{ClipCap}_{\text{CC3M}}$	84.2	80.5	90.2	83.1
$\text{ClipCap}_{\text{COCO}}$	87.3	81.5	89.8	84.1
Flamingo	90.1	86.7	92.3	88.4
BLIP-2	87.6	84.6	91.7	86.5
ALBEF_{4M}	88.5	85.4	93.7	87.6
X-VLM_{4M}	89.3	87.1	94.5	88.9
$\text{ALBEF}_{\text{14M}}$	89.4	86.4	94.7	88.6
BLIP_{14M}	49.8	48.8	47.5	48.7
PEVL_{14M}	89.4	82.9	93.9	86.2
PEVL_{GRD}	**91.2**	85.9	94.6	88.5
PEVL_{VRD}	90.1	81.1	92.3	84.8
$\text{X-VLM}_{\text{16M}}$	90.3	**88.4**	**94.6**	**90.0**
$\text{BLIP}_{\text{129M}}$	50.8	51.4	51.8	51.4
$\text{BLIP}_{\text{129M}}$-CapFilt/L	49.4	51.3	52.5	51.2
$\text{BLIP-ViT/L}_{\text{129M}}$	50.0	50.9	50.9	50.8

Table 10: Performance on the SVO-Probes benchmark according to pairwise ranking accuracy. Best results are in **bold**.

Model	Adjacency	Directional	Orientation	Projective	Proximity	Topological	Unallocated	Overall
Random	50.0 / 50.0	50.0 / 50.0	50.0 / 50.0	50.0 / 50.0	50.0 / 50.0	50.0 / 50.0	50.0 / 50.0	50.0 / 50.0
BLIP-2	59.8 / 54.9	50.0 / 43.3	52.5 / 57.1	59.8 / 63.6	56.2 / 51.2	66.4 / 67.0	75.0 / 66.7	61.2 / 61.5
ALBEF_{4M}	52.3 / 51.1	38.6 / 42.2	55.9 / **58.0**	61.7 / 60.2	56.2 / **55.3**	58.6 / 59.2	65.6 / 56.9	58.0 / 57.3
X-VLM_{4M}	57.6 / 57.7	56.8 / 43.3	59.3 / 52.7	**69.2** / 66.1	57.8 / 54.5	**71.2** / 68.4	75.0 / 62.7	66.6 / 63.0
$\text{ALBEF}_{\text{14M}}$	52.3 / 54.2	59.1 / 40.0	55.9 / **58.0**	59.8 / 62.6	46.9 / 52.0	66.8 / 58.9	71.9 / 58.8	60.2 / 58.3
BLIP_{14M}	56.8 / 49.3	56.8 / 50.0	57.6 / 47.3	42.5 / 49.3	51.6 / 48.0	45.1 / 51.8	50.0 / 41.2	47.4 / 49.7
PEVL_{14M}	47.0 / 55.3	56.8 / 48.9	57.6 / 56.2	61.9 / 60.8	51.6 / 48.8	62.4 / 57.4	71.9 / 58.8	59.3 / 57.5
PEVL_{GRD}	53.8 / 53.5	**65.9** / 50.0	59.3 / 52.7	60.9 / 59.4	60.9 / 54.5	62.7 / 60.2	75.0 / 58.8	61.1 / 57.7
PEVL_{VRD}	54.5 / 55.6	59.1 / 52.2	61.0 / 53.6	59.8 / 60.4	59.4 / 54.5	64.1 / 63.1	68.8 / 64.7	60.7 / 59.5
$\text{X-VLM}_{\text{16M}}$	**61.4** / **58.5**	**65.9** / 46.7	**64.4** / **58.0**	68.4 / **67.7**	**62.5** / 52.0	70.5 / **68.7**	**84.4** / **68.6**	**67.9** / **64.3**
$\text{BLIP}_{\text{129M}}$	44.7 / 41.2	43.2 / 52.2	52.5 / 53.6	53.6 / 45.4	53.1 / 49.6	50.2 / 49.7	40.6 / 37.3	50.5 / 46.9
$\text{BLIP}_{\text{129M}}$-CapFilt/L	57.6 / 49.3	36.4 / 57.8	47.5 / 53.6	45.9 / 45.5	48.4 / 47.2	48.5 / 51.1	37.5 / 41.2	47.7 / 48.7
$\text{BLIP-ViT/L}_{\text{129M}}$	56.1 / 51.8	29.5 / **58.9**	49.2 / 52.7	46.9 / 48.5	53.1 / 43.9	49.8 / 51.8	46.9 / 47.1	48.7 / 50.3

Table 11: Dev/Test results on the VSR Random dataset. Best results are in **bold**.

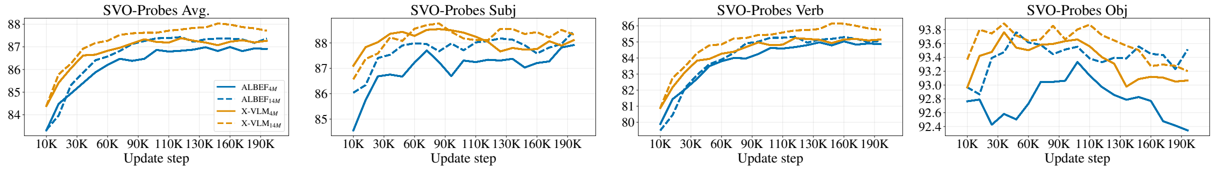

Figure 4: Training dynamics on SVO-Probes subtasks. Random performance is 50%.

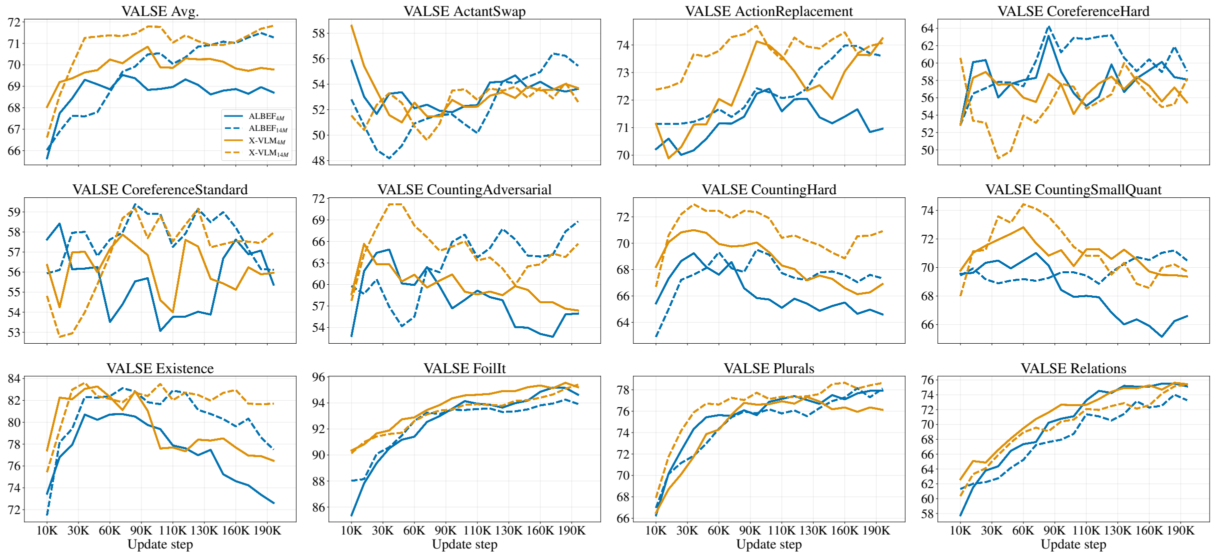

Figure 5: Training dynamics on VALSE subtasks. Random performance is 50%.

as the authors.[16] Training our models takes around 1.5 days on Cloud TPUv4 (a 2x2x2 slice). We evaluate our models on both fine-grained benchmarks (SVO-Probes, VALSE and VSR) and on two zero-shot, coarse retrieval tasks (Flickr30K and COCO).

B Results

B.1 Results by Subtask

Table 6 compares overall performance of our evaluated models (Section 3) with the state-of-the-art models in each of four fine-grained benchmarks (Section 2). Results for each subtask are reported in Tables 7 to 11.

In addition to the core discussion in Section 4, we note that FLAMINGO achieves the overall best performance on VALSE; and that the coarse-grained BLIP-2 model performs remarkably well on our range of fine-grained tasks, especially on VALSE, VSR and Winoground. This could be due to a number of factors, such as a larger ViT encoder, the usage of visual queries and the different formulations for the ITC and ITM objectives. We leave a deeper investigation of large VLMs to future work.

Moreover, we also note that CLIPCAP well on VALSE `spatial relations` and `action` subtasks, wherein its GPT-2 backbone already performs better than most VLMs. This is further proof of the efficacy of adapting strong LMs for V&L tasks.

B.2 Full Dynamics of Fine-grained Tasks

Figures 4 to 7 display pretraining dynamics for our re-implemented ALBEF_{4M}, ALBEF_{14M}, X-VLM_{4M}, and X-VLM_{14M} models. For better visualisation, our curves have been smoothed by a 0.6 factor through exponential moving average.

Finally, Figure 8 shows how performance on Winoground evolves when pretraining an ALBEF_{4M} model.[17] Looking at overall performance, we see that a model's score can vary by more than 4pp from one epoch to the next. While longer pretraining seems beneficial, some subtasks, such as `Linguistic:Both` and `Visual:Series`, fluctuate considerably; and after 20 epochs, the Image score starts decreasing on other subtasks, such as `Linguistic:Object` and `Visual:Symbolic`.

[16] The X-VLM authors trained for 200K steps of image captioning data, not counting batches of detection datasets. We count each batch towards the final number of steps, hence effectively training for fewer steps than Zeng et al. (2022).

[17] We note that we used an image resolution of 224×224 pixels, and a batch size of 256 (instead of 512) as we pretrained on a GCP instance with 4× A100 GPUs (instead of the 8× A100 GPUs originally used by the authors).

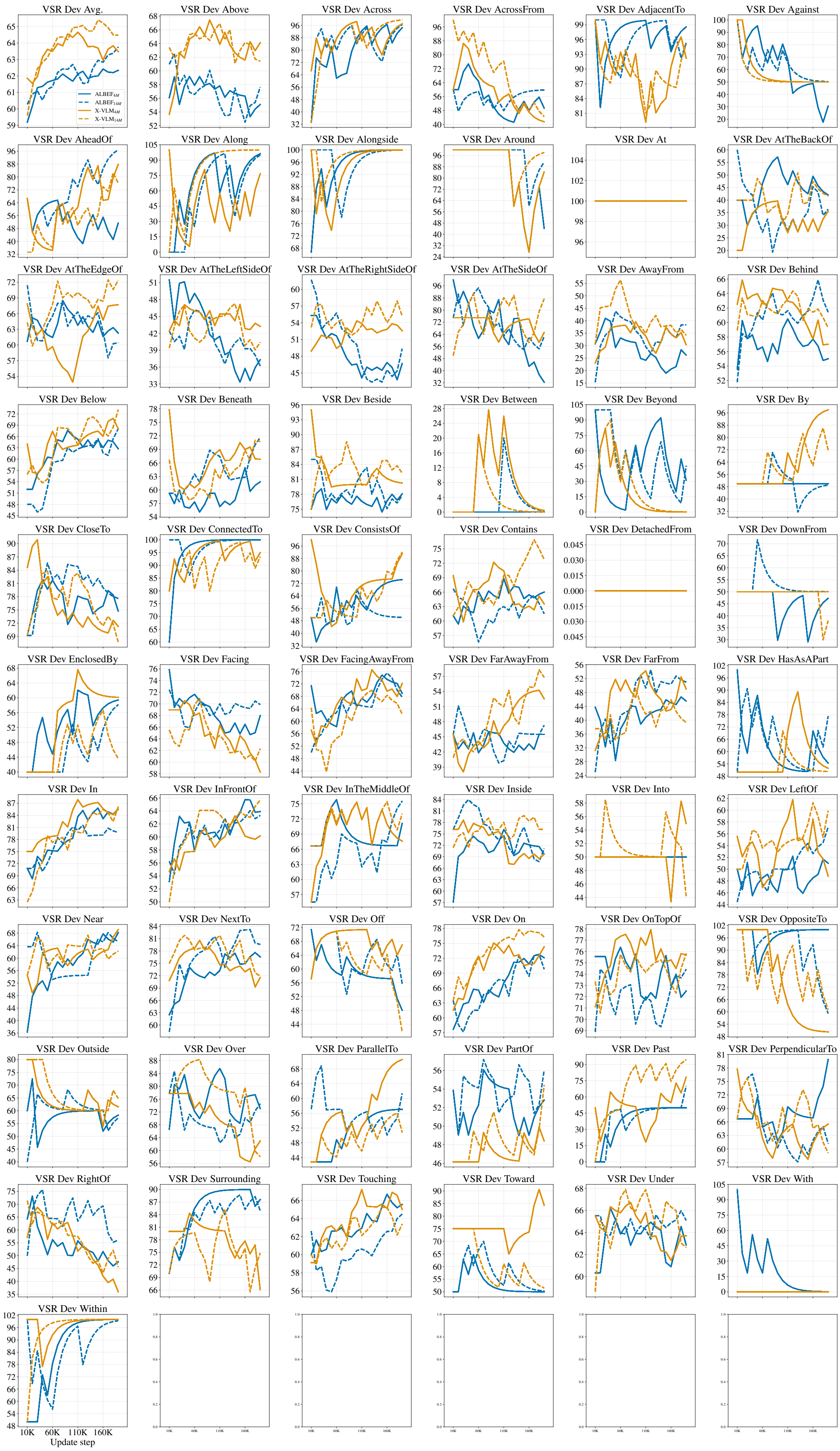

Figure 6: Training dynamics on VSR Random dev set subtasks. Random performance is 50%.

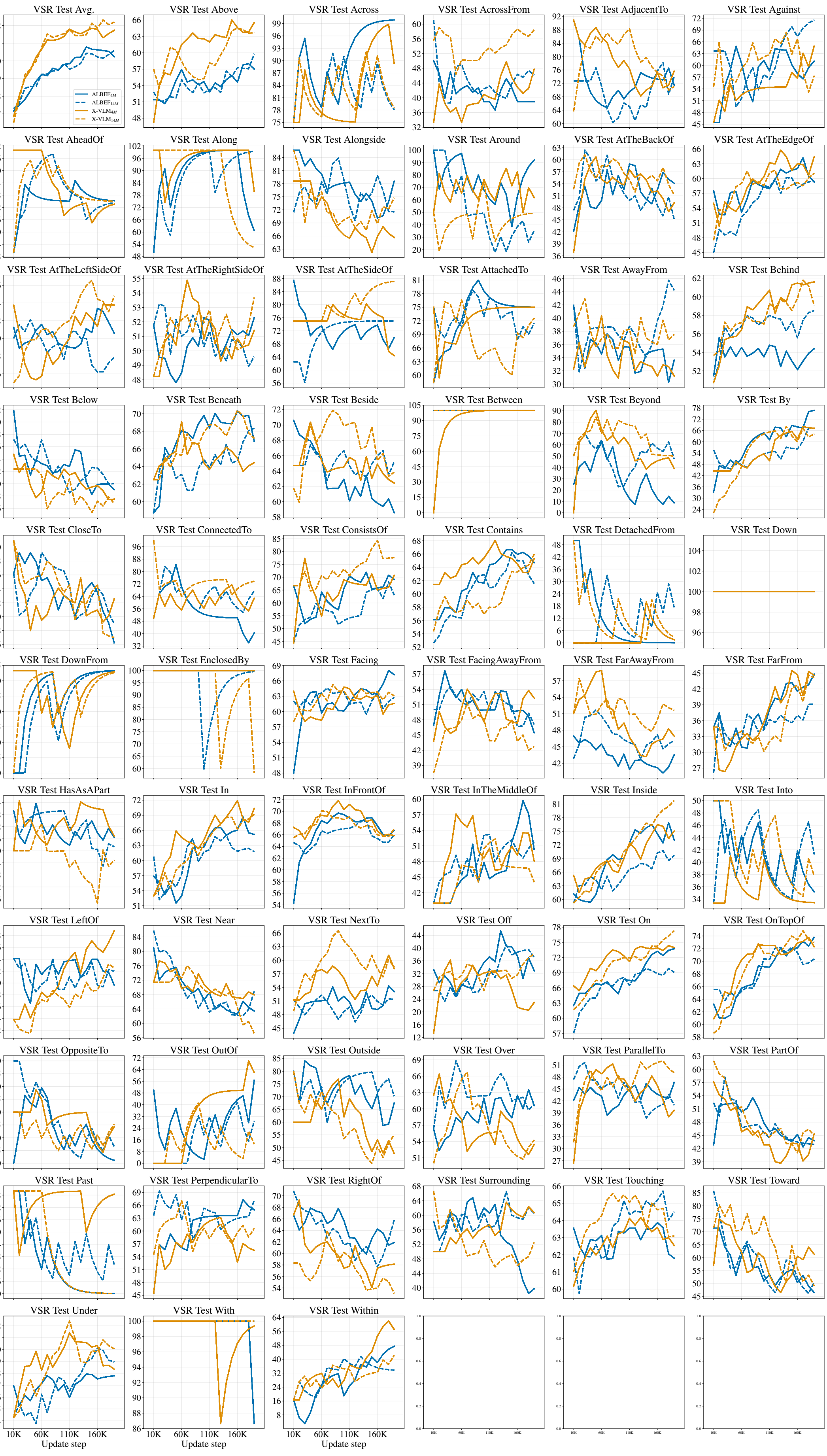

Figure 7: Training dynamics on VSR Random test set subtasks. Random performance is 50%.

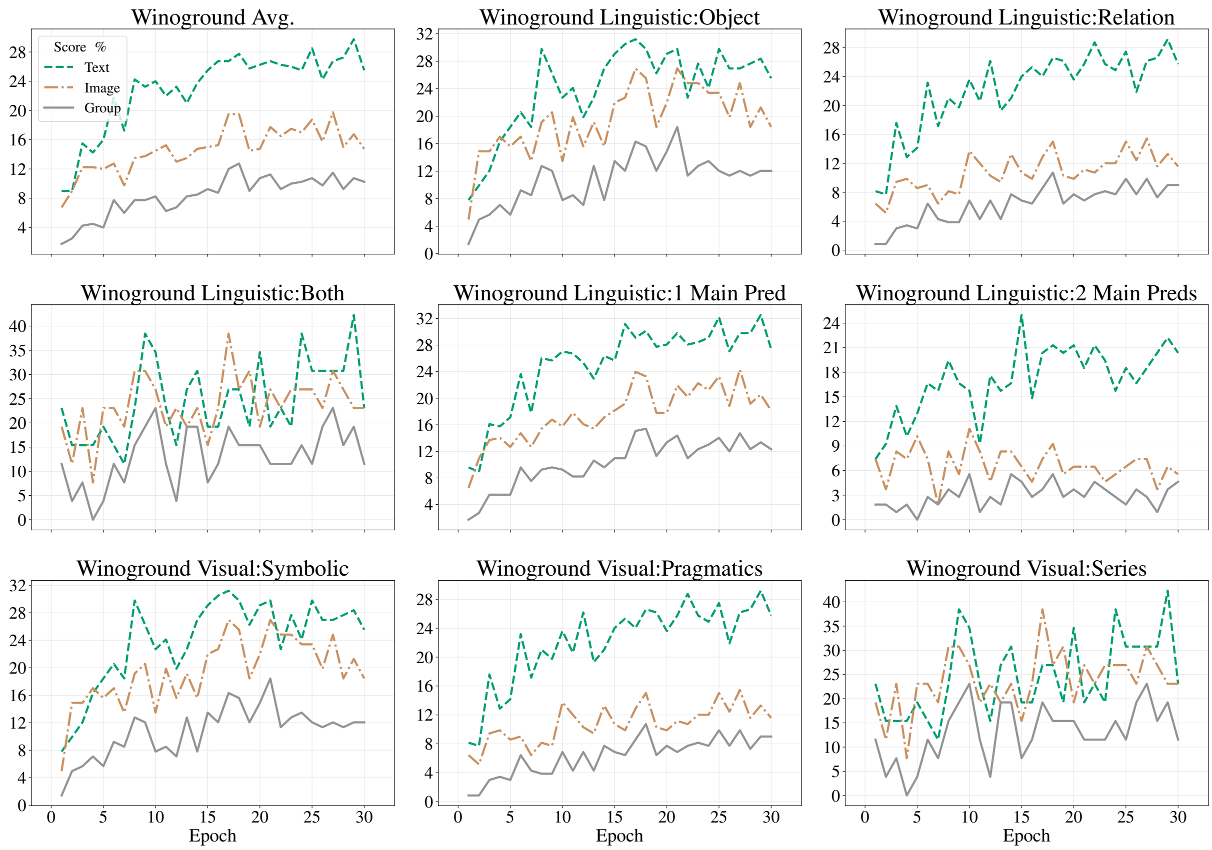

Figure 8: Training dynamics on Winoground subtasks of $\text{ALBEF}_{4\text{M}}$ pretrained with the official codebase on GCP. Random performance is 25% for Text Score and Image Score, and 16% for Group Score.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
8

☑ A2. Did you discuss any potential risks of your work?
9

☑ A3. Do the abstract and introduction summarize the paper's main claims?
7

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3

☑ B1. Did you cite the creators of artifacts you used?
3

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
2

C ☑ Did you run computational experiments?

5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
A

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
5

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
A

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.