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

Emanuele Bugliarello | Laurent Sartran | Aishwarya Agrawal | Lisa Anne Hendricks | Aida Nematzadeh

Google DeepMind

UNIVERSITY OF COPENHAGEN

Coarse-grained vs Fine-grained Tasks

Coarse-grained Image Retrieval

A person is riding a horse.

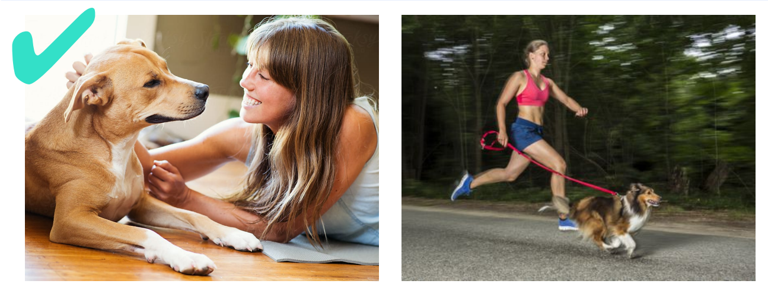
Fine-grained VALSE

pieces	existence	plurality	counting	relations	actions	coreference
instruments	*existential quantifiers*	*semantic number*	*balanced, adversarial, small numbers*	*prepositions*	*replacement, actant swap*	*standard, clean*
caption (blue) / foil (orange)	*There are no animals / animals shown.*	*A small copper vase with some flowers / exactly one flower in it.*	*There are four / six zebras.*	*A cat plays with a pocket knife on / underneath a table.*	*A man / woman shouts at a woman / man.*	*Buffalos walk along grass. Are they in a zoo? No / Yes.*
image						

Fine-grained VSR

Caption: The cow is **ahead of** the person
Label: FALSE

Fine-grained SVO-Probes



A woman **lying** with a dog

Fine-grained Winoground

(a)

(b)

some plants surrounding a lightbulb **(a)**

a lightbulb surrounding some plants **(b)**

Baselines

Coarse-grained

- **ALBEF** (baseline)
- **BLIP** (~ALBEF w/ autoregressive LM)

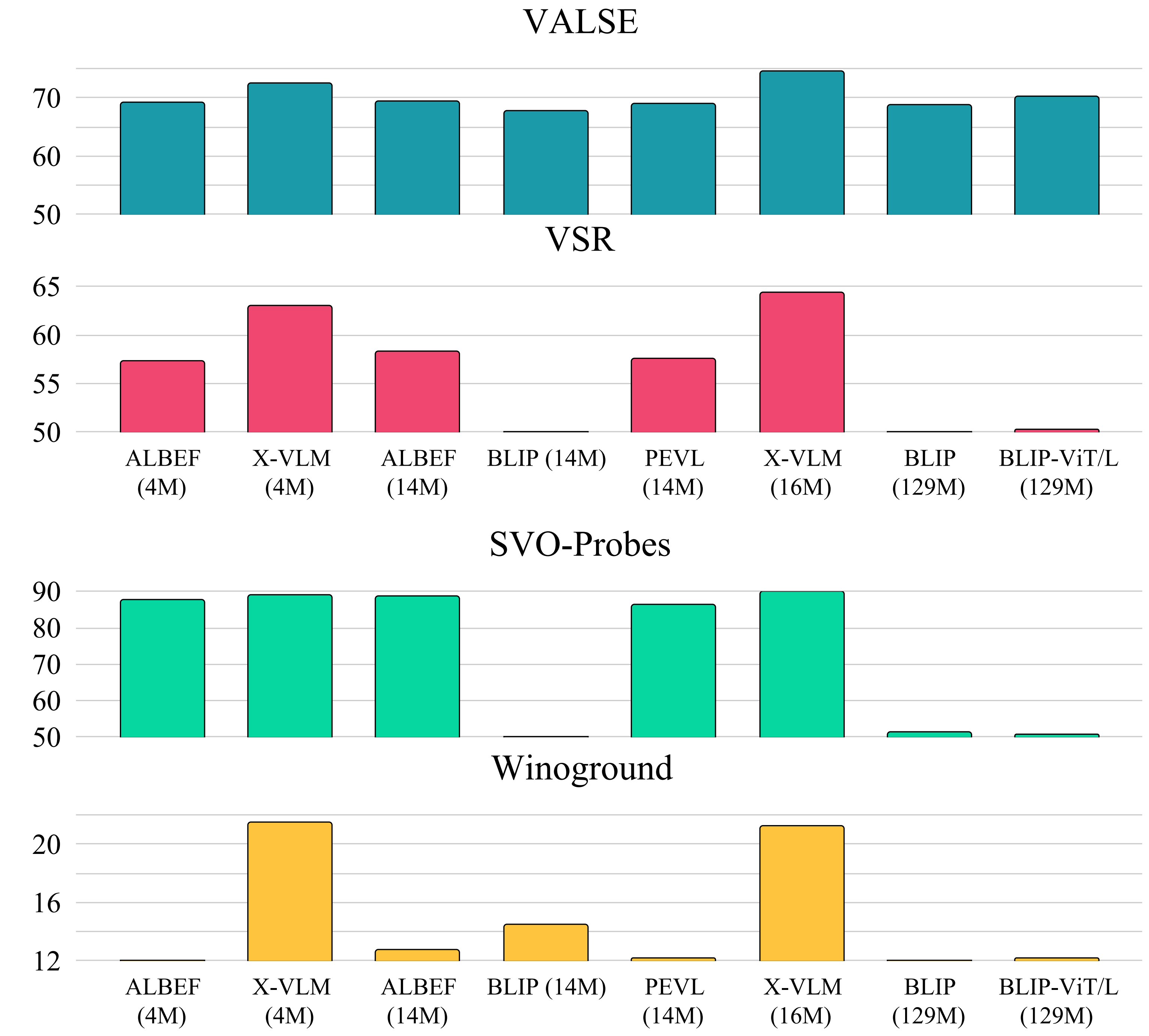
Fine-grained

- **PEVL** (ALBEF + bbox MLM)
- **X-VLM** (ALBEF + bbox regression)

Conclusion

- X-VLM > models with more data and params
- Localisation losses can be crucial
- Fine-grained skills are learned at different times
- Modelling spatial relations is promising!

Which models perform well on fine-grained tasks?

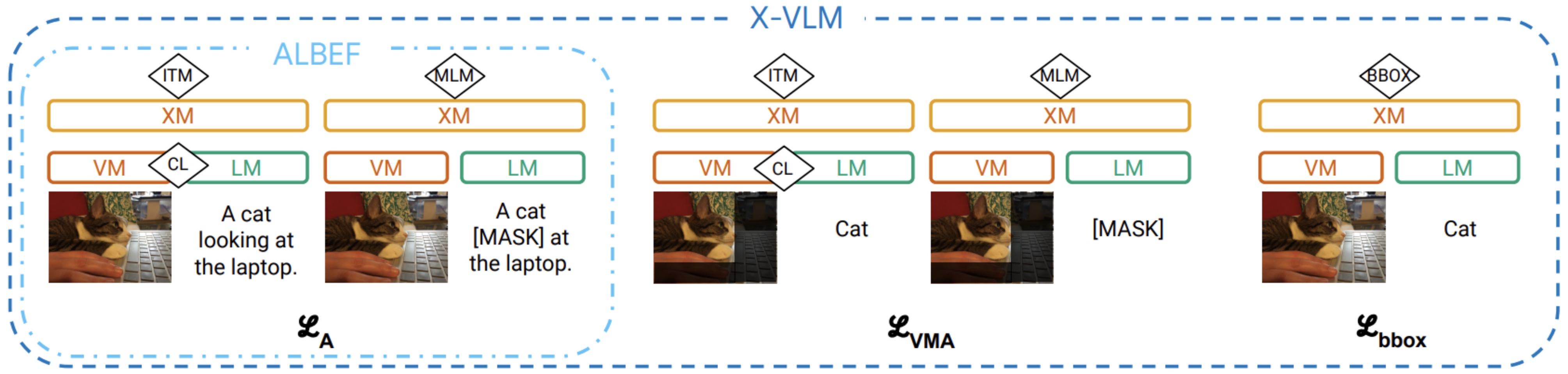


- Localisation can help fine-grained understanding
- But the localisation loss matters!
- More data does not help as much as modelling

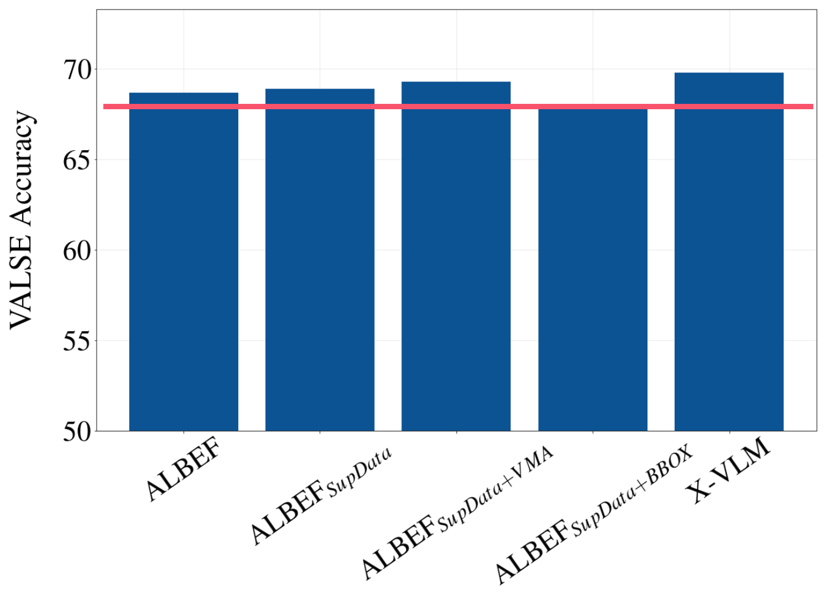
Data and Losses for Fine-grained Tasks (controlled setup)

X-VLM adds 2 additional losses to ALBEF

- BBOX: Regress object bounding box coordinates
- VMA: Same as ALBEF but applied on Visually-Masked image regions



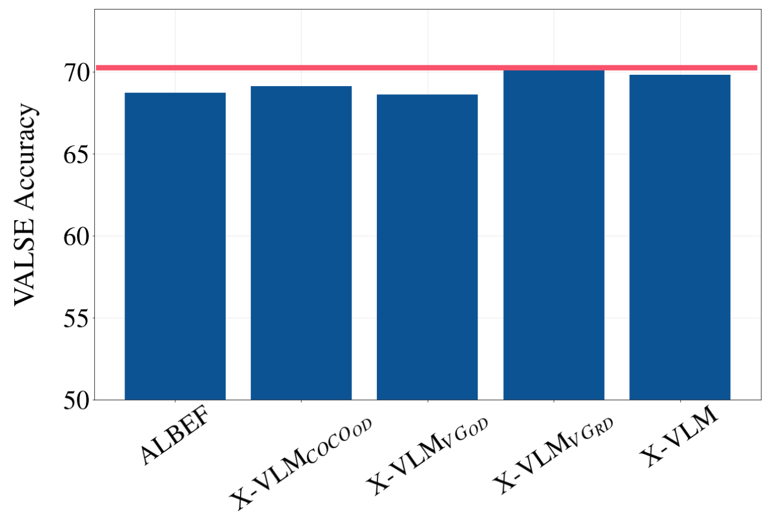
- Just adding supervised data does not help
- VMA is slightly more helpful than BBOX
- VMA+BBOX is best



X-VLM adds 3 supervised datasets to ALBEF

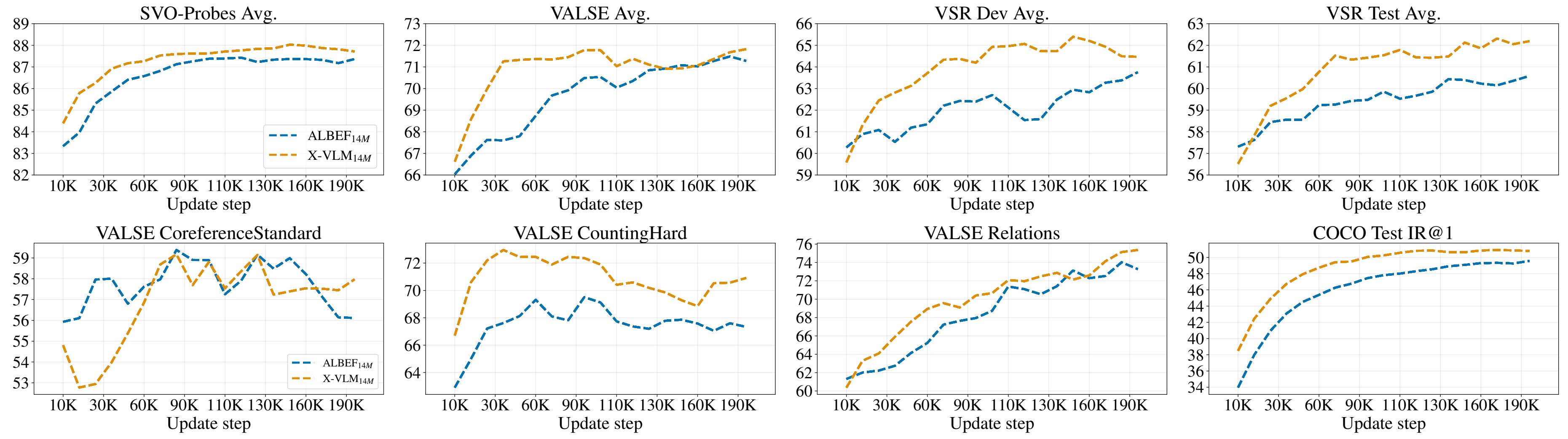
Object detection
- COCO$_{OD}$
- VG$_{OD}$

Region description
- VG$_{RD}$



- VG$_{RD}$ is the most useful dataset

Dynamics of Fine-grained Tasks



- Performance can fluctuate during training, even becoming ***worse***
- A single checkpoint might not be adequate for all skills!

Skill	Datasets	Correlation (Spearman/Pearson)
Action Replacement	VALSE Action Replacement + SVO-Verbs	55 / 67
Actant Swap	VALSE Actant Swap + SVO-Subjects	-13 / -11
Spatial Relations: Overall	VALSE Spatial Relations + VSR Average	75 / 65
Spatial Relations: Topological	VALSE Spatial Relations + VSR Topological	<-40

- Performance on similar tasks does ***not*** always correlate