Google DeepMind

UNIVERSITY OF COPENHAGEN

Weakly-Supervised Learning of Visual Relations in Multimodal Pretraining



Emanuele Bugliarello

Aida Nematzadeh

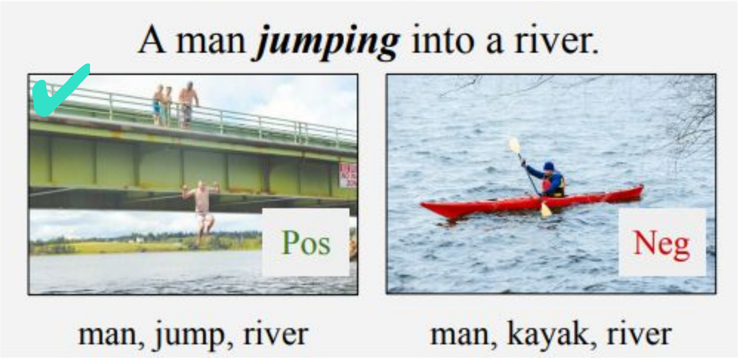
Lisa Anne Hendricks

EMNLP 2023

VLMs Struggle with Fine-grained Tasks

Strong vision-language models still struggle with fine-grained understanding

Fine-grained Verb Understanding

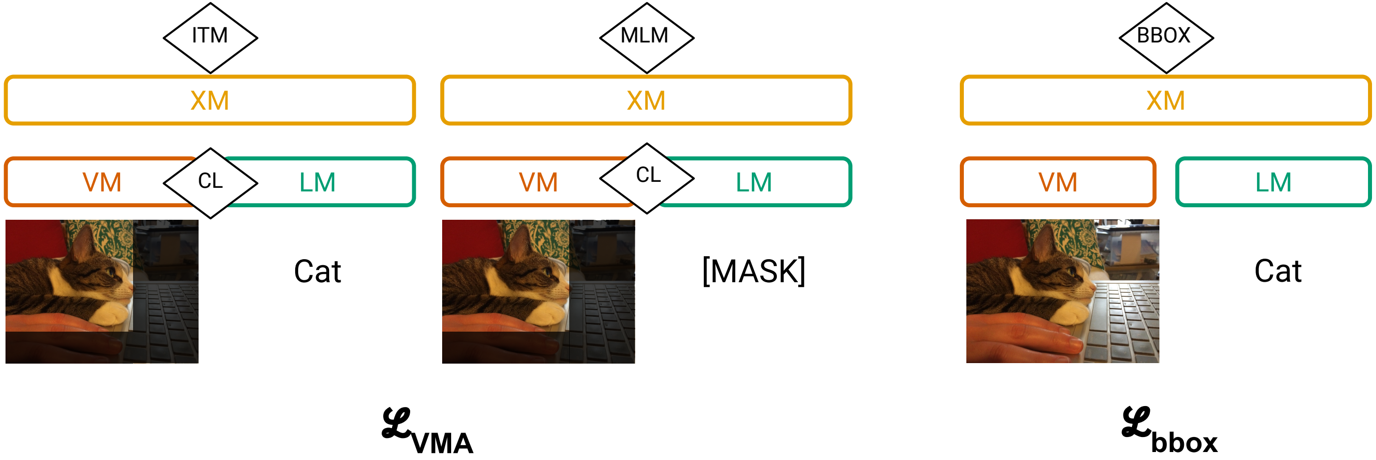


Fine-grained VSR

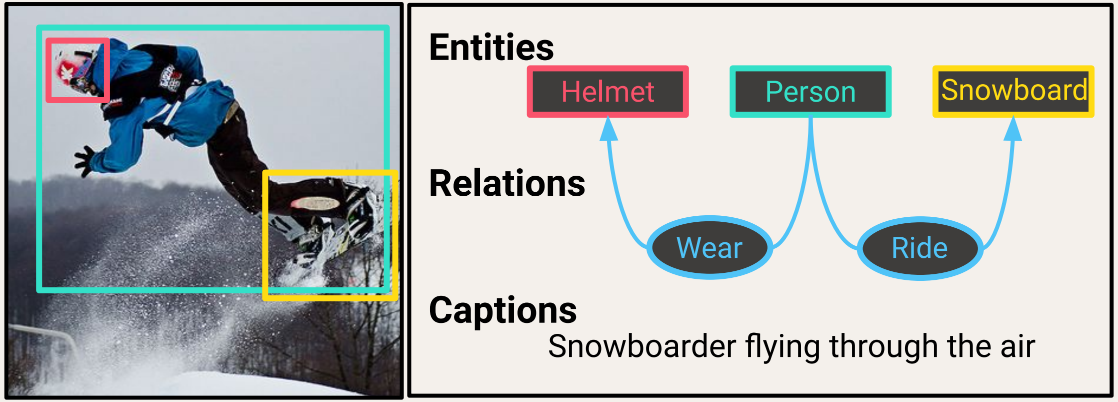
Caption: The cow is **ahead of** the person
Label: FALSE

… but supervised localisation helps

X-VLM (Zeng+ ICML'22), a model with localization supervision, outperforms larger models trained on more data on fine-grained tasks (Bugliarello+ ACL'23)



Can modelling ***visual relations*** improve fine-grained understanding?



Supervised Visual Relations for Fine-grained Understanding

- How can we incorporate visual relation data into multimodal pretraining?

- Does modelling visual relations impact task performance?

- How do our two new contributions impact task performance?

Method 1: Verbalised Scene Graphs (VSG)

Data-to-text strategy

Method 1: Verbalised Scene Graphs (VSG)

Data-to-text strategy

1. Sample K scene graph triplets

Method 1: Verbalised Scene Graphs (VSG)

Data-to-text strategy

1. Sample K scene graph triplets
2. Sort them on the subject location

Method 1: Verbalised Scene Graphs (VSG)

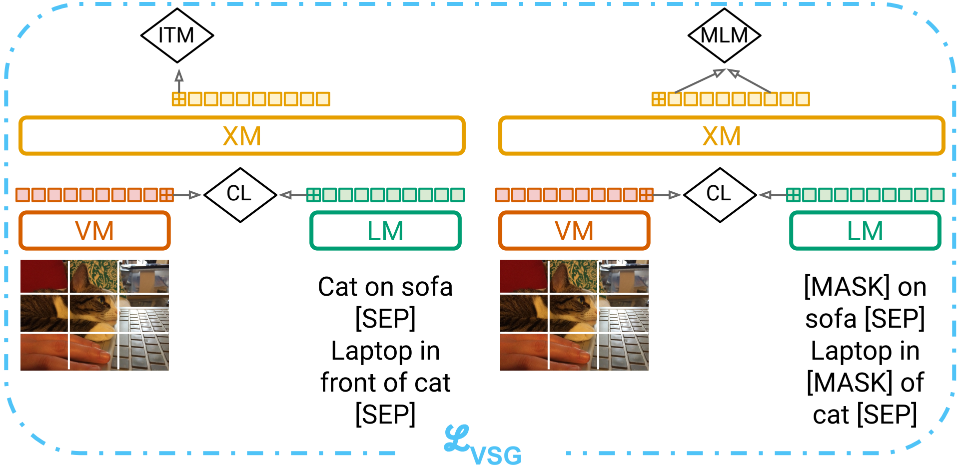
Data-to-text strategy

1. Sample K scene graph triplets
2. Sort them on the subject location
3. Verbalise into a caption: “`[CLS]` $s_1 r_1 o_1$ `[SEP]` … $s_K r_K o_K$ `[SEP]`”

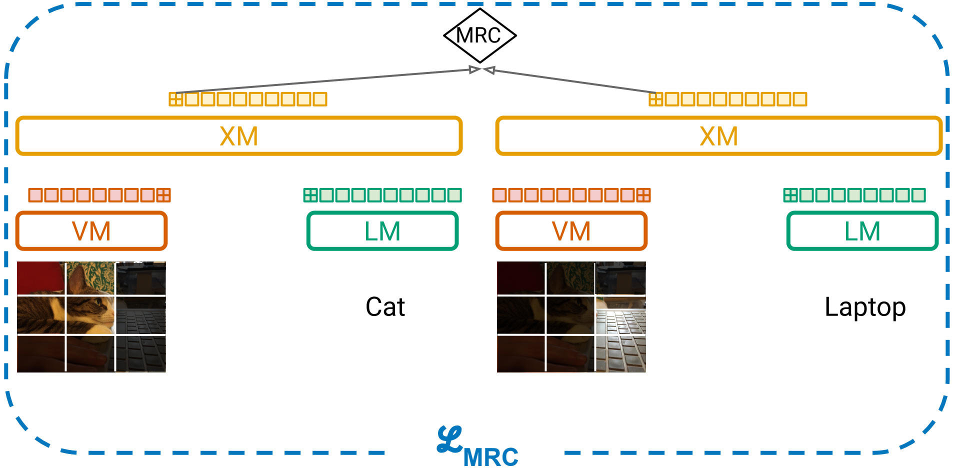
Method 1: Verbalised Scene Graphs (VSG)

Data-to-text strategy

1. Sample K scene graph triplets
2. Sort them on the subject location
3. Verbalise into a caption: "`[CLS]` $s_1 r_1 o_1$ `[SEP]` … $s_K r_K o_K$ `[SEP]`"
4. Apply standard (*e.g.*, ALBEF) image—text losses

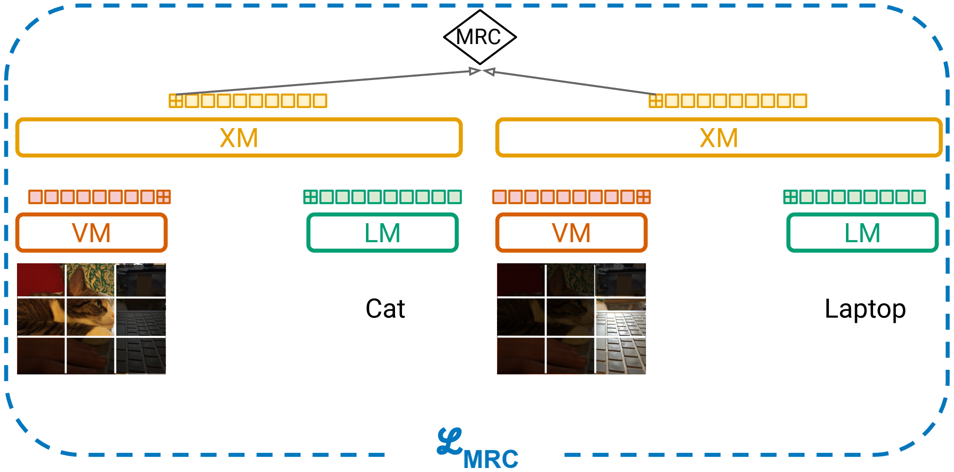


Method 2: Masked relation classification (MRC)



Pretraining cross-entropy objective

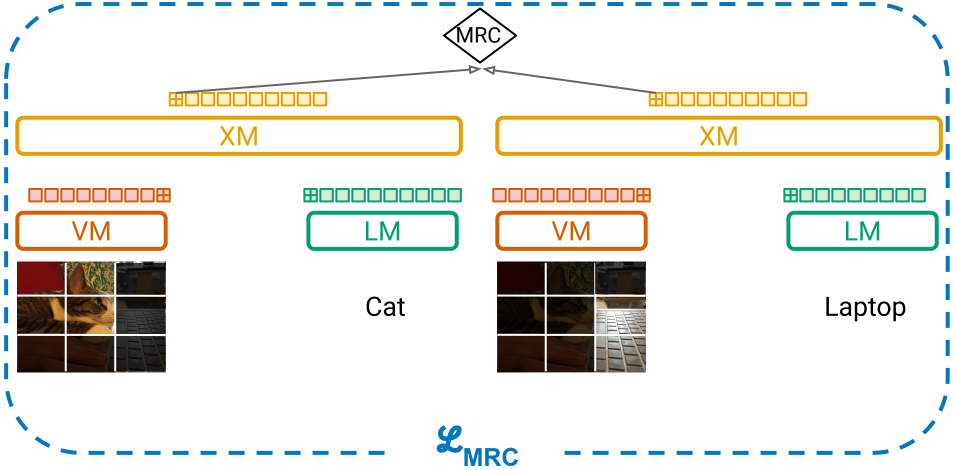
Method 2: Masked relation classification (MRC)



Pretraining cross-entropy objective

1. Encode a triplet's Subject and Object independently (by masking their visual contexts)

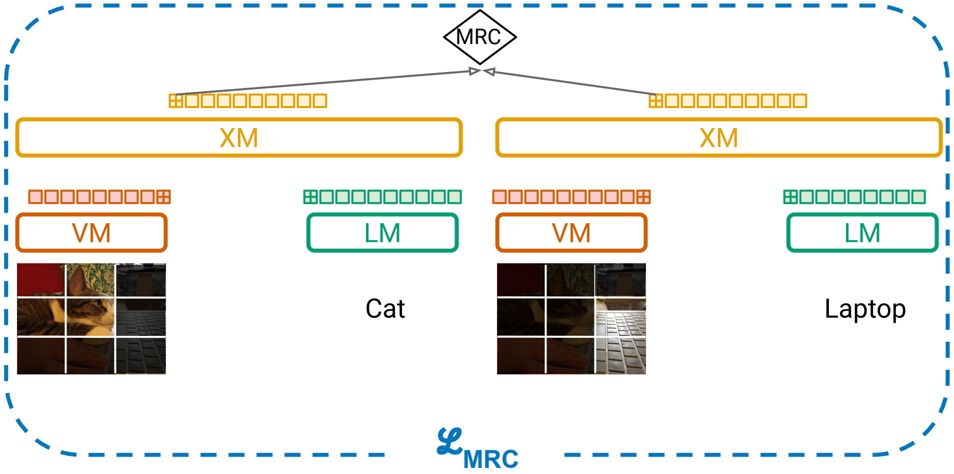
Method 2: Masked relation classification (MRC)



Pretraining cross-entropy objective

1. Encode a triplet's Subject and Object independently (by masking their visual contexts)
2. Pool their final cross-modal representations ([CLS] token)

Method 2: Masked relation classification (MRC)



Pretraining cross-entropy objective

1. Encode a triplet's Subject and Object independently (by masking their visual contexts)
2. Pool their final cross-modal representations ([CLS] token)
3. Concat pooled representations and map them to V outputs (relation labels) with an MLP

Supervised Visual Relations for Fine-grained Understanding

- How can we incorporate visual relation data into multimodal pretraining?
  Two new methods: verbalised scene graphs & masked relation classification

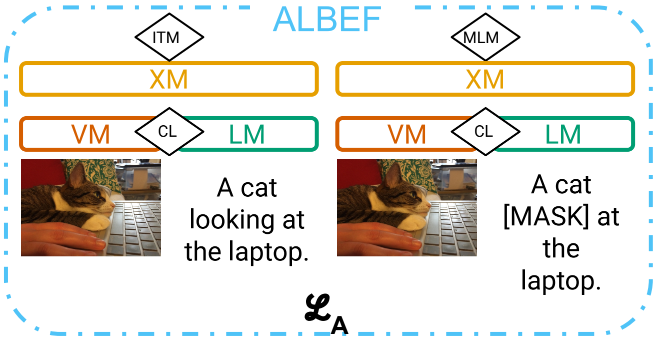
- Does modelling visual relations impact task performance?

- How do our two new contributions impact task performance?

Experimental Setup: Models

Trained on 3M and 13M data points

Experimental Setup: Models

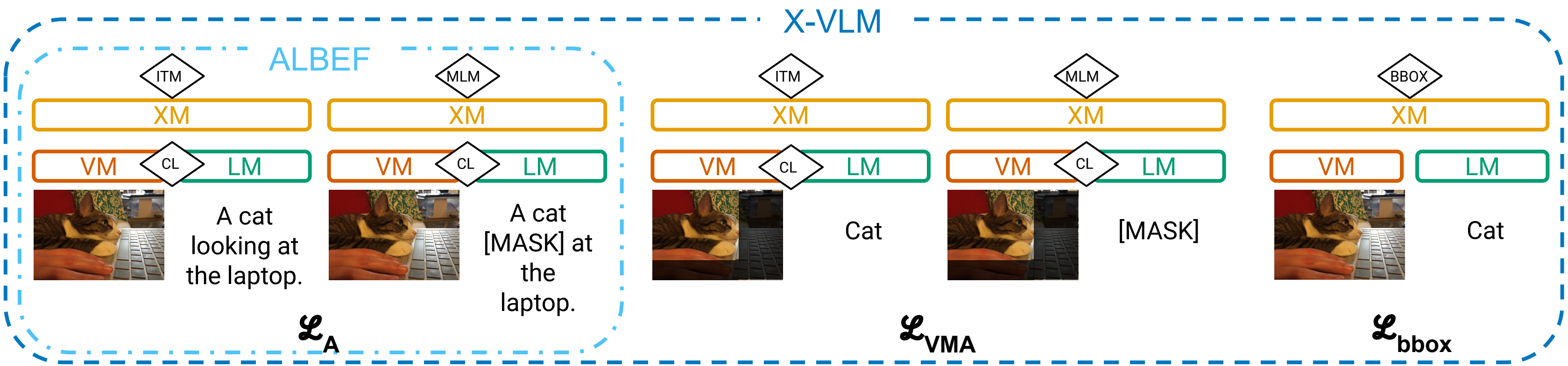


Baselines

ALBEF (coarse-grained)

Trained on 3M and 13M data points

Experimental Setup: Models



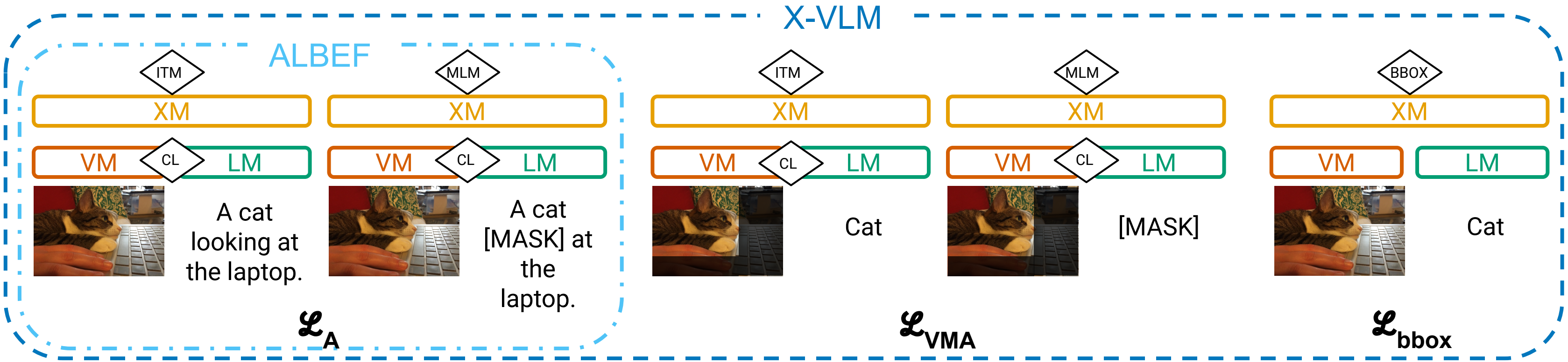
Baselines

ALBEF (coarse-grained)

X-VLM (fine-grained: ALBEF+bbox prediction)

Trained on 3M and 13M data points

Experimental Setup: Models



Baselines

ALBEF (coarse-grained)

X-VLM (fine-grained: ALBEF+bbox prediction)

Relation-enhanced (ours)

ReALBEF (ALBEF + VSG + MRC)

ReX-VLM (X-VLM + VSG + MRC)

Trained on 3M and 13M data points

Experimental Setup: Zero-Shot Tasks

Fine-grained SVO-Probes

A woman **lying** with a dog

Fine-grained VSR

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

Coarse-grained Image Retrieval

A person is riding a horse.

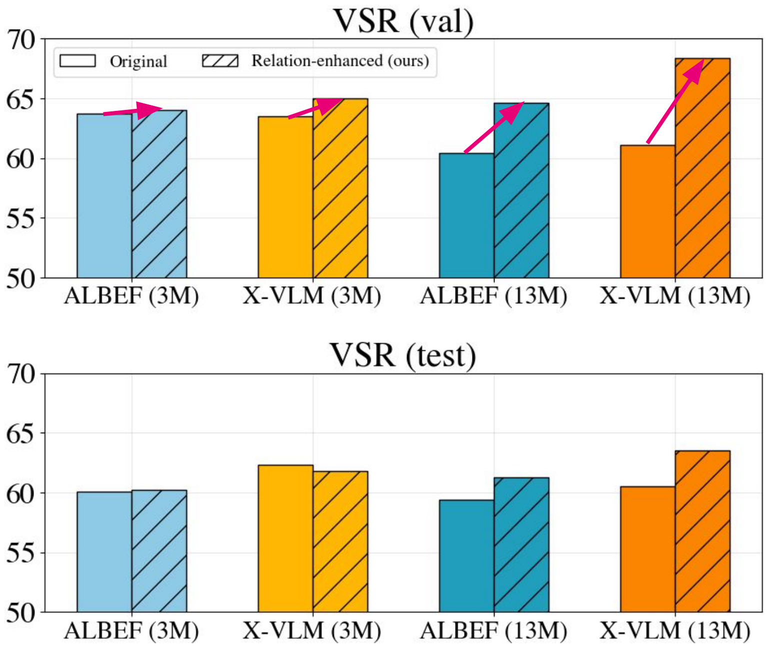
Fine-grained VALSE

pieces	existence	plurality	counting	relations	actions	coreference
instruments	*existential quantifiers*	*semantic number*	*balanced, adver-sarial, 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 Dense Image Retrieval

A person with long hair and beige sweater is smiling and riding …

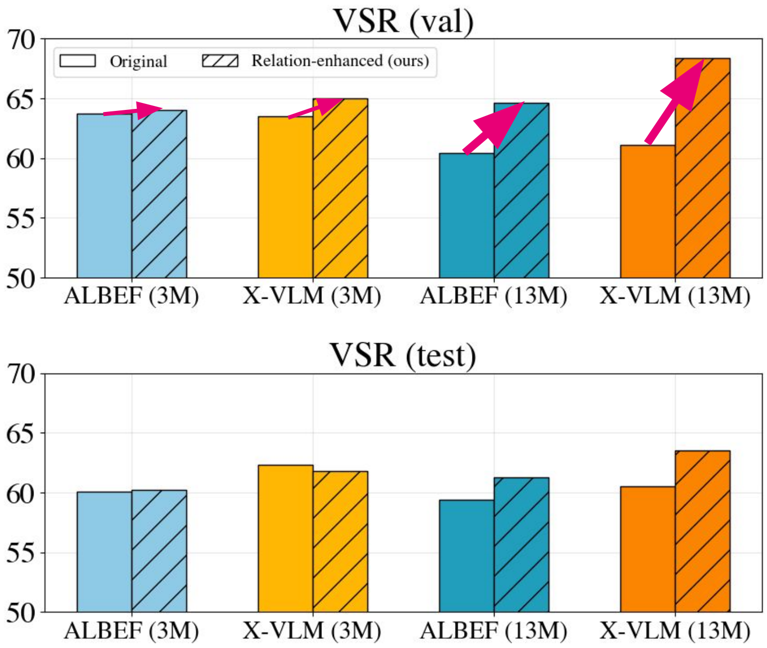
Results: Spatial Reasoning



- Generally, spatial reasoning improves when including VSG and MRC

Caution: VSR val/test performance do not always correlate!

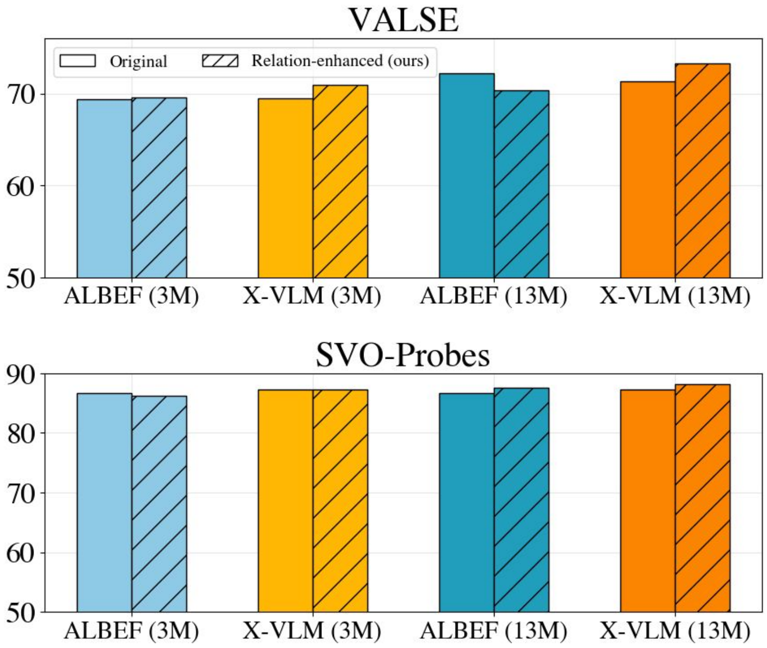
Results: Spatial Reasoning



- Generally, spatial reasoning improves when including VSG and MRC
- Gains of our approaches increase when pretraining on more data (13M vs. 3M)

Caution: VSR val/test performance do not always correlate!

Results: Other Fine-grained Tasks



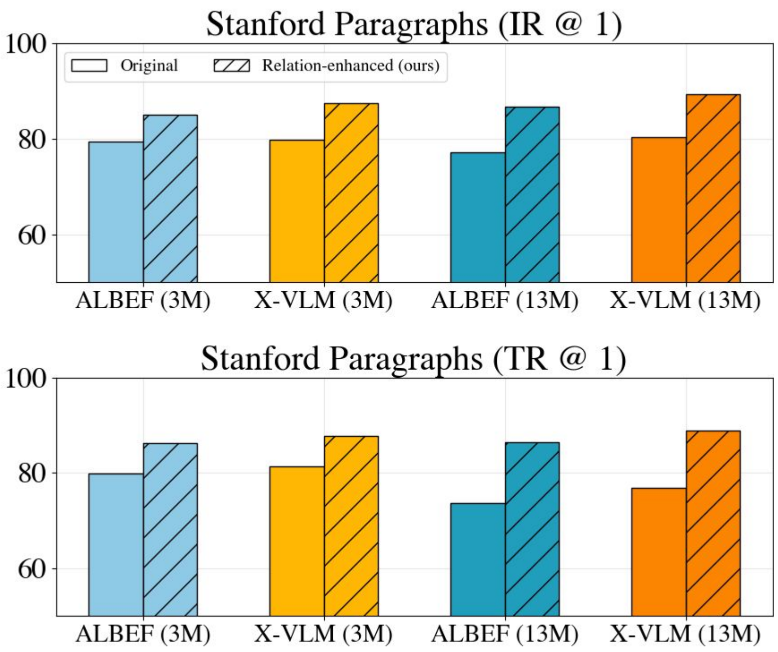
- ReX-VLM (13M) performs best across all the fine-grained tasks

  ⇒ Relations are useful even when only being a tiny percentage of pretrain data

- ReALBEF models are on par with ALBEF

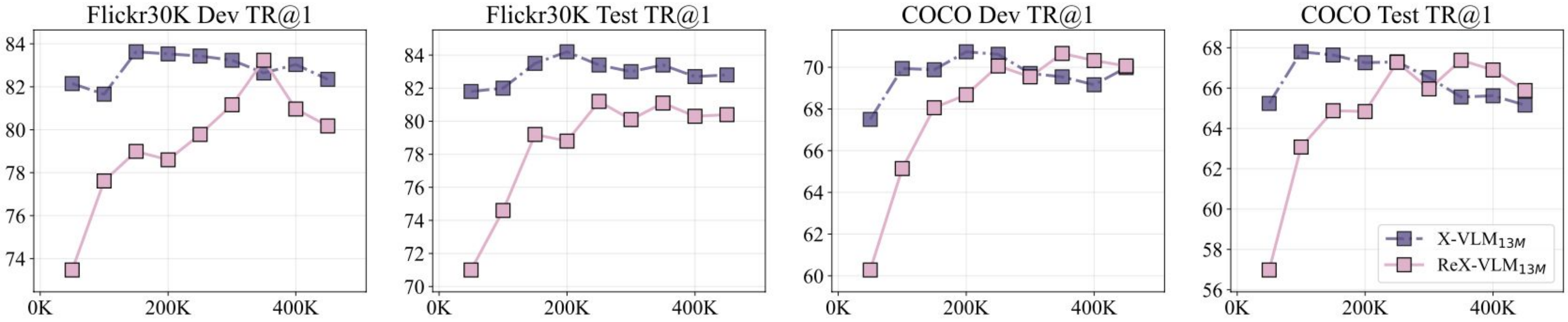
  ⇒ Harder to learn relations w/o localisation

Results: Fine-grained Dense Image–Text Retrieval



- Test our models for the ability to understand long fine-grained descriptions
- Our relation-enhanced models gain from +5.6pp to +12.8pp on this task

Results: Coarse-grained Image–Text Retrieval



- ALBEF and X-VLM quickly top out
- ReALBEF and ReX-VLM achieve comparable performance later in training

- Check our paper for more results exploring checkpoint selection strategies!

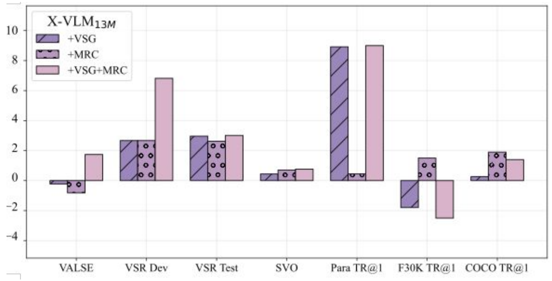
Supervised Visual Relations for Fine-grained Understanding

- How can we incorporate visual relation data into multimodal pretraining?
  Two new methods: verbalised scene graphs & masked relation classification

- Does modelling visual relations impact task performance?
  Better on fine-grained tasks & comparable for coarse-grained tasks

- How do our two new contributions impact task performance?

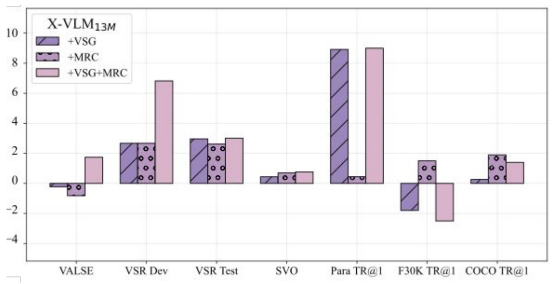
Ablations

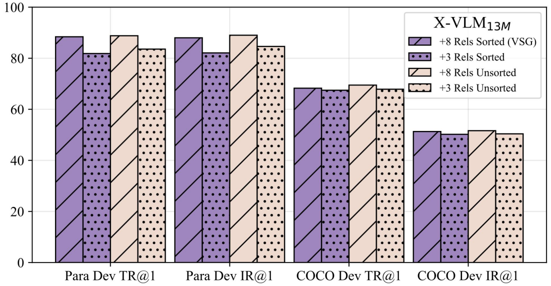


Combining VSG and MRC often leads to the best performance

VSG is key to perform well on image–paragraph retrieval

Ablations





Combining VSG and MRC often leads to the best performance

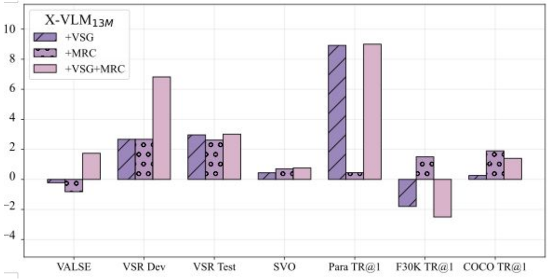
VSG is key to perform well on image–paragraph retrieval

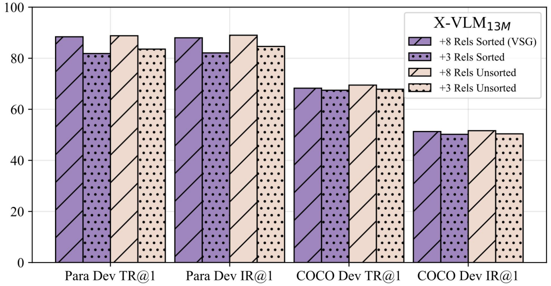
On Stanford Paragraphs

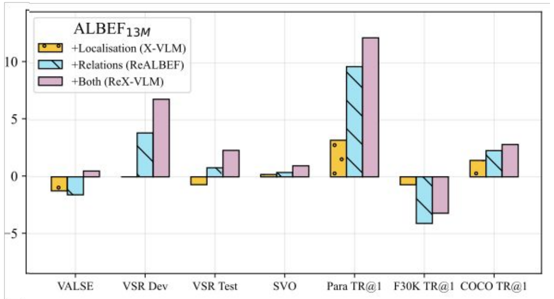
Larger #relations is important

Sorting the relations is not

Ablations







Combining VSG and MRC often leads to the best performance

VSG is key to perform well on image–paragraph retrieval

On Stanford Paragraphs

Larger #relations is important

Sorting the relations is not

Localisation + relations is best

Relations > localisation at scale

Supervised Visual Relations for Fine-grained Understanding

- How can we incorporate visual relation data into multimodal pretraining?
  Two new methods: verbalised scene graphs & masked relation classification

- Does modelling visual relations impact task performance?
  Better on fine-grained tasks & comparable for coarse-grained tasks

- How do our two new contributions impact task performance?
  Both VSG and MRC are important for best performance

Conclusions

Two new ways to use scene graph data in multimodal pretraining

Improvements on fine-grained tasks

- Small supervised datasets are useful!
- More data can probably help ⇒ automatic data generation for future work

Depending on checkpoint selection strategy, models can achieve comparable performance on coarse-grained tasks

- Open questions: balancing performance across tasks & checkpoint selection