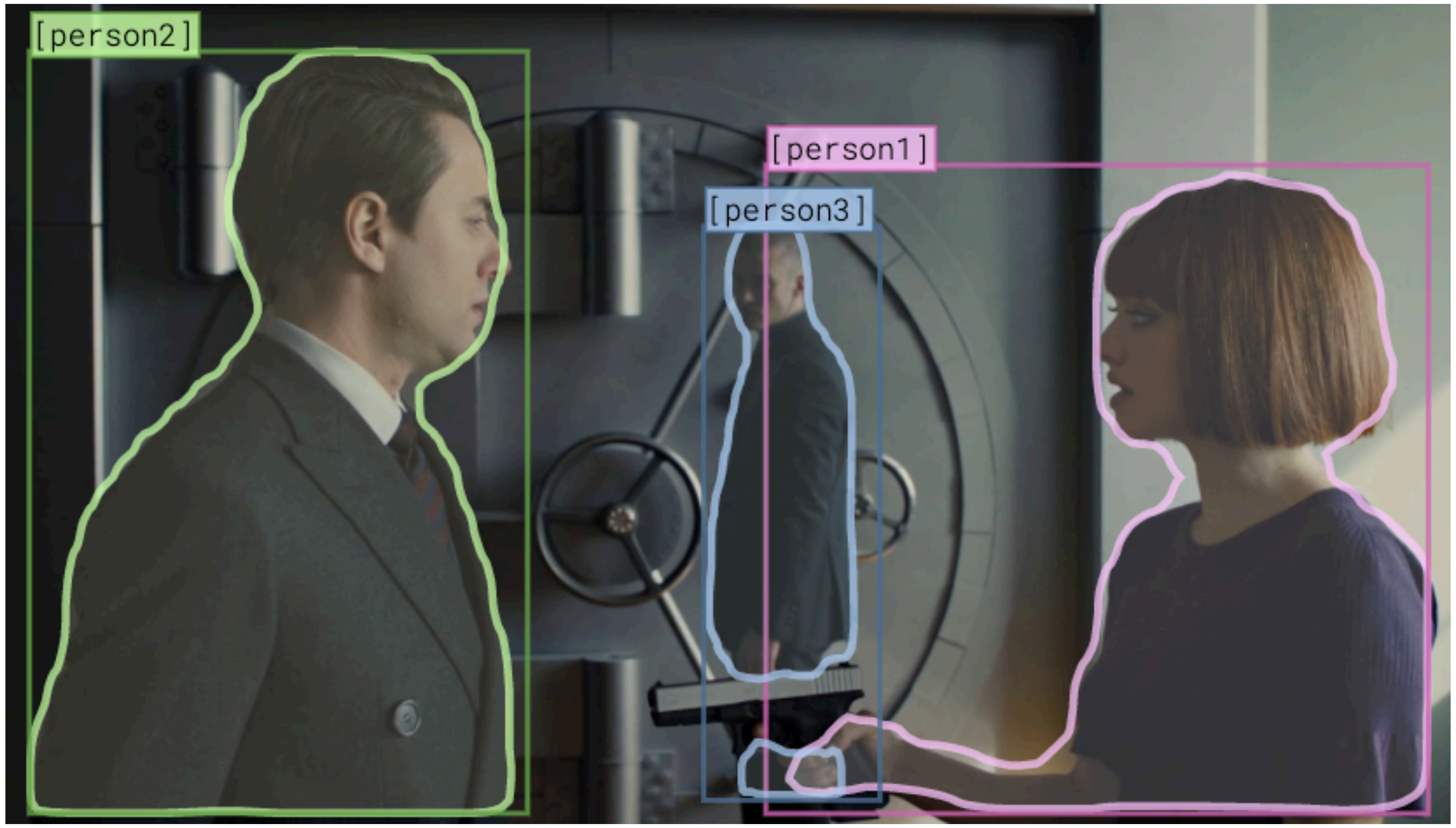
ICLR

VL-BERT: Pre-training of Generic Visual-Linguistic Representations

Weijie Su*, Xizhou Zhu*, Yue Cao, Bin Li, Lewei Lu, Furu Wei, Jifeng Dai

University of Science and Technology of China;   Microsoft Research Asia

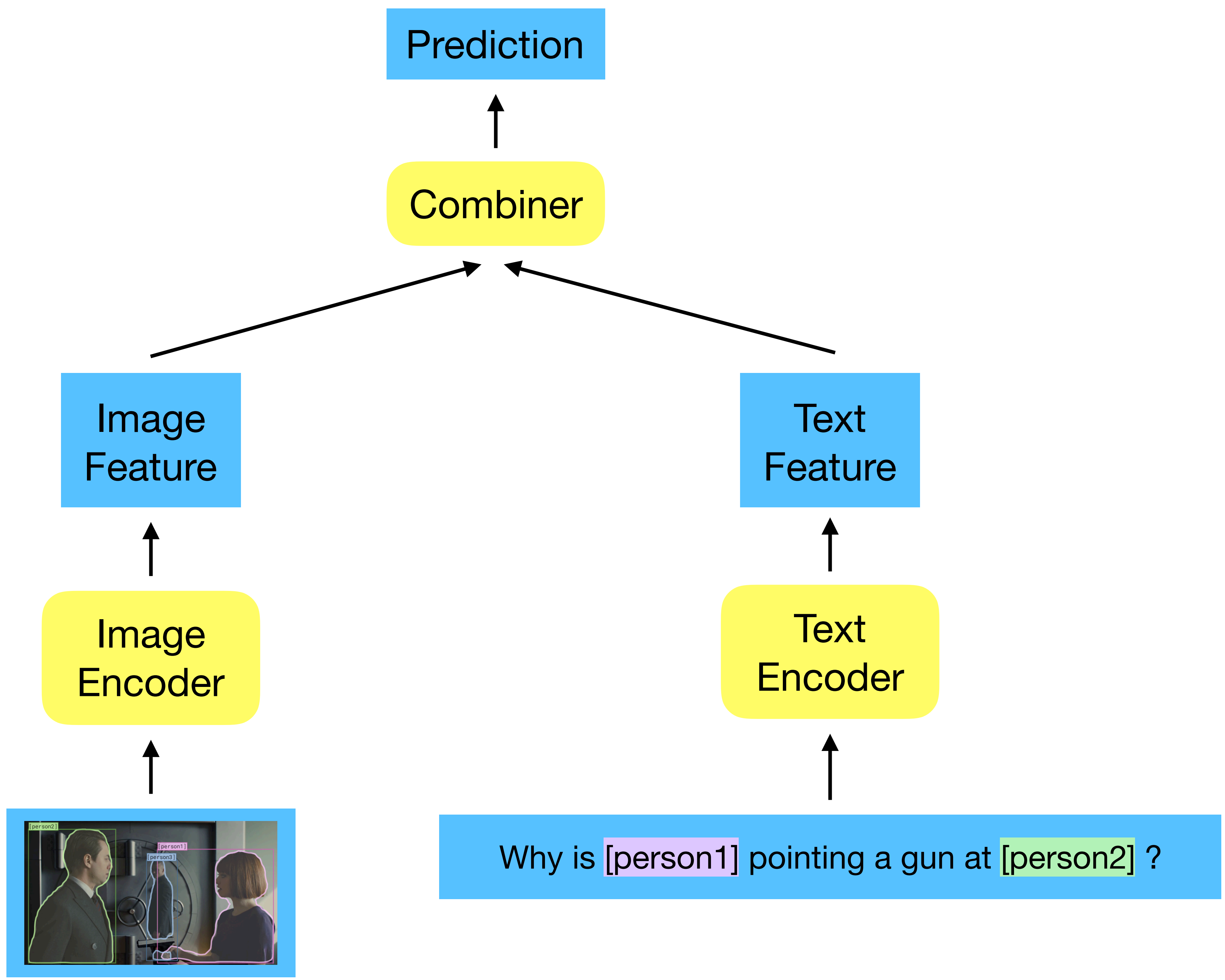
An Example of Visual-Linguistic Tasks



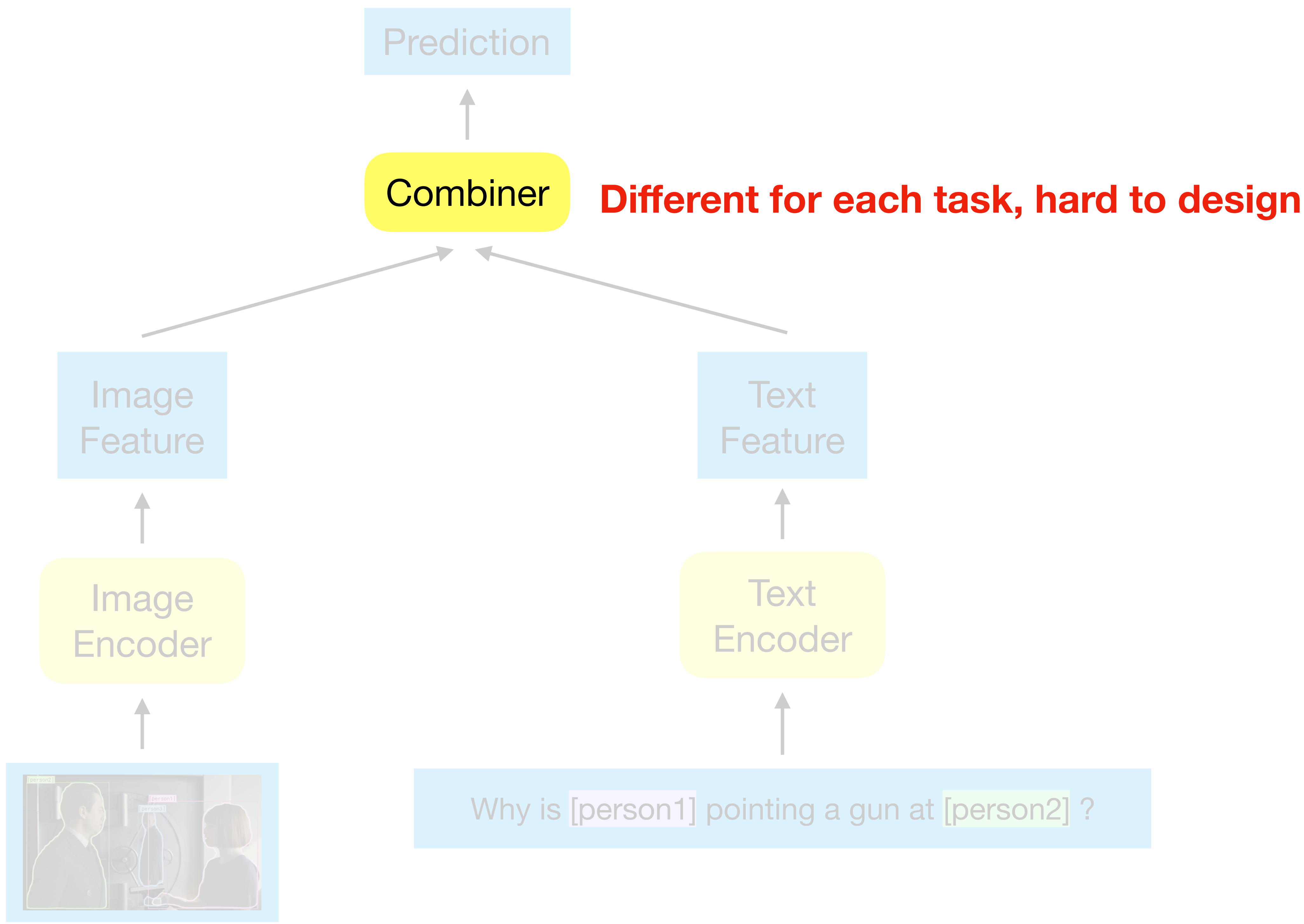
Question Why is [person1] pointing a gun at [person2] ?

Answer [person1] and [person3] are robbing the bank and [person2] is the bank manager

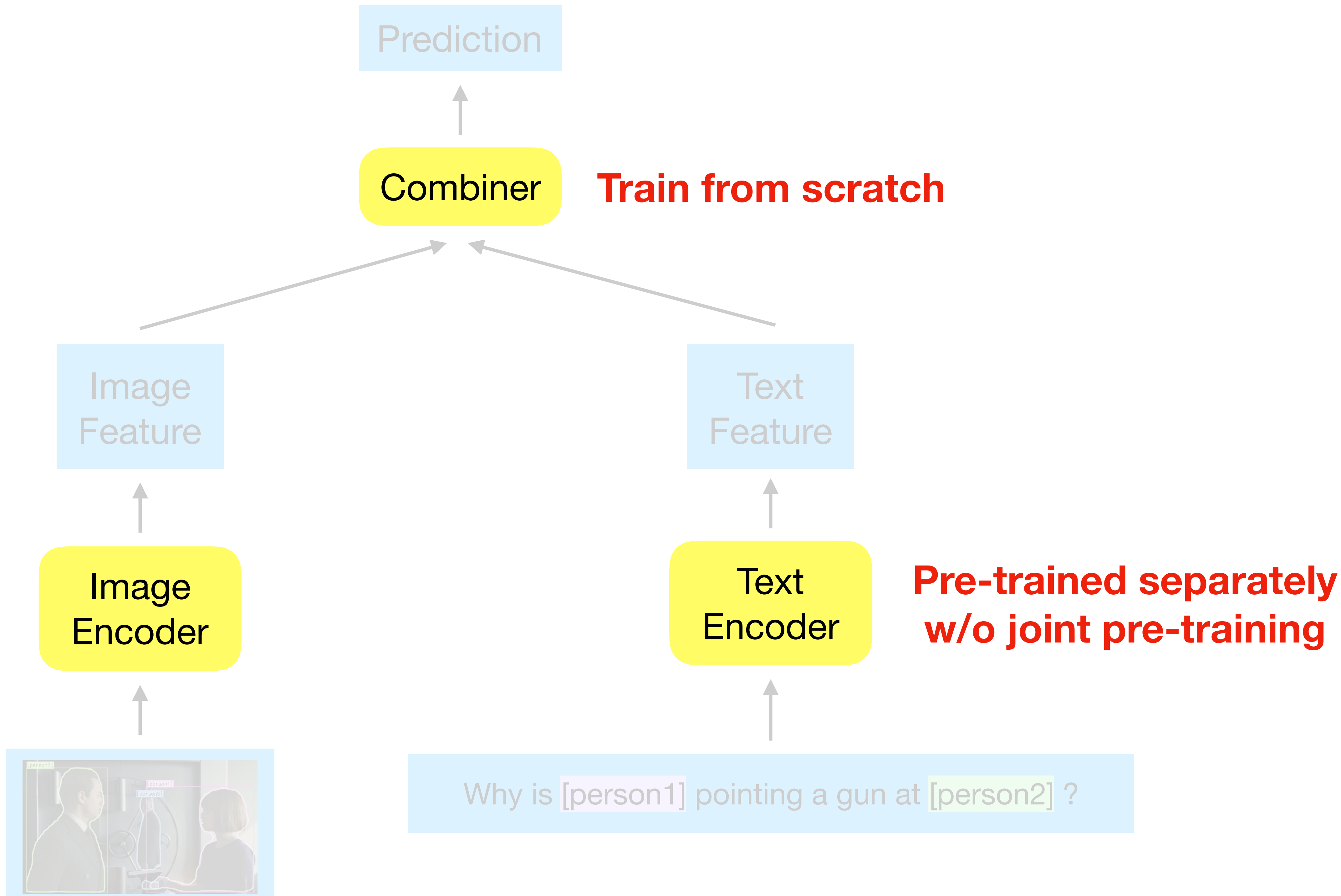
Previous Paradigm



Problem (I) High Design Cost



Problem (II) Overfitting



Inspiration

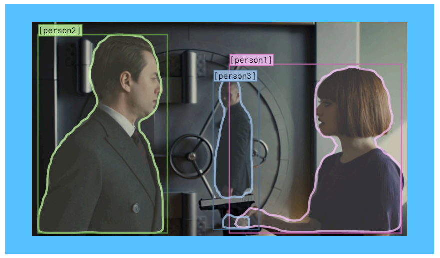
- Transformer is a unified and powerful architecture in NLP
- It can aggregate and align word embedded features
- MLM based pre-training in BERT enhances the capability

Solution

Prediction

VL-BERT

Task-agnostic
+
VL Pre-training



Why is [person1] pointing a gun at [person2] ?

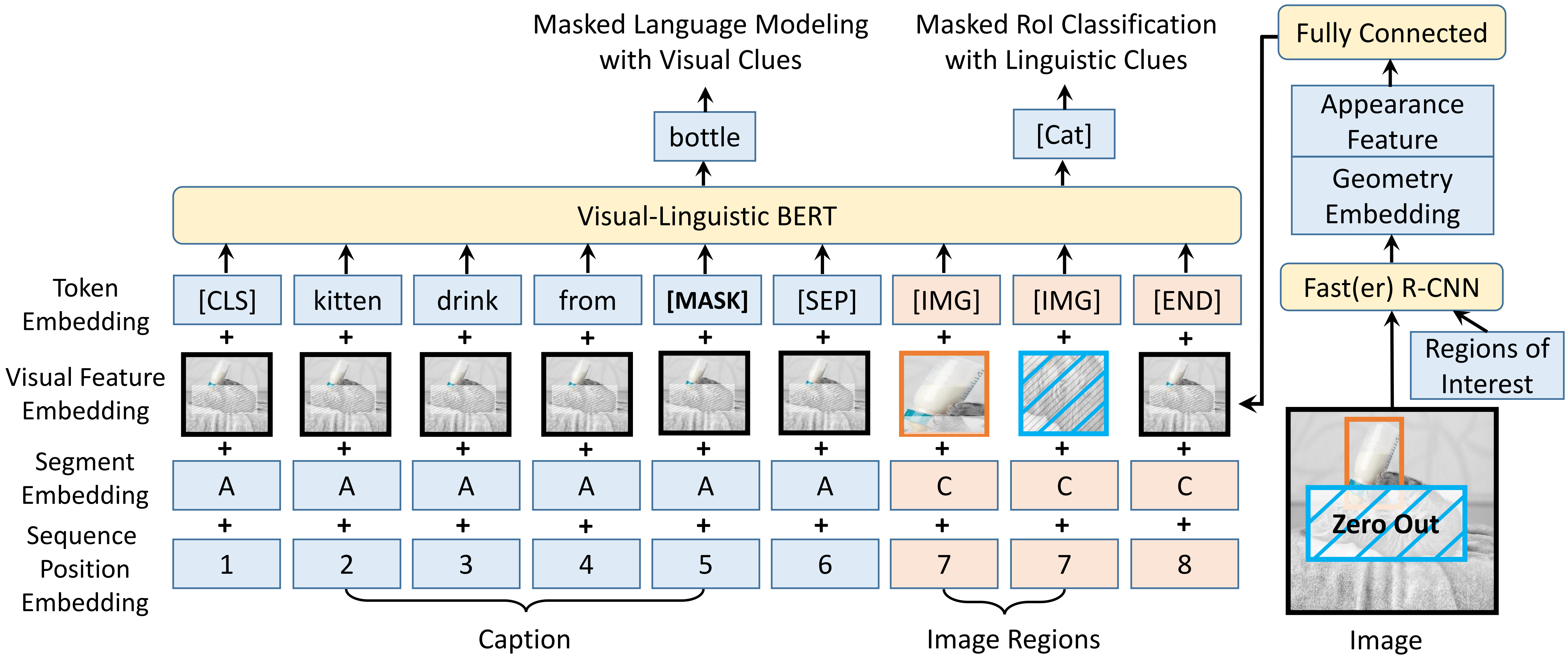
Lots of concurrent works in just 3 weeks!

	Method	Architecture	Visual Token	Pre-train Datasets	Pre-train Tasks	Downstream Tasks
Published Works	VideoBERT (Sun et al., 2019b)	single cross-modal Transformer	video frame	Cooking312K (Sun et al., 2019b)	1) sentence-image alignment 2) masked language modeling 3) masked visual-words prediction	1) zero-shot action classification 2) video captioning
Works Under Review / Just Got Accepted	CBT (Sun et al., 2019a)	two single-modal Transformer (vision & language respectively) + one cross-modal Transformer	video frame	Cooking312K (Sun et al., 2019b)	1) sentence-image alignment 2) masked language modeling 3) masked visual-feature regression	1) action anticipation 2) video captioning
	ViLBERT (Lu et al., 2019)	one single-modal Transformer (language) + one cross-modal Transformer (with restricted attention pattern)	image RoI	Conceptual Captions (Sharma et al., 2018)	1) sentence-image alignment 2) masked language modeling 3) masked visual-feature classification	1) visual question answering 2) visual commonsense reasoning 3) grounding referring expressions 4) image retrieval 5) zero-shot image retrieval
	B2T2 (Alberti et al., 2019)	single cross-modal Transformer	image RoI	Conceptual Captions (Sharma et al., 2018)	1) sentence-image alignment 2) masked language modeling	1) visual commonsense reasoning
	LXMERT (Tan & Bansal, 2019)	two single-modal Transformer (vision & language respectively) + one cross-modal Transformer	image RoI	‡ COCO Caption + VG Caption + VG QA + VQA + GQA	1) sentence-image alignment 2) masked language modeling 3) masked visual-feature classification 4) masked visual-feature regression 5) visual question answering	1) visual question answering 2) natural language visual reasoning
	VisualBERT (Li et al., 2019b)	single cross-modal Transformer	image RoI	COCO Caption (Chen et al., 2015)	1) sentence-image alignment 2) masked language modeling	1) visual question answering 2) visual commonsense reasoning 3) natural language visual reasoning 4) grounding phrases
	Unicoder-VL (Li et al., 2019a)	single cross-modal Transformer	image RoI	Conceptual Captions (Sharma et al., 2018)	1) sentence-image alignment 2) masked language modeling 3) masked visual-feature classification	1) image-text retrieval 2) zero-shot image-text retrieval
	Our VL-BERT	single cross-modal Transformer	image RoI	Conceptual Captions (Sharma et al., 2018) + BooksCorpus (Zhu et al., 2015) + English Wikipedia	1) masked language modeling 2) masked visual-feature classification	1) visual question answering 2) visual commonsense reasoning 3) grounding referring expressions

‡ LXMERT is pre-trained on COCO Caption (Chen et al., 2015), VG Caption (Krishna et al., 2017), VG QA (Zhu et al., 2016), VQA (Antol et al., 2015) and GQA (Hudson & Manning, 2019).

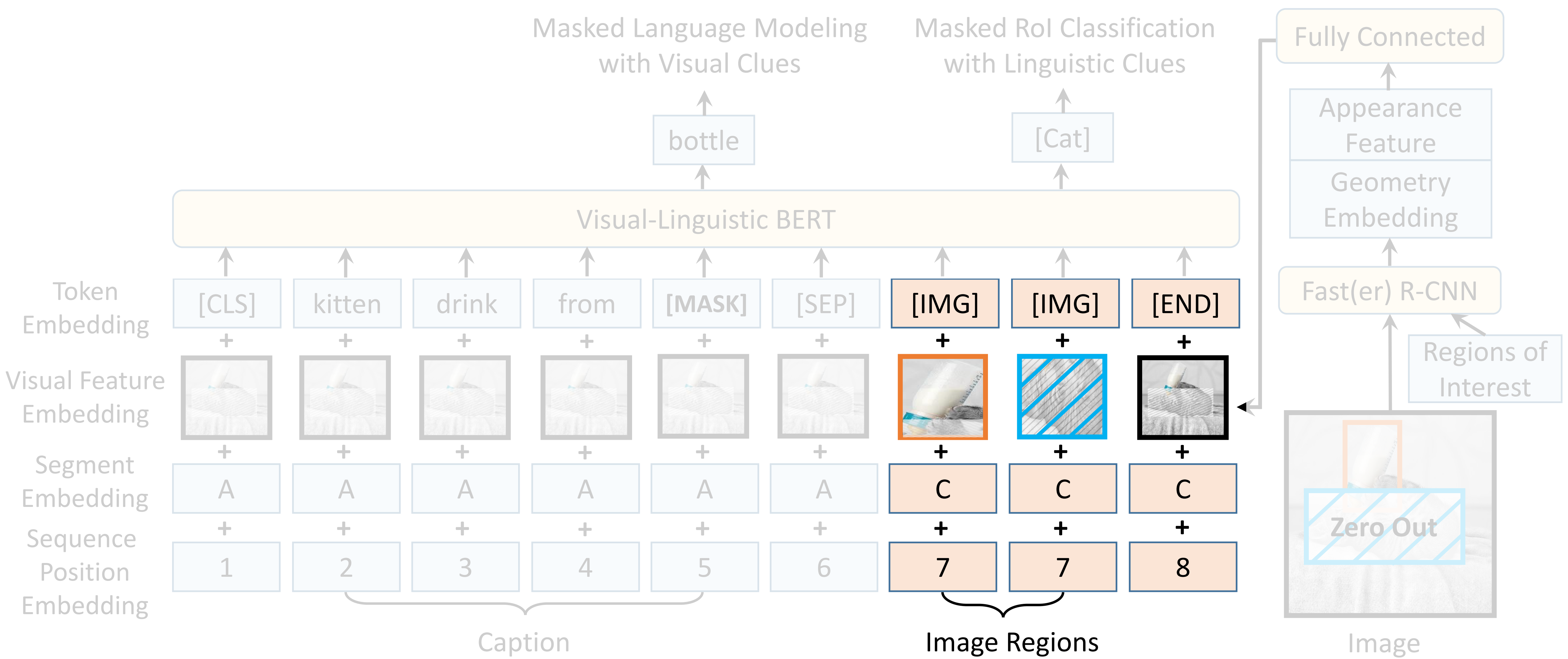
Comparison among our VL-BERT and other concurrent works for pre-training generic visual-linguistic representations

Model Architecture



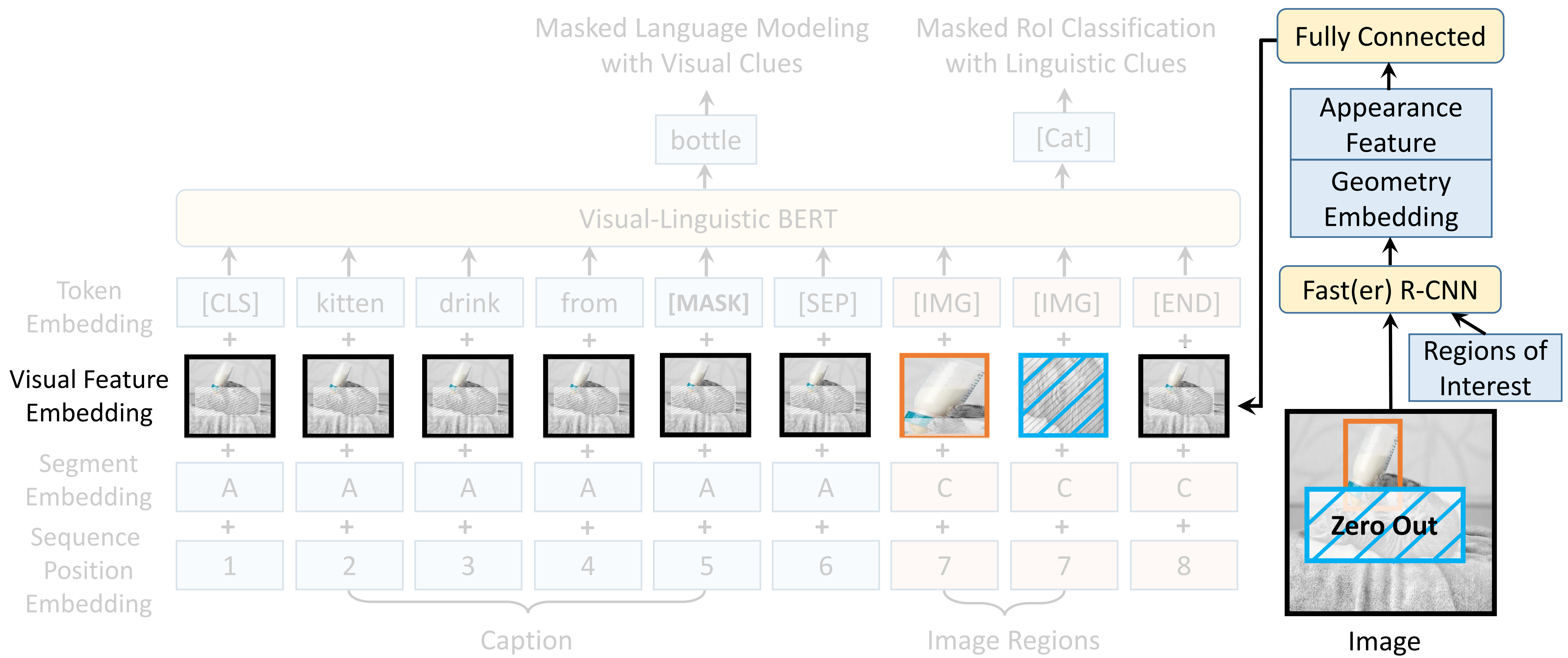
Modification (I)

Add Image Regions in Input Sequence



Modification (II)

Add Visual Feature Embedding

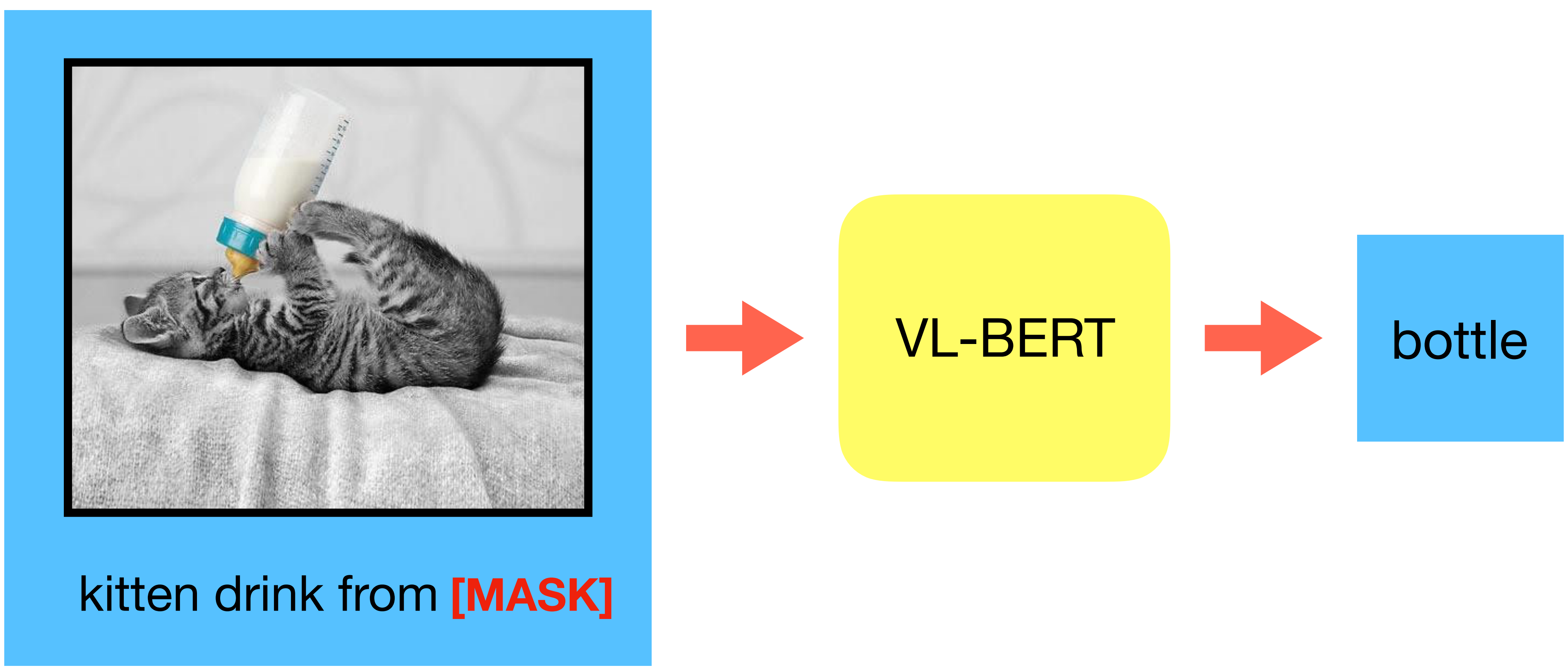


Pre-training Datasets

- Visual-Linguistic Corpus: Conceptual Captions
  - Harvested from the Internet
  - ~3M **image-text** pairs
- Text-only Corpus: English Wikipedia & BooksCorpus
  - Improve generalization over **long and complex sentences**

Pre-training Tasks #1

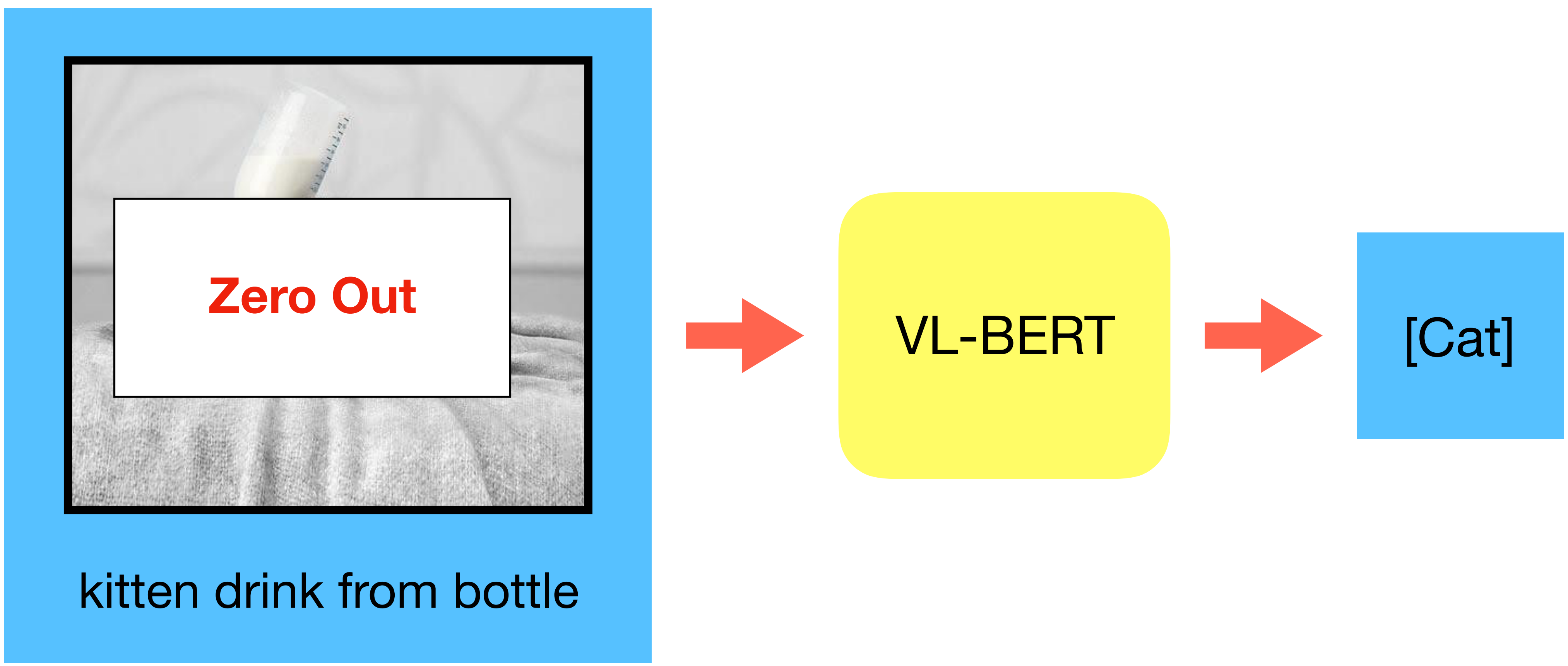
Masked Language Modeling with Visual Clues



P.S. For samples from text-only corpus, it degenerate to original MLM in BERT.

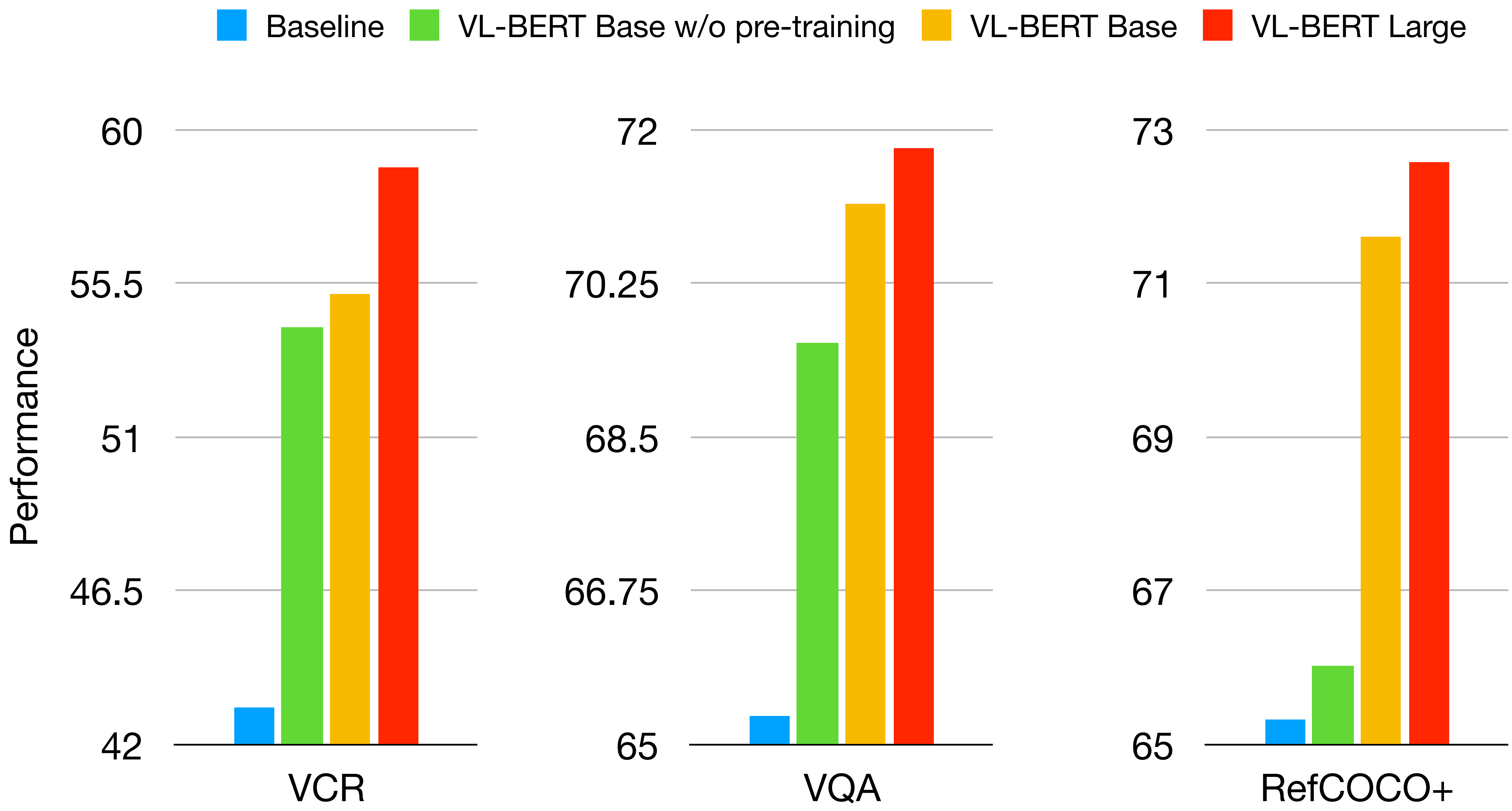
Pre-training Tasks #2

Masked RoI Classification with Linguistic Clues

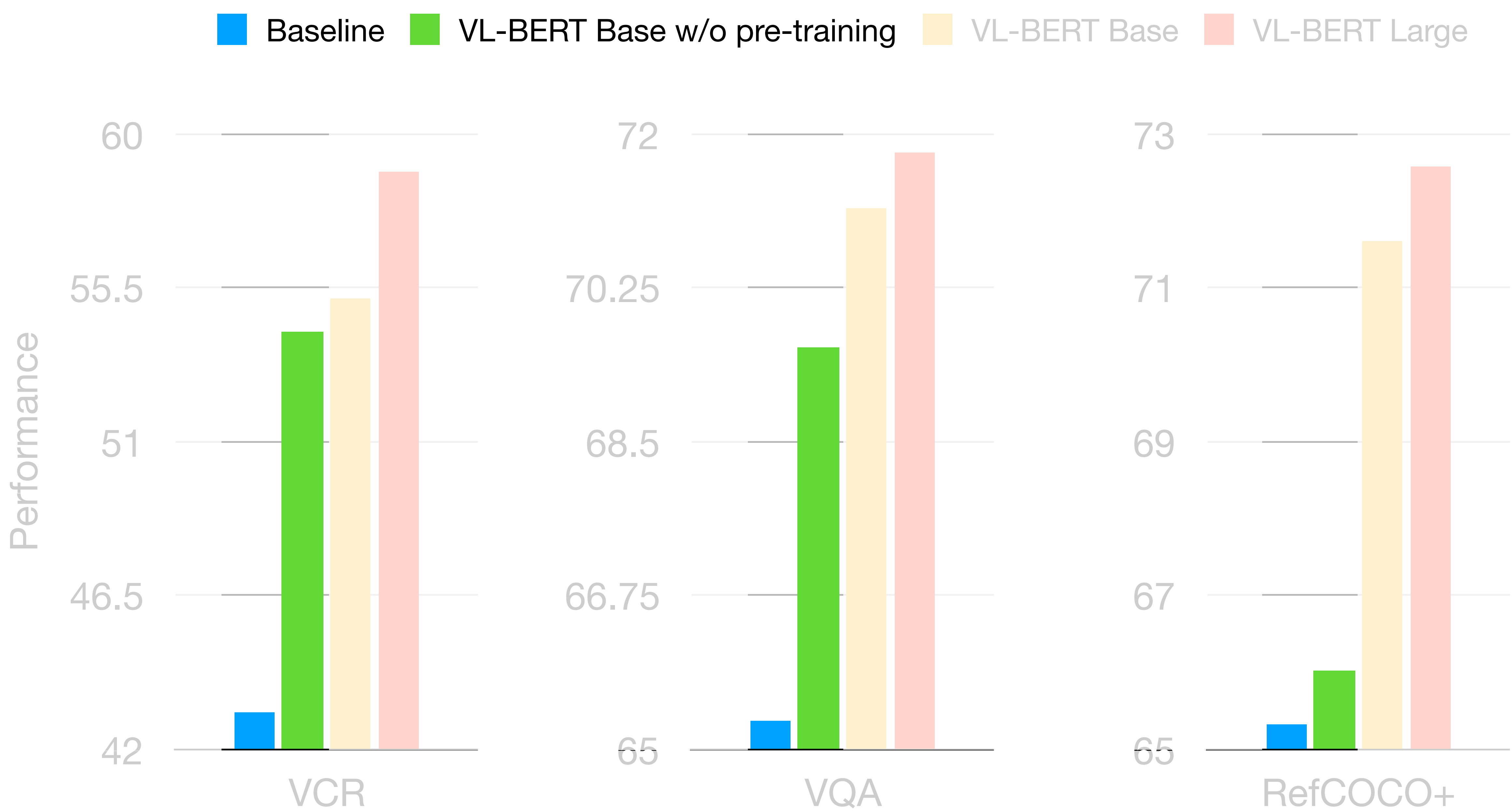


P.S. This task is not used in text-only corpus.

Results on Downstream Tasks

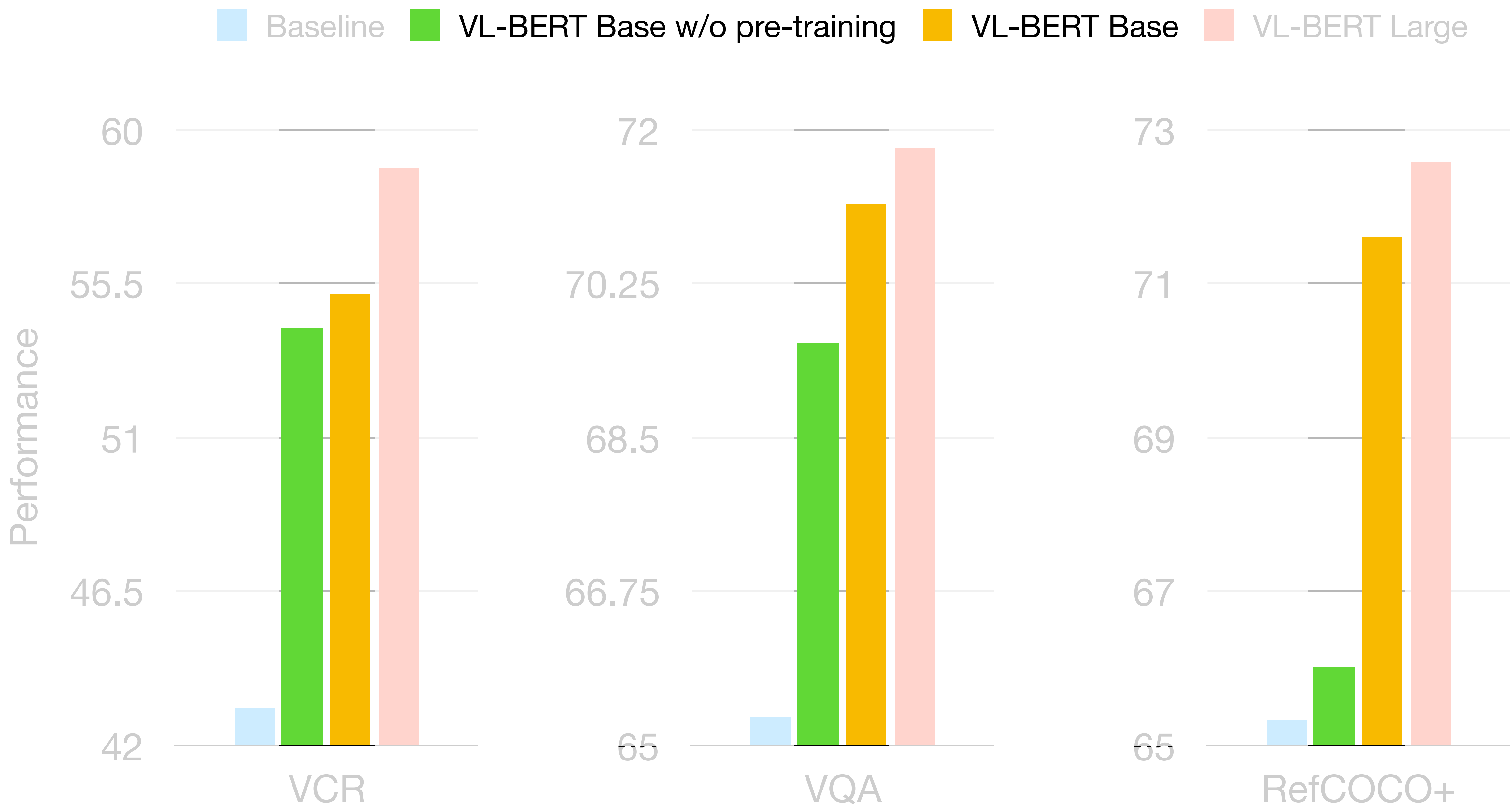


Results on Downstream Tasks



Our generic representation surpasses task-specific baseline by a large margin

Results on Downstream Tasks



Pre-training further enhances the capability

Conclusion

- A new pre-trainable generic representation for VL tasks
- Pre-training procedure can better align VL clues
- Future work: seek better pre-training tasks, benefit more downstream tasks (e.g., Image Caption Generation)

VL-BERT: Pre-training of Generic Visual-Linguistic Representations